Complex networks approach to modeling online social systems

The emergence of computational social science

TESIS DOCTORAL

Przemyslaw A. Grabowicz

Directores:
Dr. Víctor M. Eguíluz
Dr. José J. Ramasco

Ponente:
Prof. Maxi San Miguel

Universitat de les Illes Balears, 2013

Complex networks approach to modeling online social systems
The emergence of computational social science
Przemyslaw A. Grabowicz
Tesis presentada en el Departamento de Física de la Universitat de les Illes Balears

PhD Thesis

Directores: Dr. Víctor M. Eguíluz, Dr. José J. Ramasco

Tesis doctoral presentada por Przemyslaw A. Grabowicz para optar al título de Doctor, en el Programa de Física del Departamento de Física de la Universitat de les Illes Balears, realizada en el IFISC bajo la dirección de Dr. Víctor M. Eguíluz y Dr. José J. Ramasco.

Visto bueno
Director de la tesis
Dr. Víctor M. Eguíluz

Visto bueno
Director de la tesis
Dr. José J. Ramasco

Visto bueno
Ponente
Prof. Maxi San Miguel

Doctorando
Przemyslaw A. Grabowicz

To my parents.

Acknowledgments

This thesis is a result of the studies that I have performed under the supervision of Victor M. Eguiluz and Jose J. Ramasco. Victor has been supervising me from the beginning of my doctoral studies, while Jose has been supervising me in the second half of my doctoral studies. Suggestions and comments of both of them have been very helpful throughout my studies and I am thankful to them for their commitment.

This dissertation has been developed thanks to the support of CSIC JAE Predoc program and research projects of the Institute of Interdisciplinary Physics and Complex Systems in Palma de Mallorca. Parts of this thesis have been developed during my research stays, which were allowed by the support of the JAE Predoc and hospitality of Esteban Moro, Fil Menczer, and Alejandro Jaimes, whom I am thankful to for being open to research collaborations with me.

Furthermore, I would like to thank all my collaborators that have contributed to our publications. Apart from the above mentioned scholars, I thank to Luca Maria Aiello, Bruno Goncalves, Luca Chiarandini, Michele Trevisiol, and Josep Puyol for their contributions. I also thank Luis Fernandez Lafuerza, Gideon Zenz, and Dario Taraborelli for their helpful comments and discussions on some parts of the studies.

Finally, I thank all my friends from the institute who made it more enjoyable to perform the studies in the nice environment. Especially, I would like to thank Karin for occasionally interrupting me and insisting on having a tea with me. Also, I would like to thank Luis, Pedro, and Adrian for making noises and having fun while I was working. Naturally, everybody needs to take a break sometimes. I thank all the students staying in the room number 7 of our institute for creating the nice working atmosphere.

Resumen

La presente tesis está dedicada a la descripción, análisis y modelado cuantitativo de sistemas complejos sociales en forma de redes sociales en internet. Mediante el uso de métodos y conceptos provenientes de ciencia de redes, análisis de redes sociales y minería de datos se descubren diferentes patrones estadísticos de los sistemas estudiados. Uno de los objetivos a largo plazo de esta línea de investigación consiste en hacer posible la predicción del comportamiento de sistemas complejos tecnológico-sociales, de un modo similar a la predicción meteorológica, usando inferencia estadística y modelado computacional basado en avances en el conocimiento de los sistemas tecnológico-sociales. A pesar de que el objeto del presente estudio son seres humanos, en lugar de los átomos o moléculas estudiados tradicionalmente en la física estadística, la disponibilidad de grandes bases de datos sobre comportamiento humano hace posible el uso de técnicas y métodos de física estadística. En el presente trabajo se utilizan grandes bases de datos provenientes de redes sociales en internet, se miden patrones estadísticos de comportamiento social, y se desarrollan métodos cuantitativos, modelos y métricas para el estudio de sistemas complejos tecnológico-sociales.

Los grupos juegan un papel fundamental en sistemas sociales, como muestran numerosos estudios sociológicos. Por ello, buena parte de este trabajo se centra el estudio de grupos humanos. La presente tesis contribuye al campo emergente de la ciencia social computacional en los siguientes aspectos:

- Descripción y modelado de la evolución temporal del tamaño de grupos.
- Análisis de patrones estadísticos de interacción de personas dentro y a través de de grupos.
- Desarrollo de métodos de inferencia estadística para la diferenciación de grupos según su tipo.

- Introducción de un modelo de creación de enlaces acoplado a movilidad que da lugar a la formación de redes sociales con propiedades estadísticas geográficas y estructurales realistas.

Las distribuciones de probabilidad de tipo ley de potencias y "cola larga" son características de los sistemas complejos, entre ellos las redes sociales. Usualmente, su existencia se explica como consecuencia del mecanismo "rich-gets-richer" (el que es rico se hace aún más rico) y con modelos basados en crecimiento preferente. Sin embargo, en muchos casos, el crecimiento de los elementos de un sistema dado no está determinado únicamente por este mecanismo, sino que también depende de propiedades intrínsecas de los elementos. La variación de estas propiedades entre los elementos es un origen de heterogeneidad en el sistema. El efecto de la heterogeneidad puede ser más importante que el mecanismo "rich-gets-richer" y sin embargo dar lugar a las mismas distribuciones de tipo ley de potencias. De hecho, en el Capítulo 2 se muestra que las propiedades estadísticas de grupos declarados por los usuarios en la red social Flickr pueden explicarse a partir de un modelo basado únicamente en heterogeneidad.

A continuación, pasamos del estudio de grupos declarados por usuarios al estudio de grupos de individuos detectados mediante métodos basados en teoría de grafos. La identificación de grupos es uno de los intereses centrales en ciencia de redes. En los últimos años se han desarrollado numerosos algoritmos para la detección de comunidades en redes. Una pregunta natural concierne la relevancia de los grupos detectados con dichos métodos. En el Capítulo 3 se muestra, mediante el uso de bases de datos de la red social Twitter, que las interacciones dentro de y entre los grupos detectados dan lugar a propiedades estadísticas no triviales que corresponden a predicciones de la teoría de Granovetter. Esto es, las interacciones de tipo personal suceden considerablemente más a menudo de lo esperado dentro de los grupos, mientras que las interacciones de tipo transmisión de información suceden más a menudo entre grupos. Además, los usuarios que pertenecen a varios grupos actúan como puentes entre ellos, y los enlaces sociales de estos usuarios son usados más frecuentemente para la difusión de información.

Es importante señalar que en las redes sociales en internet los grupos pueden ser identificados de varias formas. Por un lado, los grupos pueden ser creados y declarados explícitamente por los propios usuarios, de modo que la existencia y composición de estos grupos puede inferirse directamente de los datos. Por otro lado, se pueden usar algoritmos de detección de comunidades para identificar los grupos a partir de la estructura de la red de enlaces. En el Capítulo 4, se comparan los conjuntos de grupos obtenidos mediante estos dos métodos y se muestra que la coincidencia entre ellos es mayor que la esperada a partir de una asignación aleatoria.

Además, la comparación es extendida mediante la consideración de la naturaleza y el tipo de los grupos, esto es, si están basados en identidad común (grupos

tópicos) o en enlaces comunes (grupos sociales). En este capítulo también se estudia la manera de clasificar los grupos en estos dos tipos usando una gran base de datos procedente de Flickr. Para ello, se desarrollan nuevas métricas basadas en las teorías de la identidad común y del enlace común y se muestra que predicen el tipo de grupo con gran precisión. Finalmente, se muestra que los grupos detectados son de tipo social más a menudo que los grupos declarados.

La última parte de la tesis se centra en las propiedades espaciales de redes sociales en internet. De hecho, las relaciones sociales y la localización espacial están intrínsecamente entrelazadas, ya que a menudo la gente con la que interaccionamos y mantenemos relaciones se localizan geográficamente cerca de nosotros. En el Capítulo 5, es introducido un modelo que acopla la creación de enlaces sociales y la dinámica espacial de una población. El modelo simula el movimiento de los usuarios y crea enlaces entre ellos cuando se encuentran geográficamente cercanos, imitando las interacciones cara a cara. Las predicciones del modelo son comparadas con grandes bases de datos de las redes sociales Twitter, Brightkite, and Gowalla que incluyen localización espacial. El modelo reproduce varias propiedades estadísticas de la red social y la distancia geográfica entre los usuarios. Varias componentes del modelo son analizadas para identificar los mecanismos más importantes y entender su impacto en la red generada y sus propiedades espaciales. Por ejemplo, se muestra que la tendencia de pares de nodos enlazados a un tercero a estar enlazados entre sí, puede derivarse del hecho de que los nodos coinciden temporal y espacialmente.

Esta tesis está formada por una Introducción, reproducciones de cuatro de mis publicaciones, Epílogo y Apéndice. Los Capítulos 2, 3, 4, and 5 son reproducciones, respectivamente, de las siguientes publicaciones:

- Grabowicz, P. A. and Eguíluz, V. M. (2012). Heterogeneity shapes groups growth in social online communities. Europhysics Lett., 97(2):28002.

- Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., and Eguíluz, V. M. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. PLoS One, 7(1):e29358.

- Grabowicz, P. A., Aiello, L. M., Eguíluz, V. M., and Jaimes, A. (2013a). Distinguishing topical and social groups based on common identity and bond theory. In Proceedings of The Sixth ACM International Conference on Web Search and Data Mining - WSDM '13, page 627, New York, New York, USA. ACM.

- Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., and Eguíluz, V. M. (2013b). Entangling mobility and interactions in social media. Submitted, preprint: arXiv:1307.5304.

Preface

This thesis is devoted to quantitative description, analysis, and modeling of complex social systems in the form of online social networks. Statistical patterns of the systems under study are unveiled and interpreted using concepts and methods of network science, social network analysis, and data mining. A long-term promise of this research is that predicting the behavior of complex techno-social systems will be possible in a way similar to contemporary weather forecasting, using statistical inference and computational modeling based on the advancements in understanding and knowledge of techno-social systems. Although the subject of this study are humans, as opposed to atoms or molecules in statistical physics, the availability of extremely large datasets on human behavior permits the use of tools and techniques of statistical physics. This dissertation deals with large datasets from online social networks, measures statistical patterns of social behavior, and develops quantitative methods, models, and metrics for complex techno-social systems.

Groups play a fundamental role in social systems, as shown by numerous sociological studies. Thus, a good part of this dissertation focuses on groups. The thesis contributes to the emerging field of computational social science in the following respects:

- It describes and models the temporal evolution of group sizes;
- It analyzes the statistical patterns of interactions of people in the landscape of groups;
- It develops methods of statistical inference for distinguishing types of groups;
- It introduces a model of coupled mobility and link formation that produces social networks with realistic geographic and structural statistical properties.

Power-law and heavy-tailed distributions are ubiquitous in complex systems, including social networks. Typically, they are explained with the rich-gets-richer rule and models based on preferential growth. However, in many cases, the growth of elements of the given system is not only driven by this mechanism, but it also depends on the intrinsic quality of the elements. This property of the elements, also known as intrinsic fitness, is a source of heterogeneity. The impact of heterogeneity may in fact prevail over the rich-gets-richer phenomenon and nevertheless drive the system to similar power-law distributions. In fact, in Chapter 2 we show that statistical properties of user-declared groups in Flickr can be explained with a model based solely on heterogeneity.

Next, we move the focus of the thesis away from user-declared groups to groups of people detected with graph-based methods. The detection of groups is one of the focal interests in network science. Numerous community detection algorithms have been developed in recent years. A natural question is what is the importance of groups found with such methods. In Chapter 3, using a dataset from Twitter, we show that user interactions within the detected groups and between them yield non-trivial statistical features that correspond to predictions of the Granovetter's theory. Namely, personal interactions happen considerably more often than expected inside groups, and information transmission interactions happen more often between groups. Moreover, users who belong to several groups act as bridges between them, and social links of such users are more frequently used for the diffusion of information.

Note that in online social networks groups can be identified in a few ways. On the one hand, groups can be created and declared explicitly by the users themselves, and subsequently directly retrieved from the data. On the other hand, community detection algorithms can be used to identify them from the network structure. In Chapter 4, we directly compare the two sets of groups showing that indeed the overlap between the two is higher than expected by random chance. Furthermore, we extend the comparison by considering the nature and type of groups, i.e., whether they are based on common identity (topical) or on common bond (social). We investigate how to classify groups into these two types using a large dataset from Flickr. We introduce metrics based on the common identity and common bond theories and show that they predict the group type with high accuracy. Finally, we show that the detected groups are more often social than the declared groups.

In the last part of the thesis, we switch the focus from groups to spatial properties of online social networks. In fact, social relationships and physical location are inextricably entangled. The people we interact and maintain relations with are often those that stay close to us geographically. In Chapter 5, we introduce a model that couples social link creation and the spatial dynamics of a population. The model simulates the movements of users and creates links when they are

physically close to each other by imitating face-to-face interactions. The model is tested against large geo-localized data from Twitter, Brightkite, and Gowalla. It reproduces several statistical properties of the social network and the physical distance between people. We investigate different components of the model to identify its most important ingredients and to understand their impact on the generated network and its geography. For instance, we show that triadic closure can be achieved by means of spatio-temporal co-occurrences with friends.

This dissertation consist of Introduction, reproductions of four of my publications, Outlook and Appendix. Chapters 2, 3, 4, and 5 reproduce the following publications, respectively:

- Grabowicz, P. A. and Eguíluz, V. M. (2012). Heterogeneity shapes groups growth in social online communities. Europhysics Lett., 97(2):28002.
- Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., and Eguíluz, V. M. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. PLoS One, 7(1):e29358.
- Grabowicz, P. A., Aiello, L. M., Eguíluz, V. M., and Jaimes, A. (2013a). Distinguishing topical and social groups based on common identity and bond theory. In Proceedings of The Sixth ACM International Conference on Web Search and Data Mining - WSDM '13, page 627, New York, New York, USA. ACM.
- Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., and Eguíluz, V. M. (2013b). Entangling mobility and interactions in social media. Submitted, preprint: arXiv:1307.5304.

Contents

Chapter 1

Introduction

If we knew what it was we were doing, it would not be called research, would it?

— Albert Einstein

Historically, social systems have been studied in sociology, frequently using results of self-reported surveys conducted on small samples of population. Nowadays, information about large fraction of population is gathered unobtrusively due to trances left by the users of online services and mobile devices, allowing the quantitative revision and advancement of sociological theories. In this thesis, we develop methods, models and metrics that contribute to the emerging field of computational social science. In this chapter, the foundations of this field are explored including complex networks, social network analysis, and data mining. We describe concepts, definitions, and methods of each of the fields related to this dissertation.

First, we introduce the mathematical framework of network theory used in our studies. Definitions of graph-related quantities are provided and their values in real social networks are given.[1] Additionally, topological and spatial properties of social networks are described. Later, we will introduce a model that reproduces these properties. Next, we introduce models of growth of networks, random models of networks, and the problem of community detection.[2] Among the growth models, preferential growth and heterogeneity models are reported. We will compare the two families of models in the study of groups' growth. Also, a special emphasis is put on community detection, as we will present a couple of studies of groups found with such methods.

[1] Word "network" is used in this thesis interchangeably with word "graph".

[2] Word "group" is used in this thesis interchangeably with word "community". Words "cluster" and "module" are related but used only in the context of networks.

Then, we introduce sociological theories related to social networks and groups, which are exploited in order to pose research questions and to interpret results of the forthcoming analyses. Basic mechanisms of tie formation are listed; the strength of a tie is defined; the relation between network structure, tie strength, and information diffusion is introduced. We will measure the strength of ties in the study of interaction patterns in groups found by community detection algorithms. Also, two types of groups based on the theories of common identity and common bond are described. Based on these theories, we will propose metrics that allow the characterization and the prediction of group types.

Finally, in the last section of this chapter we characterize and review structural properties of three established OSNs. Two of these networks (i.e., Flickr and Twitter) will be analzed in our studies. We list common features of these OSNs and show how they can be abstracted to declared social links and different types of pairwise interactions corresponding to, e.g., personal communications and information diffusion. We will use this abstraction in the study of interaction patterns in the landscape of groups. Furthermore, in OSNs, users can create and declare groups on their own. These two sets of groups are introduced to the reader. Then, we will describe and model their growth, and compare them with groups detected by a clustering algorithm.[3]

Groups constitute the common topic of this dissertation. We will present the results of our studies starting from the description of the growth of declared groups in Flickr (Chapter 2). Then, we will extend the scope of the studies by focusing on groups detected with graph-based clustering algorithms (Chapter 3). We will show that such groups are correlated with statistical patterns of user interactions in Twitter. Next, the two sets of groups, i.e., the declared and detected groups in Flickr, will be directly compared in terms of their membership composition (Chapter 4). We will introduce a method of statistical inference to find if a given group is topical or social and apply this method to the two sets of groups. The dissertation concludes with a discussion of our contributions to computational social science and an outlook for the future (Chapter 6).

1.1 Computational social science

In recent years, a new scientific discipline has emerged dubbed *computational social science* (Lazer et al., 2009; Watts, 2007; Miller, 2011; Giles, 2012; Conte et al., 2012). It is a field that connects several other disciplines, namely mathematics and physics (through the fields of statistics, graph theory, statistical physics,

[3] "Community detection algorithms" are also known as graph-based "clustering algorithms". In this thesis, these terms are used interchangeably.

complex networks, and through computational modeling), sociology (through the field of social network analysis), and computer science (through the fields of data mining and machine learning). The term "computational social science" was coined by social network analysts and network scientists. However, the field is known as *social computing* in computer science. Nowadays, a part of the research in the field of social computing is happening at Internet and phone companies such as Facebook, Yahoo, Microsoft, and Telefónica, but public research in this field increases with data availability. Computational social science is the center of the focus of this dissertation.

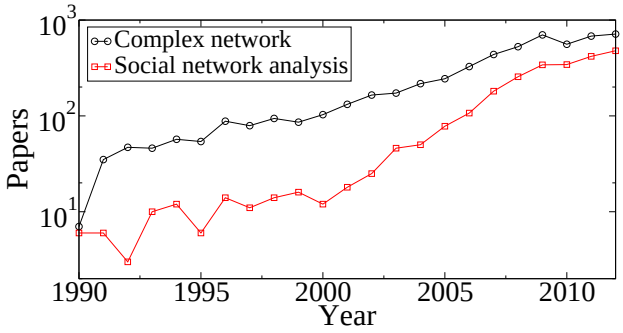


Figure 1.1: Number of papers found by searching for the topics "complex network" and "social network analysis". Retrieved in September 2013 from `http://apps.webofknowledge.com`.

Almost contemporarily, in the last fifteen years, the field of network science, also known as complex networks, has experienced a period of exponential growth (black curve in Figure 1.1). The evolution of this field is coupled with the growth of social network analysis, which also has experienced a decade of rapid growth (red curve in Figure 1.1). The two fields began to be used as fuel for new computational models based on complex networks (Macy and Willer, 2002; Epstein, 2006; Buchanan, 2009; Vespignani, 2009; Schweitzer et al., 2009).

One of the reasons why network science has emerged so rapidly is the increasing availability of data from growing technological networks, particularly from the Internet. In fact, the two seminal and most cited papers of network theory (Barabási and Albert, 1999; Watts and Strogatz, 1998) used examples of networks that either represented technological systems (the power grid) or were created in a collaborative effort mediated through the Internet (the actor collaboration graph, the World Wide Web).[4] Along with the growth of technological

[4] The only network used in one of the two papers that is not related to technology is a neural network.

networks, the amount of data stored online started to grow rapidly, causing the development of the data mining field. Currently, the amount of data stored online is estimated to be around 2.7 billion terabytes,[5] almost half of which is generated by users (Gantz and Reinsel, 2012). Among the services provided online, OSNs have gained extremely high popularity.[6] Nowadays, the biggest OSNs have up to a billion active users[7], and for instance the penetration in Spain reaches 42% of the population (82% of young adults) (Borondo et al., 2012). Suddenly, the amount of data on social networks started to reach enormous levels that were impossible to obtain with traditional user studies. The datasets of this size have become so important that it was given its own name, big data,[8] and caused shifts at the economic, technological, and scientific levels. The number of things that scientists can learn from big data in social networks is very promising, e.g., how different information diffuses, what is the mechanism of social influence, how social conventions are formed, or how to detect bias of opinions.

The mutual influence of network science, social network analysis, data mining, computational modeling, and the paradigm of big data has caused the rise of computational social science. In what follows, we dissect the three main components that triggered the emergence of computational social science, i.e., network science, social network analysis, and data mining.

1.1.1 Complex networks

The exponential growth of network science started with the introduction of the small-world network model (Watts and Strogatz, 1998) and the description and modeling of scale-free networks (Barabási et al., 1999; Barabási and Albert, 1999). The field has been commenced and advanced mainly by physicists, mathematicians, and sociologists; but also by biologists, economists, and computer scientists. The growth of the field led to the development and the collection of statistical measures and methods for real-world networks, which are nowadays at the very core of network science and used in numerous other disciplines. The metrics and methods include degree distribution (Barabási et al., 1999; Barabási and Albert, 1999), modularity (Newman and Girvan, 2004; Newman, 2006), assortativity (Newman, 2002, 2003a), centrality measures (Newman, 2010), and community detection algorithms (Fortunato, 2010). These metrics and methods

[5] Given that there are around 2.4 billion Internet users worldwide, this means that there is over 1 terabyte stored online per each Internet user.

[6] Five out of the top ten most popular websites are OSNs or related sites, according to the Alexa ranking from September 2013. For a recent ranking, visit `http://www.alexa.com/topsites`.

[7] Facebook in its second quarter 2013 financial report declared 1.15 billion monthly active users and growing. See more at `http://bit.ly/1bifDuL`.

[8] In general, big data refers to data that is difficult to manage without a distributed system. Big data includes datasets from large scientific experiments and simulations, e.g., from the Large Hadron Collider and the NASA Center for Climate Simulation.

have started to be widely used in studies of economics (Jackson, 2010), biology (Maslov and Sneppen, 2002; Guimerà and Amaral, 2005), neuroscience (Eguíluz et al., 2005; Bullmore and Sporns, 2009), and ecology (Dunne et al., 2002). Complex networks are used in computational modeling, e.g., of the spread of epidemics (Pastor-Satorras and Vespignani, 2001; Hufnagel et al., 2004; Balcan et al., 2009) and influence dynamics (Kempe et al., 2003; Klemm et al., 2012). Furthermore, the community structure has been renowned as one of the key characteristics of real networks (Newman, 2010), and a whole family of graph-based clustering algorithms has been developed to detect dense modules of nodes in networks (Fortunato, 2010). Groups play a particularly crucial role in social networks (Granovetter, 1973; White and Harary, 2001).

Other focal interests in network science are currently under intensive development, e.g., time-varying networks (Holme and Saramäki, 2012), multiplex networks (Szell et al., 2010; Mucha et al., 2010; Gómez-Gardeñes et al., 2012; Gómez et al., 2013), and dynamical processes on complex networks (Barrat et al., 2008). These research lines are motivated by the characteristics of real networks. Namely, real networks vary in time, e.g., phone calls are temporal; different types of relations are present, which can be represented by multiplex networks, e.g., users of phones can either call or text each other; several dynamic processes happen on the network structure, e.g., information diffuses through telecommunications.

1.1.2 Social network analysis

The idea of treating human beings as social atoms of a larger system dates back to the first part of the nineteenth century. Several thinkers (e.g., Comte and Durkheim) argued that social systems can be modeled as physical ones, that human communities are like biological systems in that sense, and that they are made of interrelated elements. Comte hoped to found a new field of "social physics" (Borgatti et al., 2009). As a matter of fact, in the early twentieth century, Moreno developed sociometry, a quantitative method for the evaluation of an individual's role in a community through analysis of the network of relations between the members (Moreno, 1934). Since then, social systems have often been described by their network representations (Granovetter, 1973; Freeman, 1978; Coleman, 1988; Uzzi, 1996; Freeman, 2004; Butts, 2009). Nowadays, the field of social network analysis investigates the structure, interactions, attributes, and events and their outcomes in the social networks (Borgatti et al., 2009). The spiritual successor of the idea of social physics is so-called sociophysics (Cho, 2009), which is focused on developing agent-based models of various social processes (Macy and Willer, 2002; Bonabeau, 2002).

One of the main questions tackled in social network analysis is the very origin

of social ties. There exist various families of theories about why people create, maintain, and dissolve network ties. Among them, one distinguishes (Katz et al., 2004) theories of self-interest, social exchange, mutual interest, homophily, and cognitive theories. Each of these theories corresponds to a different school. The *self-interest* paradigm assumes that people form ties in a rational process that maximizes their personal preferences (Coleman, 1988). The *social exchange* theory considers that people are interdependent and that by creating social ties, they try to minimize dependence on resources from others and to maximize dependence of the others on the resources that they can provide (Emerson, 1976). The theory of *public goods* assumes that mutual profits of connected people outweigh self-interests of its members (Samuelson, 1954). The *homophily* hypothesis suggests that people who are similar form social ties and derives from the phenomenon widely known as *birds of a feather flock together* (Byrne, 1971; McPherson et al., 2001). Finally, the cognitive theories consider how an ego perceives others as a factor influencing tie formation, e.g., in case other person knows something that ego does not know (the theory of *transactive memory* introduced in Wegner (1987)) or she has positive relations with a friend of ago (the *balance theory* introduced in Heider (1958)).

Interestingly, ties connect people tightly in a social network in the sense that the path in the network between any two people who do not know each other is very short. This concept is known as six degrees of separation (Milgram, 1967). For instance, in a well-known experiment conducted on a group of individuals in the United States (Travers and Milgram, 1969), any person could be reached by any other person in the network by passing, in most cases, through fewer than six different people. One of the follow-ups to these findings was the introduction of the aforementioned small-world networks (Watts and Strogatz, 1998).

A family of studies considers the relation between the position of a person in the social network and the benefits to that person. The seminal theory of the strength of weak ties analyzes the relation between the strength of ties, their structural position, and the potential to diffuse information (Granovetter, 1973), showing that weak ties are important for spread of novel information. Another study suggests that the rise to power of the Medici family in fifteenth-century Florence is related to high betweenness centrality in the political, economic, and marriage networks of the family members (Padgett and Ansell, 1993). The theory of social capital generalizes such findings by suggesting that the structural position of a person in the network helps in getting better job offers, obtaining faster promotions, and learning about innovations in less time (Burt, 2005). Furthermore, several recent studies have investigated simple and complex contagions that describe the mechanisms of diffusion and social influence (Christakis and Fowler, 2009; Bond et al., 2012; Ugander et al., 2012; Centola and Macy, 2007; Centola et al., 2007; Centola, 2010).

Nowadays, network studies are among the most cited in sociology (Rivera et al., 2010). Given new sources of digital data the studies influence other fields as well, such as computer (Bakshy et al., 2011; Ugander et al., 2012) and political sciences (Lazer, 2011; Bond et al., 2012).

1.1.3 Data mining

In recent years, a large amount of information on human behavior is generated unobtrusively whenever people interact through modern technologies such as online services, cell phones, and mobile applications. The advent of big data in social media has opened the gates to the analysis of massive datasets on several aspects of society, e.g., information diffusion (Bakshy et al., 2012), political polarization (Conover et al., 2012), voter turnout during elections (Bond et al., 2012), and human mobility (Song et al., 2010a). It has made possible the pursuit of a computational approach to the study of problems traditionally associated with social sciences (Lazer et al., 2009; Watts, 2007; Miller, 2011; Giles, 2012). Not only it allows quantitative approaches toward traditionally qualitative theories but also enables researchers to have more precise and daring research questions and problems.

Take as an example the theory of the strength of weak ties (Granovetter, 1973, 1983). It is one of the most cited studies in sociology and has been tested by sociologists. Surveyed people were asked to identify the source of the information that led them to find their current jobs. The final result showed that people most often learn about job openings from their weak ties. These studies, however, did not take into account that the numbers of strong and weak ties of a given person differ, e.g., the number of weak ties tends to be larger than the number of strong ties. Thus, weak ties may transmit information with lower frequency than strong ties and still be more important for information diffusion due to their sheer volume (Bakshy et al., 2012). Namely, in such case, the overall amount of weak ties in the system prevails over the lower frequency of information transmission per one tie. The datasets and experiments in OSNs not only allow unobtrusive user studies but also enable calculation of information diffusion probability as a function of the number of influencers and the tie strength (Bakshy et al., 2012).

Over the last few years, big data has allowed the development of greater insights, for instance, into human mobility (Brockmann et al., 2006; González et al., 2008; Song et al., 2010a), structure of OSNs (Kwak et al., 2010; Mislove et al., 2008), human cognitive limitations (Miritello et al., 2013; Gonçalves et al., 2011), information diffusion and social contagion (Bakshy et al., 2012; Ugander et al., 2012; Leskovec et al., 2009; Lehmann et al., 2012), the importance of social groups (Grabowicz et al., 2012, 2013a; Ferrara, 2012), and how political movements emerge and develop (Borge-Holthoefer et al., 2011; Conover et al.,

2012). Such empirical findings build the skeleton of computational social science and lay the foundations for more realistic computational modeling.

1.2 Network theory and properties of social networks

In this section, we define basic concepts and properties of network science. At the end of each of the subsections, we characterize real networks in terms of these properties, focusing especially on social networks. Social networks show a rich internal structure, far from random graphs (Newman and Park, 2003). It is a broad category of networks that represent a particular relation or interaction between people, e.g., co-appearance in movies, participation in boards of directors, or co-authorship (Newman and Park, 2003; Newman, 2002, 2003b); phone calls and communications (Onnela et al., 2007a; Palla et al., 2007; Leskovec and Horvitz, 2008); or online friendship (Mislove et al., 2007; Ahn et al., 2007; Leskovec et al., 2008; Ugander et al., 2011). A detailed description of the structural properties of large OSNs will be provided in Subsection 1.6.2, after the introduction of these online services.

1.2.1 Basic concepts and definitions

A graph G is the basic entity of graph theory. It consists of a set of vertices $V(G)$ and a set of edges $E(G)$ that connect the vertices. Each edge is a pair of vertices from the set $V(G)$. A network that does not have any edge is called an *empty graph*. A graph in which every vertex is connected with every other vertex is referred to as a *complete graph*. Finally, a network that consists of the subset of vertices $V(G)$ and edges between them is called a *subgraph* of G. In sociology, an *ego network* represents a subgraph created from an individual, i.e., the *ego*, other individuals related to her, i.e. the *alters*, and the relations between them.

A graph can be mathematically represented as an adjacency matrix $A = [a_{ij}]$, where $i \in \{1, 2, ..., N\}$ and N is the number of vertices in the graph. The elements a_{ij} of the matrix define the existence or absence of an edge between the two vertices i and j, thus taking value of 1 or 0, respectively. A graph can contain self-loops, that is, edges that have the beginning and the end in the same node, thus $a_{ii} \neq 0$. In this dissertation, we only consider networks without self-loops. The edges can be directed, meaning that an edge from vertex i to j is different from an edge from j to i. In other words, in undirected networks $a_{ij} = a_{ji}$, while in directed networks a_{ij} can be different from a_{ji}. A directed graph is also called a *digraph*. Next, each edge of a graph can have a weight attached to it, represented by $a_{ij} \in \mathcal{R}$ instead of a binary value. Finally, a graph can have

multiple edges per each pair of nodes.[9] In such case, $E(G)$ is a multiset, and the graph is called a *multigraph*, or a *multidigraph* if it is directed. In this thesis, we study undirected, directed, and multidigraphs.

The vertices and edges have alternative names, depending on the context in which they appear. In computer science, they are often called nodes and links; in sociology, actors and ties. In agent-based modeling, they usually represent agents and interactions between them; while in OSNs, users and pairwise interactions or declared relations between them.

1.2.2 Degree distribution and link directionality

Properties of a graph can be directly derived from the adjacency matrix. We introduce them in this and the following subsections. For example, the total number of links L in a graph is calculated as

$$L_{\mathrm{u}} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} a_{ij}, \tag{1.1}$$

while in a directed graph

$$L_{\mathrm{d}} = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}. \tag{1.2}$$

In this thesis we will use the term L to refer to either undirected or directed edges, depending on the context. The degree k_i of a node i is the number of edges connected to that node. In a simple graph it is defined as

$$k_i = \sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ji}, \tag{1.3}$$

while in a directed graph we have two types of degrees, out-degree k^{out} and in-degree k^{in}:

$$k_i^{\mathrm{out}} = \sum_{j=1}^{N} a_{ij}, \qquad k_i^{\mathrm{in}} = \sum_{j=1}^{N} a_{ji}. \tag{1.4}$$

Average node degrees are, respectively, $k = \langle k_i \rangle = \frac{2L_{\mathrm{u}}}{N}$, $k^{\mathrm{out}} = k^{\mathrm{in}} = \langle k_i^{\mathrm{out}} \rangle = \langle k_i^{\mathrm{in}} \rangle = \frac{L_{\mathrm{d}}}{N}$.

In the case of directed networks, the average reciprocity of edges in the network is defined as the ratio of the number of edges in both directions divided by

[9] In graph theory, an undirected graph without self-loops and without multiple edges per each pair of vertices is called a *simple graph*.

the total number of edges

$$R = \frac{L^{\text{rec}}}{L^{\text{rec}} + 2L^{\text{nrec}}}, \tag{1.5}$$

where L^{rec} corresponds to the edges in two directions (reciprocated), while L^{nrec} in only one direction (non-reciprocated). Note that in a multidigraph multiple reciprocated and non-reciprocated edges are possible for a pair of vertices. A directed network is easily converted into its undirected counterpart by replacing all reciprocated and non-reciprocated edges with undirected edges. This procedure is called symmetrization. The total number of directed edges is $L_{\text{d}} = L^{\text{rec}} + L^{\text{nrec}}$, while the total number of undirected edges after symmetrization is $L_{\text{u}} = L^{\text{rec}}/2 + L^{\text{nrec}}$. In such case, the following relation takes place

$$k = \frac{1}{1+R}(k^{\text{out}} + k^{\text{in}}), \tag{1.6}$$

where k corresponds to average degree of the converted undirected network.

We denote the distribution of degrees as $P(k)$. The first moment of $P(k)$ is the average degree. The complementary cumulative distribution function is defined as

$$\text{CCDF}(k) = P(k^* \geq k) = 1 - P(k^* < k) = 1 - \sum_{k^*=k_{\min}}^{k-1} P(k^*), \tag{1.7}$$

where $k_{\min}$ corresponds to the minimum degree. Most real networks consist of many nodes with small degrees and few nodes with extremely high degrees that are called *hubs*. In this context, the kind of distribution that decays slowly with the degree, slower than an exponentially decreasing function, is called *heavy-tailed* distribution (Newman, 2010; Boccaletti et al., 2006; Albert and Barabási, 2002; Dorogovtsev and Mendes, 2003). The tail of the distribution corresponds to the few nodes that have extremely large degree. One of the most studied distributions of this kind is the power-law distribution

$$P(k) = \frac{\alpha - 1}{k_{\min}} \left(\frac{k}{k_{\min}} \right)^{-\alpha}, \tag{1.8}$$

where α corresponds to the exponent of the power-law. The CCDF holds the same power-law functional form with its exponent equal to $\alpha - 1$. The power-law distributions are straight lines if plotted in log-log scale. In real systems, often just the tails of distributions are fitted with a power-law. Examples of heavy-tailed and power-law degree distributions in real networks are shown in Figure 1.2. Note that all moments $m \geq \alpha - 1$ of the power-law distribution diverge, e.g., when $2 < \alpha < 3$ then the mean exists, but the variance and the higher-order moments

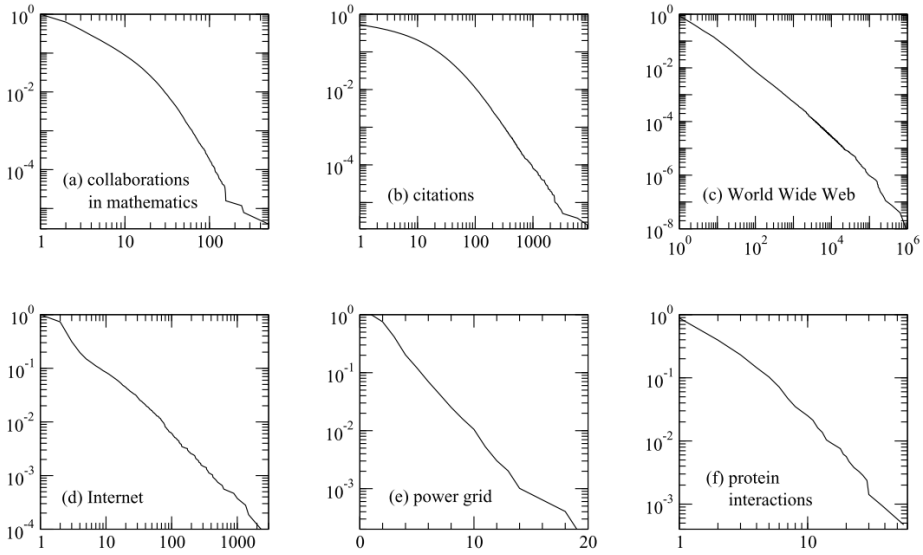


Figure 1.2: Examples of: (a,b) heavy-tailed, (c,d,f) power-law, and (e) exponential complementary cumulative degree distributions in real networks. The x-axis is degree k and the y-axis is CCDF(k). Note that all the figures are plotted in a log-log scale, apart from (e), which is plotted in the log-linear scale. The networks shown are: (a) the collaboration network of mathematicians; (b) citations to scientific papers; (c) a large subset of the World Wide Web; (d) the Internet at the level of autonomous systems; (e) the power grid of the western United States; (f) the interaction network of proteins in the metabolism of the yeast S. Cerevisiae. Adapted from (Newman, 2003b).

diverge. Networks with power-law distribution do not have a scale, meaning that the distribution does not change its form under degree transformation consisting of multiplication by a common factor (Barabási and Albert, 1999).

The degree distribution in social networks tends to be broad and usually has a heavy-tail decaying as a power-law or a log-normal function with a cutoff at some value of the number of friends (Newman, 2003b; Boccaletti et al., 2006; Mislove et al., 2007; Ahn et al., 2007; Leskovec and Horvitz, 2008), with the exception of an exponentially decaying distribution in a network of face-to-face proximity (Isella et al., 2010). The exponent of the power-law distributions ranges from around 2 in OSNs with directed links (Kumar et al., 2006; Kwak et al., 2010) to 5, or even 8.4 in the tail, in a network of mobile phone calls (Lambiotte et al., 2008; Onnela et al., 2007a). In and out-degrees are correlated and have similar

distributions in social networks (Ahn et al., 2007; Mislove et al., 2007). However, in general they differ (Kwak et al., 2010).

1.2.3 Triangles and clustering

The clustering coefficient measures the probability that two vertices sharing a common neighbor (a vertex to which both nodes are linked) are connected. This property can be quantified as the global clustering coefficient

$$C = \frac{\Delta}{\Lambda}, \tag{1.9}$$

where Λ is the number of all triads in the network and Δ is the number of closed triads. A *triad* is a sequence of 3 nodes i, j, k such that the central node j is connected to both extreme nodes i and k. A closed triad is a triad that has also an edge between i and k. Note that in directed networks the order in the sequence matters, e.g., while a triad i, j, k is closed the triad k, j, i can remain open, if there is an edge from i to k but no otherwise. This definition of the clustering coefficient is valid for both undirected and directed networks. The local clustering coefficient c_j of a node j is defined as

$$c_j = \frac{\Delta_j}{\Lambda_j}, \tag{1.10}$$

where Λ_j and Δ_j are the corresponding numbers of triads centered on the node j. In this case, a global value of the clustering coefficient may be obtained averaging the local c_j over all the nodes of the network, although one should note that it is different from the global clustering coefficient C. Usually, the local clustering coefficient decreases with degree in social networks (Ugander et al., 2011). Because of this, the two coefficients differ, particularly so in networks with heavy-tailed distributions. In such cases, the value of the global clustering coefficient is dominated by the high-degree nodes, and the local clustering coefficient is the preferred choice.

One of the most important feature distinguishing social networks from other types of networks is their high level of clustering, also known as transitivity (Newman, 2003b; Newman and Park, 2003; Watts and Strogatz, 1998; Mislove et al., 2007; Ahn et al., 2007; Leskovec and Horvitz, 2008; Ugander et al., 2011). At the structural level, a high clustering coefficient indicates the presence of many triangles in the network. At the social or personal level, this means that friends of an individual tend to be connected between themselves too, i.e., friends of our friends tend to be our friends too. The process that leads to creation of such structure in social networks is called *triadic closure* (it is described in Subsection 1.5.1). Several network models are based on the assumption of triangle closure, some of which are discussed in the latter part of the thesis (Subsection 1.4.4).

1.2.4 Average shortest path

If in a given subgraph there exists a path between every pair of vertices, then this subgraph is called a *connected component*. The component that contains the highest number of vertices is called the largest connected component. In directed networks we distinguish between *weakly and strongly connected components*. The former is defined as a set of nodes to which a network crawler can arrive while crawling the component. The latter, i.e., strongly connected component, is defined as a set of nodes to which a network crawler can arrive and from which the crawler can reach all other nodes in the component as well.

An important property characterizing the structure of complex networks is the shortest path between nodes. A *path* between two nodes i and j is a sequence of vertices $P_{ij} = (i, ..., j)$ such that each of the vertices in the sequence is connected to the vertices that appear after it and before it. The length of the path is defined as the number the nodes in the sequence minus one. The *shortest path* is the path that has the least number of vertices in the sequence. The average shortest path is the average shortest path length between any two nodes of a connected component.

Almost all users of OSNs are connected in the largest connected component (Leskovec and Horvitz, 2008; Ugander et al., 2011). Some of the studies also point out that OSNs contain a densely connected core (Mislove et al., 2007; Leskovec and Horvitz, 2008) consisting in clusters of high-degree nodes that hold the network together. Such cores provide paths for connecting distinct parts of the network and reduce the average shortest path. The importance of the shortcuts for decreasing the network path-length was highlighted in (Watts and Strogatz, 1998).

Low average shortest path in comparison to the number of nodes is a characteristic property of most of real networks (Newman, 2010; Boccaletti et al., 2006; Albert and Barabási, 2002). This property has been of special interest in sociology and it is popularly known as the small-world effect or the concept of six degrees of separation, which suggests that the distance between any two people in a social network is on average very short (it is close to six or smaller in the experimental study of Travers and Milgram, 1969). This characteristic is also present in OSNs (Mislove et al., 2007; Ahn et al., 2007; Leskovec and Horvitz, 2008; Ugander et al., 2011).

1.2.5 Community structure and modularity

At the macroscopic level, a high density of triangles is related to the existence of community structure in the network. In networks with high value of clustering coefficient community structure emerges without any additional ingredients included ((Foster et al., 2011)). The community structure means that the network

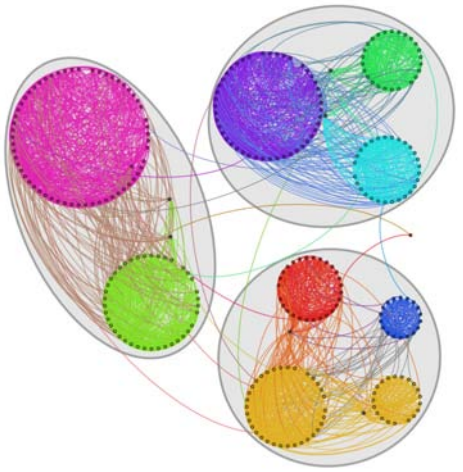


Figure 1.3: An example of an hierarchical community structure. Each cluster is depicted with different color. Adapted from (Lancichinetti et al., 2011).

is built of modules. The modules are often called *clusters* and are sets of nodes that are well connected internally (intra-connected), while being just loosely connected between each other (inter-connected). An example of a network with a clear community structure is illustrated in Figure 1.3.

Naturally, some networks have less and others have more modular structure. The most popular method of quantifying this property is the *modularity* Q of an undirected network that is partitioned in g groups (Newman and Girvan, 2004; Newman, 2006):

$$Q = \sum_{g=1}^{N_{\text{groups}}} \left[\frac{L_g^{\text{int}}}{2L} - \left(\frac{K_g}{2L} \right)^2 \right], \tag{1.11}$$

where the L_g^{int} is the number of internal links of the group g, while K_g is the total degree of nodes in that group, and L is the total number of links. The modularity compares the number of internal links with the expected number of internal links by subtracting the two numbers. The expected number is the average number of internal links in random graphs with fixed degree K_g per each group. The modularity takes values from -1 to 1. It is supposed to be positive if the network has an underlying community structure and if the partition captures it. If the network is not modular or if the partition does not capture it, then the modularity is close to zero. Finally, the modularity takes a negative value if the community structure is present in the network, but the network is parti-

tioned into groups that are anti-correlated with the community structure, i.e., the connections are more abundant between groups than inside groups, given the expectation in a random graph as a reference point. The problem of detection of communities in networks is one of the focal interests in complex networks. The methods that partition networks into clusters are known as community detection algorithms or clustering algorithms. Some of the first methods of this kind were based on modularity optimization (Blondel et al., 2008; Newman and Girvan, 2004; Newman, 2006). We elaborate on modularity optimization and the problem of community detection in the next section, and we analyze results of several clustering algorithms for various OSNs in the following chapters of this dissertation.

It was suggested that in biological networks modularity has evolved to reduce connection costs (Clune et al., 2013). Such costs include creating and maintaining links, and transmitting along them; it increases as a function of connection length, e.g., in neural networks or genetic and metabolic pathways. In sociology the modules are called *communities* or *groups*. Existence of communities in social networks is considered by sociologist to have fundamental relevance (Granovetter, 1973; Burt, 2005), e.g., for the diffusion of information. In the following parts of this dissertation, we show that groups correlate with the activity of the users of an OSN.

1.2.6 Assortativity, homophily and similarity measures

Nodes within a network that are one hop away from a central node are called the *neighbors* of that node. The second neighbors are two hops away from the node etc. Often the neighboring nodes are related to each other, e.g., share certain attributes or properties in common, are connected to the same groups of nodes, or have related functions. Here, we introduce a set of measures that capture correlations and similarities between connected nodes.

From the point of view of network topology the similarity between nodes may be expressed as the correlation between the degrees of neighboring nodes, which is known as *assortative mixing*, or as the *rich-club effect*. In assortative networks nodes with high degree tend to be connected to other nodes with high degree, and low-degree nodes tend to be connected to other low-degree nodes. In disassortative networks it is the opposite. The assortativity coefficient r is defined as a Pearson correlation coefficient of node degrees

$$r = \frac{\sum_{ij} (a_{ij} - k_i k_j / 2L) k_i k_j}{\sum_{ij} (k_i \delta_{ij} - k_i k_j / 2L) k_i k_j}, \tag{1.12}$$

where δ_{ij} is the Kronecker delta, which is 1 if $i = j$ and 0 otherwise. If the assortativity coefficient is positive, then the network is assortative, if negative,

then the network is disassortative.

To measure the similarity between non-scalar properties of nodes, we need to use a different method than the Pearson correlation coefficient. Here, we present two other methods of measuring the similarity: the Jaccard coefficient for a pair of sets and the cosine similarity of a pair of vectors.

The Jaccard coefficient measures the similarity between two sets A and B, and is defined as the cardinality of the intersection divided by the cardinality of the union of the sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1.13}$$

This coefficient takes values from 0 to 1. It is close to 0 when the two sets do not share almost any elements in common, and it is close to 1 when they share most of their elements.

The cosine similarity measures the similarity between two vectors $\vec{A}$ and $\vec{B}$, and it is defined as a dot product of the vectors divided by product of their magnitudes

$$S_{\mathrm{C}}(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\|\vec{B}\|}. \tag{1.14}$$

In general, the cosine similarity is used to measure the similarity between various attributes of nodes, which often are represented as a vector.

Social networks are assortative (Newman, 2002; Newman and Park, 2003), in contrast to networks of other types, e.g., information, technological, or biological networks (Newman, 2003b). Naturally, the node degree is not the only property in which neighbors are similar. People who are connected in OSNs tend to have similar age, live in close locations, and have similar interests (Palla et al., 2007; Ugander et al., 2011; Leskovec and Horvitz, 2008; Schifanella et al., 2010). It is also predicted that people who belong to the same community, namely the same well-connected group of people, talk about similar topics, which has an important impact on information and innovation diffusion in social networks (Granovetter, 1973). In general, the similarity between characteristics of connected individuals is called homophily and it is present broadly in social networks. It is popularly known as *birds of a feather flock together* phenomenon (McPherson et al., 2001).

1.2.7 Spatial properties of networks

Real networks often represent systems of elements that exist in a physical space. In such cases, one can add an attribute to each node that localizes the nodes in physical space. In most practical cases, the space is a two-dimensional space with the Euclidean distance. In general, we call a graph a *spatial network* if its nodes are located in a space equipped with a metric. In spatial networks there is usually a cost associated with the length of edges, which has strong effects on the

topological structure of such networks, e.g., wiring costs determine placement of neurons in animals (Chen et al., 2006; Rivera-Alba et al., 2011) leading to highly modular networks (Clune et al., 2013). Several studies explore the mechanisms of spatial link creation and try to identify the cost functions for various spatial networks (for a recent review, see Barthélemy, 2011).

Physical space plays also an important role in social networks. People tend to maintain relations and interact with geographically close peers (Liben-Nowell et al., 2005). This is reflected by the decay of the linking probability $P_l(d)$ with physical distance, which is measured as the number of links $L(d)$ at a given distance d divided by the total number of pairs of people who are separated by this distance

$$P_l(d) = \frac{L(d)}{\text{pairs}(d)}. \tag{1.15}$$

An example of the dependence of this property on the distance is shown in Figure 1.4. This probability decays as a power-law of the physical distance in social networks, with the exponent depending on the characteristics of the social system: between 0.7 and 1 in online friendships (Liben-Nowell et al., 2005; Scellato et al., 2011; Grabowicz et al., 2013c) and around 2 in phone call records (Lambiotte et al., 2008). Furthermore, in OSNs a plateau is observed after the initial power-law decay, for distances over 200-500 km,[10] which implies that above a certain distance interactions are independent of the distance. Some further aspects of the relation between geography and online social contacts have been studied. For instance, the probability that a link at a given distance closes a triangle decays with the distance and saturates for distances above 40-200 km (Lambiotte et al., 2008; Grabowicz et al., 2013c). Furthermore, the Jaccard similarity of neighbors of connected users decays with the distance as well (Volkovich et al., 2012). Both the Jaccard similarity and the link reciprocity appear to decay slowly with distance following $-\log(d)$.

Geographic constraints affect not only the structure of spatial networks but also the processes that take place on these networks, such as social interactions, epidemic spreading, and network navigation. In disease spreading it is important to take geography, social networks, and virus properties into account to make predictions. The flow of people between cities is usually described by the gravity model (Zipf, 1946; Krings et al., 2009; Balcan et al., 2009). This model assumes that the flow of people is proportional to product of population sizes of the two cities and inversely proportional to square of the distance between them. Note that the gravity model corresponds to the power-law with exponent equal 2 in $P_l(d)$. A comparison of the results of the gravity model and the more complex radiation model against census data was presented in (Simini et al., 2012).

[10] The exact value of this distance depends on the online service and the geography of the country for which it is measured (Grabowicz et al., 2013c).

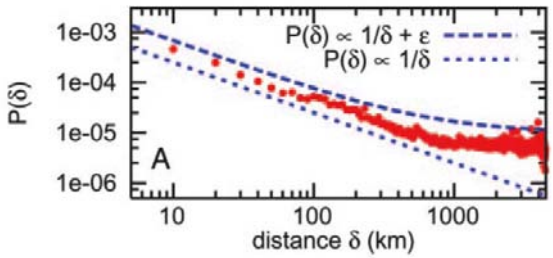


Figure 1.4: Probability of an online friendship as a function of distance in LiveJournal, adapted from (Liben-Nowell et al., 2005).

1.3 Selected methods of complex networks

Random models of networks play a fundamental role. Features extracted from real networks are compared to the values in corresponding random null models to understand how a real network differs from random graphs and if these differences are statistically significant. For instance, random graph models are used in the definition of modularity or in clustering methods.

1.3.1 Models generating random networks

Random graph model is an *ensemble* of graphs, defined as a set of distinct graphs with probabilities of their appearance in the model. Random graphs are created under certain assumptions, e.g., the number of edges is given, or the distribution of degrees is fixed, while links are placed at random. In this subsection, we introduce two fundamental random graph models, namely classic random graph and configuration models, a method of network randomization, and two more complex, yet still basic, models of preferential attachment and small-world network. The graphs created with classic random graph and configuration models are expected to have clustering coefficients, modularity, and assortativity close to zero in the limit of large number of nodes.

The *Erdős–Rényi* (ER) model (Rapoport, 1957; Erdos and Rényi, 1960) is a classic example of a random graph model: a network is constructed by randomly connecting pairs of nodes with probability p, independently of any other factor. The degree distribution in ER graphs is thus binomial:

$$P(k) = \binom{N \quad 1}{k} p^k (1 \quad p)^{N \;\; 1 \;\; k}, \tag{1.16}$$

where N is the total number of vertices in the graph. Naturally, this distribution becomes Poissonian

$$P(k) = \frac{(Np)^k \mathrm{e}^{-Np}}{k!}. \tag{1.17}$$

as $N \to \infty$ while $Np = \text{const}$. These distributions decay fast with the degree, what corresponds to the lack of hubs in the network. Interestingly, the ER model is equivalent to a system maximizing Gibbs entropy under a simple Hamiltonian controlling the number of edges (Park and Newman, 2004).

Figure 1.5: An illustration of the rewiring procedure.

Whereas in the ER model the degree distribution has a particular functional form, in the *configuration model* (Bender and Canfield, 1978; Molloy and Reed, 1995) a degree sequence, and consecutively degree distribution, is given a priori. In this random model, the connections are created between pairs of nodes with fixed degrees by randomly selecting disconnected edges of different nodes and linking them. (Naturally, the total sum of degrees has to be even to connect all the edges.) This process leads to creation of self-loops and multiple edges per a pair of vertices, which often is not desired. However, the average number of self-loops and multiedges is a constant as the network becomes large, meaning that they can be neglected in the limit of large graphs (Newman, 2010). A similar procedure can be applied to randomize existing networks (Maslov and Sneppen, 2002). It consists of rewiring the links in the way that preserves the degrees of nodes. It proceeds as follows (Figure 1.5). First, two edges between distinct nodes are randomly chosen; one between nodes i and j, another one between k and l. Next, one of the ends of each of the two edges is swapped, so that now i is connected with k, and j is connected with l, while the old edges are destroyed. Note that such method can produce multiple edges per each pair of nodes, unless an explicit instruction forbids it. This rewiring method, also known as reshuffling, is widely used in the studies involving complex networks as a mean of network randomization that maintains the degree distribution intact.

A more complex, yet still basic, random models generating graphs include the model introduced by Barabási and Albert, i.e., the *BA model* (Barabási et al., 1999) and *small-world* network (Watts and Strogatz, 1998). The former is based on the rule of preferential attachment. To describe it shortly: at each time step,

one node is introduced to the system with m edges that get connected to existing nodes in the system with probability proportional to the degree of the nodes. The model is initiated with $m_0 > m$ vertices present in the system that form a fully connected graph. As a result, a network with power-law distribution of node degrees with the exponent $\alpha = 3$ emerges. In that latter small-world network model a regular lattice is taken as a base with a high clustering coefficient and low average shortest path length (Figure 1.6). A percentage p of all connections in the regular lattice is rewired randomly to control the values of clustering coefficient and average shortest path length. The larger the p, the lower is the clustering coefficient, the higher gets the shortest path length, and the more similar the small-world network becomes to ER random graphs.

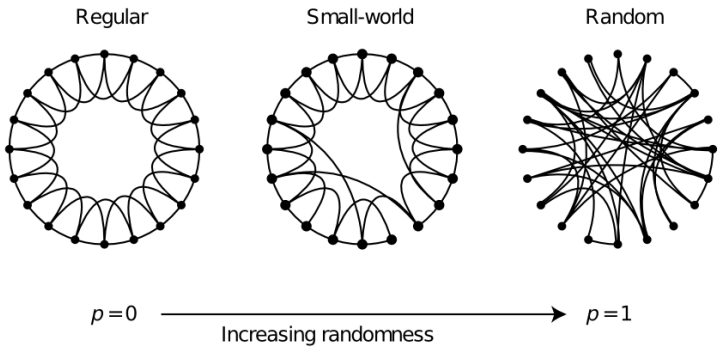
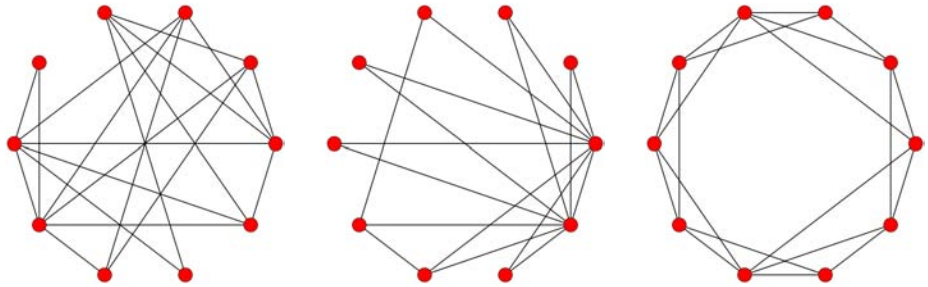


Figure 1.6: The small-world networks. The amount of the randomness introduced by the random rewiring interpolates between a regular ring lattice and a random graph. Illustration adapted from (Watts and Strogatz, 1998).

A comparison of an ER graph, a network created with the BA model, and a small-world network is shown in Figure 1.7. The graph created with the ER model shows a random structure, while the network created with the BA model has a scale-free structure with clearly visible hubs, and the small-world network yields a regular structure with random edges, characterized by lack of hubs, a high number of triangles, and a low average shortest path. While the ER model is a plain random model, the BA and the small-world models introduce the concepts of preferential attachment and clustering into random network structure. Because of their simplicity and characteristics, these three models are widely used as a toy-models in various studies, e.g., in agent-based modeling.



(a) Erdős–Rényi graph (b) Barabási-Albert model (c) Small-world graph

Figure 1.7: A comparison of various random graph models in a circular layout. Each of the graphs contains 10 nodes and around 20 edges. The rewiring probability in the small-world network is set to 0.1. (a) The classic random graph model shows no specific structure; (b) the preferential attachment model shows hubs and heterogeneity in the degrees; and (c) the small-world network shows a high number of triangles and a couple of random shortcuts.

1.3.2 Clustering algorithms

One of the important topics in network science is the detection of communities defined as more densely connected subgraphs of the network comparing with their neighborhood. Various community detection algorithms have been developed and they continue to be developed. A recent review described available methods, which differ in techniques, features, and capabilities (Fortunato, 2010). Here, we offer an introduction to these methods.

Perhaps the most popular method of community detection consists of modularity maximization, defined by Equation 1.11. It detects communities by searching over possible partitions of the network to find the one with high modularity. Since exhaustive search over all possible partitions is usually intractable, practical algorithms are based on approximate optimization methods such as greedy algorithms, simulated annealing, or spectral optimization, offering different balances between speed and accuracy (Danon et al., 2005; Fortunato, 2010). Despite its popularity modularity optimization is not free of certain problems, such as the resolution limit (Fortunato and Barthélemy, 2007; Good et al., 2010; Fortunato, 2010), or difficulties to find the absolute maximum of the modularity due to a rough landscape of its value in the space of the possible network partitions (Good et al., 2010). Furthermore, the modularity yields positive values even in random networks (Guimerà et al., 2004), which may lead to detection of communities in such networks, even though they are not present there.

Clustering methods are distinguished by other aspects apart from the ap-

proach that they use to find groups. To start with, some of the algorithms take directionality of links into account, while others are designed only for undirected networks. To convert a directed network to an undirected one a symmetrization procedure is applied. This procedure, however, neglects information that may be important to define the groups, and it can affect the performance of the methods. Another aspect is the ability to find overlapping communities, i.e., nodes belonging to many groups. Many of the clustering algorithms assume that each node belongs just to one community, which is not realistic, for instances, in the context of social communities. Finally, the accuracy of the methods varies. Clustering methods are compared using benchmark networks in which the groups are defined a priori; the level of disorder is increased by introducing random connections; and the methods are tested by measuring to which point they recover the planted groups. One of the best performing algorithms in this sense is OSLOM (Lancichinetti et al., 2011). It is a local optimization method using order statistics. It estimates statistical significance of each group, defined as a probability of finding the cluster in a random null model, namely the configuration model. OSLOM takes directionality of links into account and detects overlapping communities (see Appendix I). We use it for group detection in the studies presented in the following chapters.

1.4 The growth of complex networks

This section describes models of the growth of complex networks and systems. The methods based on preferential growth and intrinsic heterogeneity are used in complex systems, including complex networks. The methods based on triangle closing are used particularly for modeling social networks.

1.4.1 Preferential growth

Many features of complex systems are characterized by heavy-tailed distributions (Newman, 2005; Saichev et al., 2009). This property is typically perceived as a symptom of the rich-gets-richer principle, from which the so-called preferential growth stems. Imagine a system of elements with a certain property that can grow incrementally for each of the elements. The common idea behind preferential growth models is that a property of the elements grows proportionally to its current value, i.e., the elements that are big grow faster than elements that are small. Typically, in preferential growth models a constant number of increments are introduced to the system at each time step. The increments are used to create new elements and/or to increase the value of the growing property of the

existing elements accordingly to the preferential growth rule. Such models are usually the first approach to explain heavy-tailed distributions in many different systems (Simon, 1955; Albert and Barabási, 2002). In the case of networks, preferential attachment models were among the first studies igniting the field of network science (Barabási and Albert, 1999; Barabási et al., 1999; Huberman and Adamic, 1999; Dorogovtsev et al., 2000; Bornholdt and Ebel, 2001). The first such model, namely the aforementioned BA model, was introduced in (Barabási et al., 1999).

The preferential attachment model is simple, which is a desirable feature, but may cause some concerns. In models of preferential growth, the time unit is directly coupled to the number of new arriving elements, which complicates the comparison of the dynamics of these models with real data. Some other drawbacks include the lack of diversity between the elements and strong correlation between age of elements and their size (Adamic and Huberman, 2000; Klemm and Eguíluz, 2002b) (Figure 1.8). For the presented systems, the model reproduces the dependence between time and average node degree, but it does not explain the high fluctuations of the degree present in the scatter plot, nor the anti-correlation of degree growth with the age of the nodes. Because of these issues, the basic preferential growth model is typically used as a simple toy-model for the generation of networks with power-law distribution of degrees. On the other hand, it is also successfully used as a component of more complex models simulating the growth of real social networks (Mislove et al., 2008; Leskovec et al., 2008).

1.4.2 Heterogeneity

In many real systems, especially in social systems, individuals or elements are diverse. This factor may be one of the reasons why heavy-tailed distributions are so commonly found in complex systems, i.e., due to the diversity in the intrinsic properties of the elements of the system. In this direction, some models of growth incorporating heterogeneity in the form of fitness, hidden variables, or rankings have been proposed (Bianconi and Barabási, 2001b; Caldarelli et al., 2002; Söderberg, 2002; Boguñá and Pastor-Satorras, 2003; Ratkiewicz et al., 2010). A detailed discussion of a fitness-based model can be found in the chapter devoted to the description of groups' growth in an OSN. In general, in this family of models the growth rate of elements depends on an intrinsic property that characterizes them. Whereas in preferential attachment models the growth is often proportional to the current size of the elements, in fitness models it is usually proportional to the intrinsic fitness of each element. Typically, the fitness is a random variable specific for each element drawn from a given distribution. If the distribution is broad, then high heterogeneity is present in the system. How-

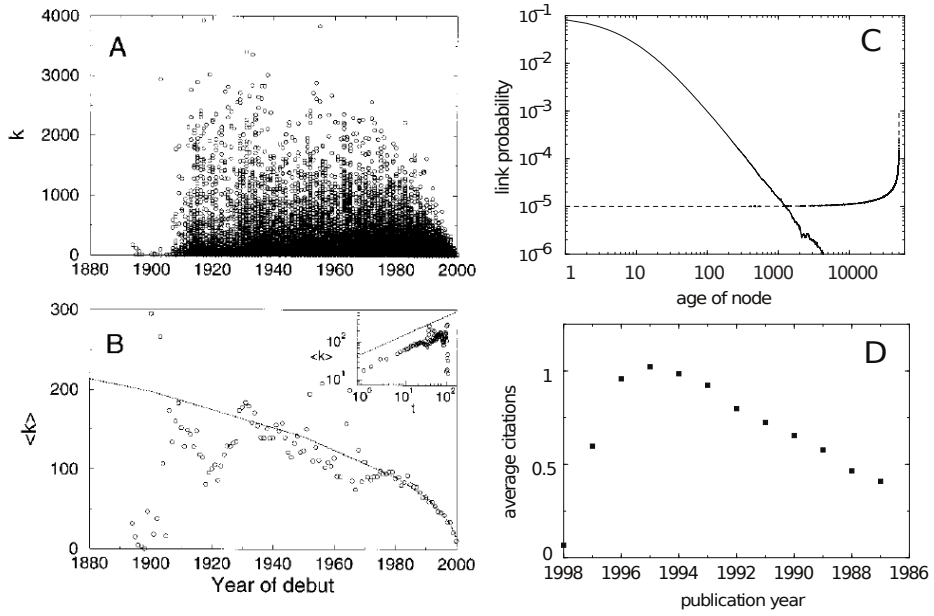


Figure 1.8: The correlation between age and degree of nodes in a network of: (A, B) actors starring together in the same movies, (C, D) scientific publications connected by citations. The figures shown are: (A) A scatter plot of the degree as function of the age; (B) The average of degree as a function of the age in the real network (symbols) and in the BA model (line); (C) the dependence of the probability of degree growth on the age in a model reproducing the degree distribution and clustering (solid line) and in the BA model (dashed line); (D) The average of degree as a function of the age in the real network (symbols). The model reproduces well the average dependence (B), however, it is unable to explain neither the high fluctuations in the scatter plot (A), nor the anti-correlation of degree growth with the age of the nodes (C). Figures (A, B) are adapted from (Adamic and Huberman, 2000) and Figures (C, D) from (Klemm and Eguíluz, 2002b).

ever, there is rather little empirical work showing how intrinsic heterogeneity is distributed and what is its role in complex system growth (Garlaschelli and Loffredo, 2004; De Masi et al., 2006; Kong et al., 2008). In the next chapter, we show that fitness is log-normally distributed in a system of growing groups and directly compare results a model based on heterogeneity to a model based on preferential growth.

1.4.3 Coupling between heterogeneity and preferential attachment

Heterogeneity and preferential growth, play a crucial role in the growth of popularity of content (Salganik et al., 2006), as shown by experiments in an online platform in which users can listen to and download songs. The experiments are performed in different conditions: independent (user choice of songs is unaffected by any social factors, i.e., users do not see the number of downloads of a song), social influence (users see how many times a song was downloaded) and stronger social influence (users see a sorted list of songs in a descending order of the number of downloads). On the one hand, increasing the strength of social influence increases both inequality and unpredictability of success (measured as the number of downloads). Users tended to download songs that were already many times downloaded by others, as in preferential growth. On the other hand, the intrinsic quality measured in the independent condition is correlated with the popularity of songs measured in the other conditions. In other words, the best songs rarely did poorly, and the worst rarely did well. Therefore, the experiment shows that both preferential growth and intrinsic heterogeneity influence the popularity of the songs. In conclusion, an interplay of both mechanisms shapes the growth of popularity in such social systems, e.g., the growth of the number of citations of scientific papers (Wang et al., 2013).

1.4.4 Triangle closing

The concept of *triadic closure* derives from sociology. Due to the fact that the clustering coefficient is remarkably high in social networks, several models have been introduced in order to reproduce the high number of triangles in such networks. One of the first simple network models accounting for the high clustering is the aforementioned small-world network (Watts and Strogatz, 1998). Other random models introduce high clustering in growing scale-free networks (Holme and Kim, 2002; Klemm and Eguíluz, 2002a; Dorogovtsev et al., 2002) or in the configuration model (Holme and Kim, 2002), and allow to tune the level of clustering (Serrano and Boguñá, 2005; Toivonen et al., 2006).

A more sophisticated model based on triangle closure simulates the growth of OSNs (Leskovec et al., 2008). In this model, each node draws its arrival time using a node arrival function and its lifetime from an exponential distribution.

Next, it connects to a random node with the preferential attachment rule. Then, it draws the waiting time to create its next link from a degree-dependent power-law distribution with a cutoff. Finally, it creates its next connection by closing a triangle with a random neighbor. The last two steps are repeated until the node lifetime passed. The study has tested other linking mechanism than the preferential attachment for the first connection and the random triangle closure for all other connections, and found that these two mechanisms yield the best results. The model produces networks with the structure corresponding to the real structure of OSNs in terms of the degree distribution, the dependence of clustering coefficient on degree, and the distribution of the shortest path length between nodes (Leskovec et al., 2008). In a latter chapter of this thesis, we introduce a model of coupled mobility and network growth that induces triangle closing by spatial coincidences of people visiting their friends. In the next section, we offer the sociological definition of triadic closure and show further proofs of its existence and its relation to other phenomena found in social networks.

1.5 Sociological theories

On the one hand, OSNs are the outcome of technological development. On the other hand, they can be considered a representation of social relations and the interactions of people. In that context, one can think of OSNs as a product of a crossover of supply and demand, where supply refers to the technological capabilities, while demand refers to the needs of the human population (emergence of an information dissemination model based on personal subscriptions enabled by computer-based systems was foreseen by Brown et al., 1967). As such, OSNs are meant to satisfy social needs of people and they provide information on their social behavior. Until the end of the twentieth century, social behavior has been studied by traditional means by sociologists with self-reported surveys or small-scale field experiments. One can expect that many of the sociological theories developed for offline social systems also hold in online systems. To this end, we introduce sociological concepts and theories describing social systems in the following subsections and test some of them in the subsequent chapters (for a summary see Table 1.1).

1.5.1 Tie formation mechanisms

There are various theories explaining why social ties are formed, persist, and dissolve. An interested reader can find a broad classification in the review (Katz et al., 2004), which we have briefly introduced at the beginning of this chapter, in

Theory/Theories	Tested mechanisms or concepts	Desc.	Chap.
Relational and proximity mechanisms of tie formation	Network distance of two hops and physical coincidence dramatically increase chances of tie formation.	1.5.1	5
The strength of weak ties	Strong ties are inside groups, whereas weak ties are between groups. Weak ties facilitate information diffusion.	1.5.3	3
The diversity-bandwidth tradeoff and structural folds	The strong ties between groups diffuse more information than the weak ties between groups due to high diversity and wide bandwidth.	1.5.3	3
Identity-bond theory	Groups in social systems are based on social bonds or common identity and yield different characteristics.	1.5.4	4

Table 1.1: List of sociological theories and concepts explored and/or tested in the following chapters of the thesis.

Subsection 1.1.2. Here, we introduce the classification listed in the review of sociological research (Rivera et al., 2010). Three specific sociological mechanisms are responsible for tie formation: relational, assortative, and proximity. Naturally, this is just one of the possible classifications.

In the relational mechanism, the formation of ties is determined by the structure of the social network. One of the earliest and most important concepts of this kind is *triadic closure* (Simmel, 1950), depicted in Figure 1.9. It suggests that, if an actor has ties with two other people, then these two individuals will likely have a tie between themselves as well. This implies that there are many social triangles in social networks. Recent large-scale studies have confirmed the existence of triadic closure in both offline (Kossinets and Watts, 2006) and OSNs (Leskovec et al., 2008; Gallos et al., 2012) and compare it against other mechanisms. The triadic closure is related to the formation of groups and social cohesion (Moody and White, 2003; Forsyth, 2009), that is an inclination to actively participate in the group interactions. It is important for the development of strong social ties (Krackhardt and Handcock, 2007; Granovetter, 1973), as there exists a correlation between the strength and the number of common friends shared between two actors, and affects the emergence of positive and negative relations (Leskovec et al., 2010; Szell et al., 2010), e.g., good and bad relationships, friends and foes.

The assortative mechanism suggests that ties are formed due to the compatibility and complementarity of actors' attributes. This concept corresponds to

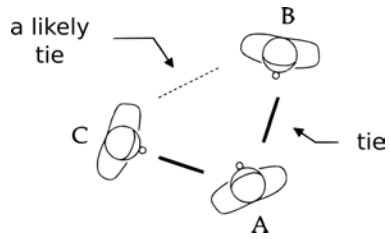


Figure 1.9: An illustration of triadic closure. According to the rule, if the actor A has ties with actors B and C, then actors B and C will likely have a tie.

homophily and heterophily. Note that actors in their choice of social ties may seek a balance between similarity on some dimensions and heterogeneity on different dimensions.

The proximity mechanism assumes that ties are formed due to physical or cultural proximity. The former concept is straightforward, as physical co-location dramatically increases the chances of interaction and the formation of a social tie, and results in the distinctive geographic properties of social networks (see Subsection 1.2.7). The latter corresponds to cultural and social environments or interests that bring actors together, and increase the chances of positive sentiment and interactions.

Note that the three mechanisms correspond to the characteristics of social networks (see Section 1.2): high clustering, assortativity/homophily, and geographic proximity; respectively. In the reminder of this dissertation, we introduce a model that realizes two of the three mechanisms, namely the triadic closure and the proximity mechanism, to create artificial social networks with realistic geographic and structural properties.

1.5.2 Strength of ties

Intuitively, strong social ties correspond to close friends, strong relationships, kinship, etc. The strength of ties and the relation between strong and weak ties were first considered at the beginning of twentieth century Simmel (1950). It was a much later work of Granovetter (1973), however, that popularized the concept. Here, we introduce the meaning and ways of measuring the strength of ties in practice.

Granovetter suggested that strong ties are characterized by four properties. First, actors connected by a strong tie spend a great deal of *time* together. Second, there is a high *emotional intensity* between them, for instance, in the form of parental or romantic feelings. Third, there exists a certain level of *trust and intimacy* between the actors. Fourth, interactions and services between the actors

tend to be *reciprocal*. Each of these indicators may be present in a tie to a certain degree. How to define and measure the strength of a tie is an open problem. For instance, Krackhardt (1992) defined alternative *philos* relationship.[11] as the one that meets the following three conditions: it contains frequent interactions, certain affection is involved, and there is a history of interactions.

Several studies investigate ways of measuring tie strength in both offline (Marsden and Campbell, 1984) and OSNs (Gilbert and Karahalios, 2009; Jones et al., 2013). Such studies are conducted through surveys asking the participants about kinship, closeness, and the intensity of their relations to contrast these properties with other indicators that are more easily quantified. It has been demonstrated that the time spent together is positively correlated with the tie strength (Marsden and Campbell, 1984). In OSNs, the number of interactions (various types of interactions are introduced in the next section) and coincidences in photos (signaling physical meetings and geographical proximity) predict the tie strength between a pair of users (Gilbert and Karahalios, 2009; Jones et al., 2013). Similar measures of tie strength have been used in other studies, e.g., aggregated duration of phone calls (Onnela et al., 2007b) or number of comments and private messages exchanged between users (Bakshy et al., 2012).

Measuring the number of interactions and photo coincidences is straightforward in online systems and can be done automatically for the whole system on a scale of millions or even a billion users. It is not straightforward in such systems to measure intimacy. To this end, intermediary metrics are used. For instance, the context of physical proximity signals intimacy; i.e., time spent together after work hours and during weekends signals intimacy, in contrast to time spent together at work during work hours (Eagle et al., 2009). Furthermore, intimate words can be identified in textual conversations and signal a strong relationship (Gilbert and Karahalios, 2009).

1.5.3 Structure, tie strength, and information diffusion

The theory known as *the strength of weak ties* deals with the relation between structure, tie strength, and the diffusion of information in social networks (Granovetter, 1973). First, a tie is characterized by its strength, which is related to the time spent together, intimacy, the emotional intensity of a relation, and reciprocity. Strong ties refer to relations with close friends or relatives, while weak ties represent links with distant acquaintances. Second, a tie can be characterized by its structural position in the network. Granovetter's theory suggests that strong ties occur between people who have many friends in common, i.e., inside social groups, while weak ties occur between actors who have few friends in com-

[11] "Philos" is the name of a type of relationship. It comes from Greek and was introduced by Krackhardt as a noun.

mon, i.e., between social groups. Third, a social tie can transport information between the two connected actors. Granovetter's theory predicts that weak ties are important for the diffusion of new information across the network because they act as bridges between groups of people, likely having different sources of information. Strong ties are predicted to be less important for information diffusion, as they tend to be located in the interior of groups between actors who have many friends in common and the same sources of information.

The advantage of connecting different groups to access novel information due to the diversity in the sources was emphasized by Burt (2005). It described the concept of bridging *structural holes*, which refers to the connections bridging the sparse areas between communities. An example of a structural hole is illustrated in Figure 1.10 (left illustration). The importance of structural overlap for the emergence of trust and intimacy and for the formation of social groups is emphasized in the theory of structural cohesion (Moody and White, 2003).

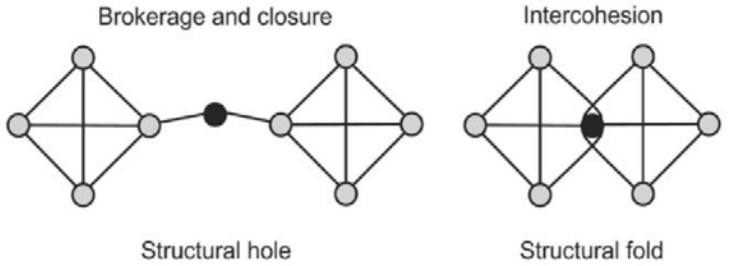


Figure 1.10: Illustration of structural holes (brokerage and closure) versus structural folds (intercohesion). The person marked with the black circle is in the brokering position in the left and right cases, but in the right case, she is also in cohesive positions with both groups; while in the left case, she is not. Illustration adapted from (Vedres and Stark, 2010).

More recent works have pointed out that information propagation depends on a *diversity-bandwidth tradeoff* (Aral and Van Alstyne, 2011) and an interplay of brokerage and social cohesion (Vedres and Stark, 2010). Diversity relates to the heterogeneity in knowledge and information held between the two communicating parties. The bandwidth of a tie is defined as the rate of information transmission per unit of time. Aral and Van Alstyne noted that weak ties interact infrequently and have low bandwidth, whereas strong ties interact more often and have high bandwidth. The authors claimed that both diversity and bandwidth are relevant for the diffusion of novel information. Since both are anti-correlated, there must be a tradeoff to reach an optimal point in the propagation of new information.

They also suggest that strong ties may be important to propagate information depending on the structural diversity, the number of topics, and the dynamic of the information.

The study of Vedres and Stark (2010) suggests that actors can be members of multiple groups, creating so-called *structural folds*, as shown in Figure 1.10 (right illustration). Such structural position guarantees the maintenance of cohesive ties between actors sharing friends in common, as well as ties that connect diverse environments, i.e., groups that have access to different sources of information. The study shows that actors in structural folds tend to have better access to novel information.

1.5.4 Common identity and common bond theory for groups

Notions of community and social groups have been widely studied in the behavioral sciences (Riger and Lavrakas, 1981; Tajfel, 1982). It has been shown that the internal dynamics of social groups emerge from the combination of complex cognitive processes, such as a sense of membership, influence between people, fulfillment of individual and collective needs inside the group, and shared emotional connections (McMillan and Chavis, 1986).

The *common identity* and *common bond* theory describes social groups along the dimensions of topicality and sociality (Prentice et al., 1994; Ren et al., 2007). Attachment to a group, as well as its permanence and one's involvement in it, can be explained in terms of a common identity or common bond. Identity-based attachment holds when people join a group based on their interest in the community as a whole or in a well-defined common theme shared by all of the members. People whose participation occurs due to identity-based attachment may not directly engage with anyone and might even participate anonymously. Conversely, bond-based attachment is driven by personal social relations with other specific members, so the main theme of the group may be disregarded. The two processes result in two different group types, which for simplicity are named in this dissertation "*topical*" for identity-based attachment and "*social*" for bond-based attachment.

In practice, groups are formed from a mix of identity and bond-based attachment, but often they lean more toward either sociality or topicality. According to the theory, the group type is related to the *reciprocity* and the *topics* of discussion. Interactions in topical groups are generally not directly reciprocated, whereas members of social groups tend to have reciprocal interactions with other members. This is related to the definition of tie strength, which involves reciprocity, according to Granovetter. In addition, discussions tend to be related to the group theme and cover specific areas in topical groups, while in social groups topics of discussion tend to vary drastically and cover multiple subjects. Further-

more, topical groups tend to be bigger than social groups, due to cognitive limit to the number of stable relations an individual can maintain (Dunbar, 1998). The limit is usually assumed to be around 150 individuals, however, other values have been also proposed (Gonçalves et al., 2011; Miritello et al., 2013).

According to the theory, topical groups are more open to newcomers and more robust to departures of members. Social groups, on the other hand, are founded on individual relationships between their members, so it is harder for newcomers to join and integrate with members that already have strong relationships with each other. One implication is that social groups are vulnerable to turnover since the departure of a person's friends may influence his own departure.

1.6 Online social networks

An increasing number of social interactions occur using OSNs as communication channels. Some OSNs have become extremely popular, reaching up to a billion active users. They differ in the character of the service they provide to online users. For instance, Twitter is mainly used to propagate and receive news, Flickr gathers amateurs and professionals in photography, Facebook is primarily a platform for keeping in touch with close friends and relatives. Albeit different, all these online platforms share an ingredient that pervades all their applications. There exists an underlying social network that allows their users to keep in touch with each other and helps to engage them in common activities or interactions leading to a better fulfillment of the service's purposes. This is the reason why these platforms share a good number of functionalities, e.g., personal communication channels, easy one-step information sharing, and news feeds containing broadcasted content (see more in Table 1.2). As a result, OSNs are an interesting field in which to study online social behavior that seems to be consistent among different online services and offline social networks.

Common functionality	Twitter	Flickr	Facebook	Sect.
Declared social network	followers	contacts	friends	1.6.3
Personal communication channel	mentions	comments	comments	
One-step information sharing	retweet	-	share	
Collaborative content assessment	favorites	favorites	likes	
Groups maintained by users	lists	groups	groups	1.6.5
Content tagging	hashtags	tags	tags	1.6.6
Geo-localized content	posts	photos	posts	

Table 1.2: Common functionalities in various globally established OSNs.

Since at the bottom of these services lies a network of declared relations and the basic interactions in these platforms tend to be pairwise, i.e., between pairs of users, a natural methodology for studying these systems is provided by network science. One of the most natural questions about OSNs concerns their relation with offline social networks (Wellman et al., 1996). It can be tackled in a few ways, although all methods have demonstrated that offline and online behaviors are related to a considerable extent. First, direct studies have been conducted comparing online relations with offline ties (Jones et al., 2013). Second, sociological theories have been tested in OSNs (Leskovec et al., 2010; Szell et al., 2010; Gruzd et al., 2011; Bakshy et al., 2012; Ugander et al., 2012). Third, various psychological (Quercia et al., 2011, 2012b) and economical (Quercia et al., 2012a; Mitchell et al., 2013) variables have been correlated with online behavior. We test sociological theories related to groups in the following chapters of this dissertation.

The next subsections describe the common infrastructure of the three mentioned OSNs, i.e., Twitter, Flickr, and Facebook. Note that in the following chapters we will present studies based on datasets from Twitter and Flickr, focusing on the features of the OSNs that constitute the abstract frame outlined in Table 1.2.

1.6.1 Description of exemplary online social networks

Twitter[12] is a micro-blogging social site. Each user has her own profile with a timeline that can be accessed by logging in. A user can write short messages of up to 140 characters, called *tweets*, which are saved in her timeline and broadcasted to the users who *follow* her. Her *followers* see tweets from her, and other users who they follow, in an integrated timeline called the *news feed*. When a new follower relation is established, the targeted user is notified, although his or her explicit permission is not required. Thus, the declared network is directed, and the connections are cheap to form. Furthermore, using modern mobile devices, the user can choose to attach to the tweet his current geographic position in the form of GPS coordinates with a single click. By default, all the tweets created by the user are publicly visible.

Flickr[13] is an image and video hosting website that implements social media features. Each user has her own profile with the timeline of all the photos uploaded by the user to date, called a *photostream*. As in the case of Twitter, the user can mark other users as *contacts*, creating directional links. The photos and videos uploaded by the contacts are broadcasted to the user and shown in his news feed. The photos can contain information about the geographic loca-

[12] Available at `http://twitter.com`.

[13] Available at `http://flickr.com`.

tion where they were taken. The users can comment on one another's photos by simply writing underneath them and mark *favorite* photos with a single click to express their satisfaction. Furthermore, the users can create, join, and administer groups. The photos can be posted in group *pools* of photos to increase their exposure. If the user does not choose otherwise, all the content uploaded by the user is publicly visible.

Facebook[14] is currently the largest OSN[15] and the second most popular website on the Internet.[16] It is a social media platform for sharing content with *friends*. A user can add other users to her set of friends, but her consent is required, and a bidirectional link is formed between the two, in contrast to Twitter and Flickr. Thus, the underlying declared network is undirected. Facebook is an integrated platform in the sense that it allows broadcasting of messages of various sizes and uploading of photos and videos. By default, all the uploaded content is not visible publicly, in contrast to Twitter and Flickr once again.

1.6.2 Structure of declared networks

The declared connections in OSNs represent a relation between the users, in the form of either undirected or directed links, i.e., directed followers in Twitter, directed contacts in Flickr, and undirected friends in Facebook. These declared networks exhibit properties found in other social networks (see Table 1.3).

The number of users in the networks is large, from several million to a billion. The average degree varies from 17 in Flickr to 190 in Facebook and it increases as the networks grow. The degree distribution in these networks is heavy-tailed, as shown in Figure 1.11. Although these are not power-law distributions, some of their parts can be approximated with a power-law, e.g., the distributions of out-degree in Flickr above and below 500 contacts (Figure 1.11C). Furthermore, the bump in the distribution for Twitter (Figure 1.11A) around 20 and the plateau for degrees up to 50 in Facebook (Figure 1.11B) correspond to certain policies of the services, i.e., recommender systems suggesting the first 20 friends, etc. The maximum degree in Facebook is around $5,000$, which corresponds to a limitation imposed in this platform. Some users in Twitter have extremely large numbers of followers. These include celebrities and large news outlets, e.g., Britney Spears, CNN, and New York Times. A study suggested that a log-normal distribution fits well the degree distribution in Twitter (Galuba et al., 2010).

Link reciprocity in OSNs differs. Low reciprocity signals that the network's social component is not strong and that the graph fulfills other roles, e.g., Twitter

[14] Available at `http://facebook.com`.

[15] It has almost twice as many users as the second largest OSN, according to GlobalWebIndex; see more at `http://zd.net/17ziSKc`.

[16] Popularity in terms of traffic according to the Alexa ranking from September 2013. For a recent ranking, visit `http://www.alexa.com/topsites`.

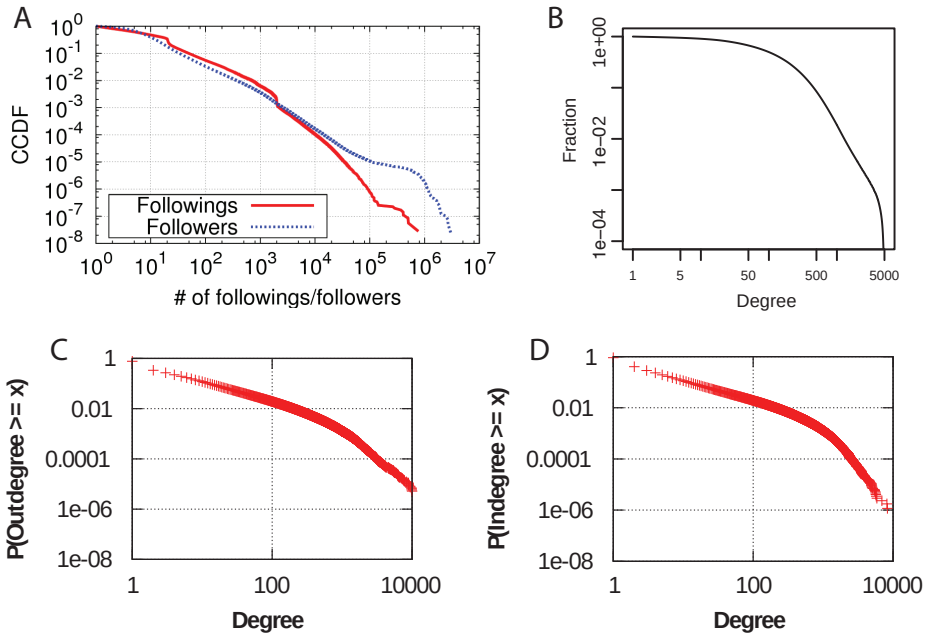


Figure 1.11: Complementary cumulative distribution of node degrees in: (A) Twitter, (B) Facebook, (C,D) Flickr (out-degree and in-degree). Adapted from (Kwak et al., 2010; Ugander et al., 2011; Mislove et al., 2007), respectively.

is considered a news media rather than a social network (Kwak et al., 2010). Most of the celebrities present in Twitter do not follow their followers in return. Facebook has a strong social component because its content is not publicly visible and the connections are bidirectional.

The average clustering coefficient in the networks is one or more orders of magnitude higher than its expected value in an ER random graph or a preferential attachment network (Mislove et al., 2007). The clustering coefficient varies with the degree of nodes; e.g., in Facebook, it takes values from 0.1 for the nodes of degree around $1,000$ up to 0.5 for the low-degree nodes. High clustering is related to the community structure; e.g., the clustering coefficient is 50% higher in the groups created by the users of Flickr (Mislove et al., 2007). Some of the largest communities correspond to countries. In fact, most of Facebook connections are internal to countries, that is, over 84% of the connections. The modularity of the Facebook friend graph partitioned into countries is 0.75, an extremely high value exposing a high-level community structure.

The average path length in the three networks is low and ranges from 4.1 in

Property	Twitter[a]	Flickr[b]	Facebook[c]	Introduction
Year measured	2009	2007	2011	-
N	42 M	6.9 M	721 M	1.2.1
k^{out} or k	35	12	190	1.2.2
R	0.22	0.62	undirected	
$\langle c_i \rangle$	0.11[d]	0.31	0.1-0.5	1.2.3
l	4.1	5.7	4.7	1.2.4
r	> 0	0.20	0.27	1.2.6

Table 1.3: Properties of declared social networks in Twitter, Flickr, and Facebook, based on the respective studies: [a](Kwak et al., 2010), [b](Mislove et al., 2007), [c](Ugander et al., 2011), [d](Java et al., 2007).

Twitter to 5.7 in Flickr, which corresponds to the concept of six degrees of separation. The strongly connected component contains a majority of nodes in each of the OSNs, i.e., 92% in Twitter (Galuba et al., 2010), 45-60% in Flickr (Kumar et al., 2006; Mislove et al., 2007), and 99.91% in Facebook (Ugander et al., 2011). One of the reasons why the average path is so short may be the existence of a dense core, as argued in (Mislove et al., 2007). The study showed that the removal of 1-10% of the nodes with the highest degree from the Flickr network fragments it into numerous smaller components. The dense core consists of many very high-degree nodes connected to each other due to the assortative mixing present in the OSNs. The assortativity coefficient takes positive values in the networks, up to the value of 0.27, meaning that nodes of similar degrees tend to be connected to each other, i.e., nodes of high degree tend to be connected to other nodes of high degree.

1.6.3 Pairwise interactions

In general, a social network is a broad term that refers to a set of actors and a set of ties between them representing some kind of relation or interaction. In fact, however, there are many types of both relations and interactions (Borgatti et al., 2009; Butts, 2009), and they usually happen on top of each other.

In OSNs, the types of relations and interactions are often specified, allowing for more in-depth studies. On the one hand, such networks have the declared connections. On the other hand, a variety of interactions are possible in these systems (for a summary, see Table 1.2). Some of them have emerged organically

through the development of online social conventions (Kooti et al., 2012)[17] and represent the natural social needs of the users. The interaction types that are the most popular and related to this dissertation are personal communications, information sharing, and content assessment.

First, personal communications can be interpreted as one type of such interactions, e.g., through *mentions* in Twitter or *comments* in Flickr and Facebook. These examples correspond to interchanges that are publicly visible. However, most OSNs have also a private communication channel that is accessible only to the parties involved. Unfortunately, such private interactions are usually not available to researchers due to privacy concerns. In either case, the communicated messages are directed to a specific person, who may choose to respond in the same way. Second, another type of interactions corresponds to information sharing, e.g., *retweets* in Twitter or *shares* in Facebook. A user can select a piece of content from a person she is connected to, and share it to her own followers, so that they can see it as well. Third, users can also interact with others by positively evaluating their content, e.g., with *favorites* in Twitter and Flickr or *likes* in Facebook. The receiver of a positive evaluation may decide to reciprocate it as a way of expressing gratitude or friendship. Each such interaction between two users can be represented by a link in a graph. Each of these interaction types has its own characteristics and its own representation in most of the established OSNs.

1.6.4 Interaction networks versus declared networks

The pairwise interactions of users of OSNs are related to their declared social relations. The comparison of a network built from declared online relations and a network built from user interactions shows several differences at the structural level. First, the actors tend to interact with much fewer people than they declare as friends, which results in smaller degrees of nodes in the interaction network (Viswanath et al., 2009; Wilson et al., 2009). Moreover, the friends they interact with change rapidly, and only about 30% of pairwise interactions in one month continue over the next month (Viswanath et al., 2009). A study of mobile phone calls shows that it is natural to maintain both long-term persistent and short-term exploratory relations, and that the exact ratio between the two varies from person to person (Miritello et al., 2013). Because the degrees are lower, the properties related to the small-world effect are also less evident. Namely, the average path length is higher (Wilson et al., 2009), and there are less densely connected cores (Chun et al., 2008).

[17] Blog posts from Twitter developers describing the first implementation to the system of social conventions for mentions/replies and retweets: `http://blog.twitter.com/2009/03/replies-are-now-mentions.html` and `http://blog.twitter.com/2009/08/project-retweet-phase-one.html`.

1.6.5 Groups

In OSNs, groups can be identified in a few ways. On the one hand, groups can be created and/or declared explicitly by the users themselves and subsequently directly retrieved from the data. On the other hand, community detection algorithms can be used to identify them from the network structure (see Subsection 1.3.2). A natural question is whether groups found with such methods are related and what their importance is. It has been found that declared groups internally tend to have higher clustering coefficients than the rest of the network (Mislove et al., 2007), so they may be correlated with the more densely connected parts of the network found by community detection algorithms. We make a direct comparison of detected and declared groups in Chapter 4.

Several aspects have been identified as positively influencing groups' growth and their persistence. The growth of declared groups is facilitated by low or medium clustering coefficient (Backstrom et al., 2006) and high internal connectivity (Taraborelli, 2011). Other work argues that flexibility of big detected groups helps them stay alive longer, while small detected groups are more persistent if their composition stays unchanged (Palla et al., 2007).

Furthermore, there exist different types of groups, i.e., in Subsection 1.5.4 we have introduced the notion of topical and social groups. The information about the type of groups is not given in OSNs. A natural question is whether declared and detected groups are topical or social.

1.6.6 Tagged content

Interactions in OSNs can have various attributes and content associated with them. For instance, some of the interactions between users of OSNs are moderated through user-generated posts or photos. A comment directed to a specific user can be posted underneath a photo that the target user has uploaded to her profile. The owner of the photo is notified and can either respond underneath her photo or through different means with a different type of interaction. These attributes and content can be leveraged to characterize the interactions beyond their structural properties and to understand better the behavior of users.

The content often has *tags* associated with it (examples are shown in Table 1.2). The tags refer to semantically compressed keywords describing the content. A system consisting of users, tags, and content pieces is called a *folksonomy* (Cattuto et al., 2007, 2008). In a latter chapter, we use tags to distinguish between the aforementioned topical and social groups. Another type of tags is the geo-localizing tags, which are simply the longitude and the latitude of geographic positions, usually localizing the piece of content in geographical space. Using them, one can draw conclusions about human mobility and the geographical properties of social networks and interactions. In the last chapters of this

thesis we use this information to analyze spatial properties of various OSNs.

1.7 Outline

In this thesis, we use methods of complex networks and complex systems to study OSNs. In the following chapters, we present four studies, three of which are concerned with groups of people, and the fourth one considers coupled spatial and structural properties of OSNs.

More specifically, Chapter 2 analyzes the growth of declared groups in Flickr. In this study, exceptionally, we do not use any network representation of the system. Instead, we focus on the time series of group sizes and on the properties of the system as a whole. The growth of groups is simulated with a model based on heterogeneity and compare its results with an alternative model based on preferential growth. We find that the model based on heterogeneity reproduces better the properties of the real system.

In Chapter 3 we extend the study of declared groups to groups detected with various graph-based clustering algorithms. Namely, we study the patterns of interactions in the landscape of groups detected in the follower network in Twitter. We distinguish between two types of interactions, i.e., personal communication and information sharing. We test if the interactions in this OSN follow the predictions of the sociological theories for offline social networks relating structure, tie strength, and information diffusion. We find that the statistical patterns of the interactions can be explained by these theories.

We explore the similarities and the difference between detected and declared groups in Chapter 4 using a dataset from Flickr. The detected groups are found with OSLOM, described in detail in Appendix I. The declared groups are created by users. First, we explore the membership overlap of the two sets of groups to check if they match each other. Second, based on the common-identity and common-bond theory, we classify statistically each group as either topical or social, using metrics quantifying reciprocity and diversity of topics of conversations. Our results are consistent with the theory. We compare the detected and declared groups in terms of their topicality and sociality.

Chapter 5 analyzes spatial and structural properties of three different OSNs, including Twitter. We measure several network properties as a function of physical distance between nodes, e.g., link probability, reciprocity, social overlap, and clustering coefficient. We introduce a model that couples mobility of agents and network growth and reproduces the statistical properties with a good accuracy. We compare its results against a triangle closing model and a random network model that connects nodes depending on the distance between them.

The dissertation is summarized in Chapter 6 with a discussion of our contributions to the emerging field of computational social science and a consideration of the importance of big data for the studies of social systems. For each of the topics of our studies we provide a broad outlook for the future research. The thesis concludes with a generic outlook for computational social science.

Chapter 2

Impact of heterogeneity on groups' growth

Heavy-tailed distributions, such as power-law distributions, are widely encountered in real systems and networks. We have introduced growth models that lead to such distributions in Subsections 1.4.1 and 1.4.2. In general, the mechanisms of preferential growth and heterogeneity are coupled, and both play a role in growth processes in real social systems, e.g., in the growth of popularity (see Subsection 1.4.3). Here, we show that groups in Flickr grow according to a model based on heterogeneity.

2.1

Introduction

Many complex systems are characterized by broad distributions capturing, for instance, the frequency of words (Zipf, 1949), the wealth of nations (Pareto, 1896), or the degree distribution of complex networks (Barabási and Albert, 1999) (for more examples in networks, see Subsection 1.2.2). Typically, this feature is explained by means of a preferential growth mechanism. In line with the rich-gets-richer principle, Gibrat's law suggests that the expected growth of a firm, a city, or social activity is proportional to its size (Gibrat, 1931; Gabaix, 1999; Rozenfeld et al., 2008; Rybski et al., 2009). However, less attention has been devoted to the time evolution of complex systems, probably due to the lack of empirical data over time (for some exceptions, see Saichev et al. (2009); Barabási et al. (2002); Palla et al. (2007); Tessone et al. (2011)). In many network growth models, the time unit is mapped to the number of new arriving elements, which makes it difficult to compare the results with real data. Moreover, many models

assume that the elements are born identical, leading to correlations between age and frequency, which are not fully supported by empirical observations (Adamic and Huberman, 2000) (as in Figure 1.8). In many real systems, especially in social systems, individuals or elements are very diverse. In this direction, some models incorporating heterogeneity in the form of fitness, hidden variables, or ranking have been proposed (Caldarelli et al., 2002; Söderberg, 2002; Boguñá and Pastor-Satorras, 2003; Fortunato et al., 2006; Ratkiewicz et al., 2010). However, there is rather little empirical work showing how intrinsic heterogeneity is distributed and its role in complex system growth (Garlaschelli and Loffredo, 2004; De Masi et al., 2006). Based on data collected on a daily basis on the time evolution of an online social system, we characterize the heterogeneity of the groups and identify the heterogeneity and the distributed birth dates as key players explaining the heavy-tailed distribution of group sizes and the apparent proportional growth of groups to their size.

2.2 Dataset

We study groups created and declared by users of Flickr. These groups are mainly used to collaboratively post photos associated with the theme of the group. We consider each group an element of the system characterized by the number of members belonging to the group (group size). We have collected two datasets containing in total over $260,000$ member-created groups in Flickr, which accounted for over 65% of all public groups existing in Flickr. The first dataset has high temporal resolution and a wide time window. It contains $9,503$ groups tracked for 350 days, between June 5, 2008 and May 20, 2009, by the publicly accessible external service called GroupTrackr.[1] The service tracked on a daily basis the number of members of the groups. The second dataset has a shorter time window and minimal temporal resolution, but it covers a larger number of groups. It contains over $260,000$ public groups for which we gathered information on the number of members, collected in two snapshots on December 18, 2009 and January 29, 2010. For these groups, we also gathered estimated information on their birth date. As an estimation of the group birth date we consider the time when the first photo was posted to the group pool, as the first photo is normally posted to the pool soon after the group's creation. The oldest groups in our dataset date back to July 16, 2004.

[1] The Web page of the tool is available at `http://nitens.org/taraborelli/webcommunities`.

2.3

Groups' growth in Flickr

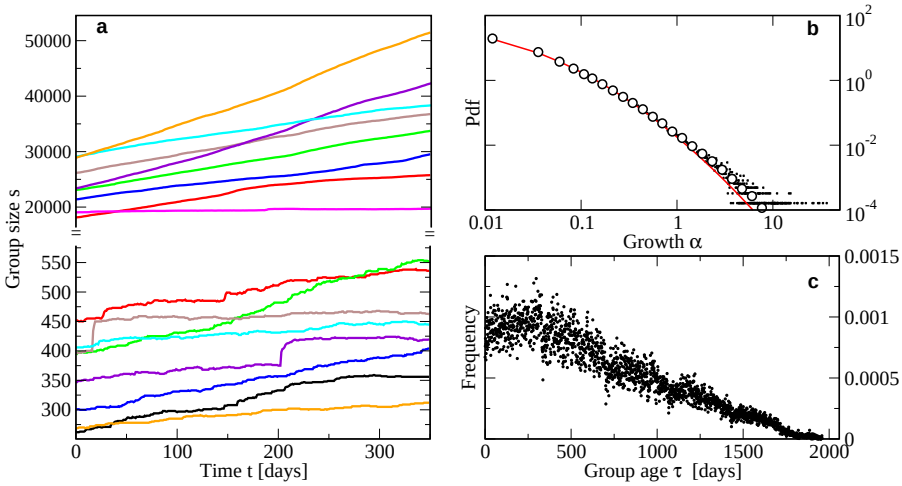


Figure 2.1: Characterizing the time evolution of online groups. (a) Time evolution of the group size for a representative sample of small and large groups. (b) Distributions of groups' growth α (open circles) with fitted log-normal distribution (line). The growth per day α is estimated based on growth over 6 weeks. (c) Distribution of group ages.

We first analyze the time evolution of groups. In Figure 2.1a, we show how typical groups grow in the number of members on a daily basis during a period of one year. As the first approach, linear growth captures the individual trend (despite evident deviations in the form of sudden jumps). We have performed a linear regression of the time evolution of the sizes of $9,503$ groups over a period of one year. For about half of these groups, the coefficient of determination R^2 has a value over 0.95, and more than 80% of the groups larger than $1,000$ have R^2 higher than 0.95. The difference comes from the fact that the larger groups are affected less by fluctuations in size. Aggregated residual plots do not show any clear trend deviating from the linear model. The time series covers a considerable part of the average lifespan of the groups. Thus, we consider that groups grow linearly over time; the size s_g of the group g evolves as

$$s_g = 1 + \alpha_g(t - t_g^0) = 1 + \alpha_g \tau_g \ , \tag{2.1}$$

where α_g is the growth per unit of time, t_g^0 is the birth date, and τ_g is the current age of group g. We estimate the two parameters for $260,000$ groups. The growth α_g for each group g is calculated as the change in its size per day over six weeks. A log-normal distribution provides the best fit to the distribution of growth values α (Figure 2.1b) with an average $\mu = \overline{\ln \alpha} = -3.62$ and the standard deviation $\sigma = 1.57$. Finally, we estimated the current ages of all groups, finding that the number of groups created daily grew (almost linearly) over time (Figure 2.1c).

2.4 Linear growth model with heterogeneous birth and growth

Based on the aforementioned findings, we propose a minimal model of the time evolution of group sizes in Flickr, a linear growth model with heterogeneous birth and growth, which we refer to as the heterogeneous linear growth model. The model proceeds as follows: at each time step t, (i) new groups are created in the system. The number of groups created at each time step increases linearly with t. Each newly created group g starts with one member, and it is assigned its own growth value α_g, drawn from a log-normal distribution. Growth value α_g remains unchanged for the simulation time; (ii) the size of each group g is increased by α_g.

We have performed numerical simulations of the heterogeneous linear growth model such that each time step of the simulation corresponds to a single day. We have simulated $1,959$ days in Flickr, from the moment when the first group from our dataset appeared. As a result of the numerical simulations, we obtain the daily evolution of the sizes of over $260,000$ artificial groups. The distribution of the final sizes of the groups reproduces with good agreement the observed distribution (Figure 2.2a). As shown in Figure 2.2a, there is a small divergence for large group sizes, which can be explained by the deviations, mostly for small groups, from the linear growth assumption. The strong fluctuations of the time evolution of sizes of the small groups (see the jumps in Figure 2.1) lead to a larger apparent growth than the real one, leading to an over-estimation of their growth α; as a consequence, the model displays a larger number of big groups than in the real system.

The average growth of groups of the same size, $\langle \alpha | s \rangle$, shows that bigger groups grow faster (Figure 2.2b) both for the real data and the model in accordance with Gibrat's law: $\langle \alpha | s \rangle \propto s$. This result is obtained even though the microscopic rules of the model do not implement the rich-gets-richer principle. The average growth is an average over all groups of a given size, with each of them growing linearly. Due to the heterogeneity and the linear growth, at a given time, larger groups

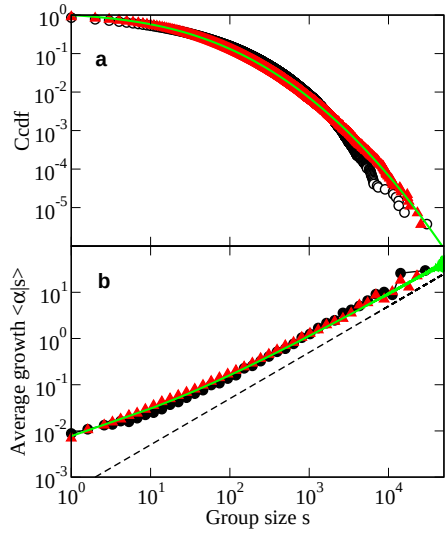


Figure 2.2: The heterogeneous linear growth model vs. real data. (a) Complementary cumulative distribution function of group sizes for the real data (circles), the heterogeneous linear growth model (filled triangles), and its analytical solution (solid line). (b) Average daily growth as a function of the initial size of the groups, estimated for the period of six weeks and averaged over all groups of a given initial size for the real data (circles), the model (triangles), and its numerical solution (line). The dashed line corresponds to the linear behavior $\langle \alpha | s \rangle \sim s$.

consist of old groups that grow slowly and younger groups that grow faster. Thus, the observation of preferential growth for groups of the same size does not reflect in this case an underlying rich-gets-richer principle; rather it is a consequence of the competition of groups with different growth values and ages.

The statistical properties of the model can be estimated analytically. From the definition, the average growth of groups of the same size is given by:

$$\langle \alpha | s \rangle = \frac{\int_{(s-1)/t}^{\infty} \alpha p_{\alpha s}(\alpha, s) d\alpha}{\int_{(s-1)/t}^{\infty} p_{\alpha s}(\alpha, s) d\alpha} \ , \tag{2.2}$$

where $p_{\alpha s}(\alpha, s)$ is the joint probability of having a group of size s and growth rate α, and $\int p_{\alpha s}(\alpha, s) d\alpha dg = 1$. The lower limit of the integral is given by Equation 2.1 and depends on s, and the maximum value of τ is limited to t if the

first group was created at time $t = 0$. We transform Equation 2.2 by replacing the joint probability $p_{\alpha s}(\alpha, s)$ with $p_{\alpha\tau}(\alpha, \tau)$ and making the assumption that τ and α are independent random variables:

$$\langle \alpha | s \rangle = \frac{\int_{(s-1)/t}^{\infty} \alpha p_\alpha(\alpha) p_\tau(\tau(\alpha, s)) \frac{\partial \tau}{\partial s} d\alpha}{\int_{(s-1)/t}^{\infty} p_\alpha(\alpha) p_\tau(\tau(\alpha, s)) \frac{\partial \tau}{\partial s} d\alpha}. \tag{2.3}$$

The numerical solution of Equation 2.3 for log-normal p_α and linear p_τ is plotted in Figure 2.2b. Similarly the distribution of group sizes:

$$p_s(s) = \int_0^{\tau_{max}} p_{s\tau}(s, \tau) d\tau \tag{2.4}$$

$$= \int_0^{\tau_{max}} p_\alpha(\alpha(s, \tau)) p_\tau(\tau) \frac{\partial \alpha}{\partial s} d\tau \ , \tag{2.5}$$

is plotted in Figure 2.2a. As shown, the solutions for both the average growth and the size distribution are in good correspondence with the results of numerical simulations, which indicates that the assumptions of independent random variables and linear growth are reasonable.[2]

2.5 Heterogeneity versus preferential growth

The heterogeneous linear growth model captures the statistical properties that commonly are attributed to the preferential growth mechanism. Thanks to the intrinsic heterogeneity, different growth patterns are permitted, even if groups have the same number of members at any point in time. An example of this effect is in Figure 2.1a, where group sizes cross each other in time, though they continue to grow as they grew before the crossing. To make a direct comparison between the two mechanisms, heterogeneity vs. preferential growth, we consider the Simon model (Simon, 1955). The Simon model was originally proposed to explain the distribution of words' frequency in a written text. At every time step, a word is added to the text: with a given probability q, it is a new word; otherwise, the word is chosen at random from the text, so the words that appear more frequently are chosen more often. We have adapted the Simon model to our system. The values of the parameters are set to obtain the same total number of groups and members as in the real case; in addition, the number of new groups created in the system at each time step of the Simon model grows linearly, to

[2] Equations (2.3) and (2.5) are easy to solve if α and τ are independent random variables and p_α is a power-law distribution. In such a case, $\langle \alpha | s \rangle \propto s$ and that $p_s(s)$ is a power-law as well.

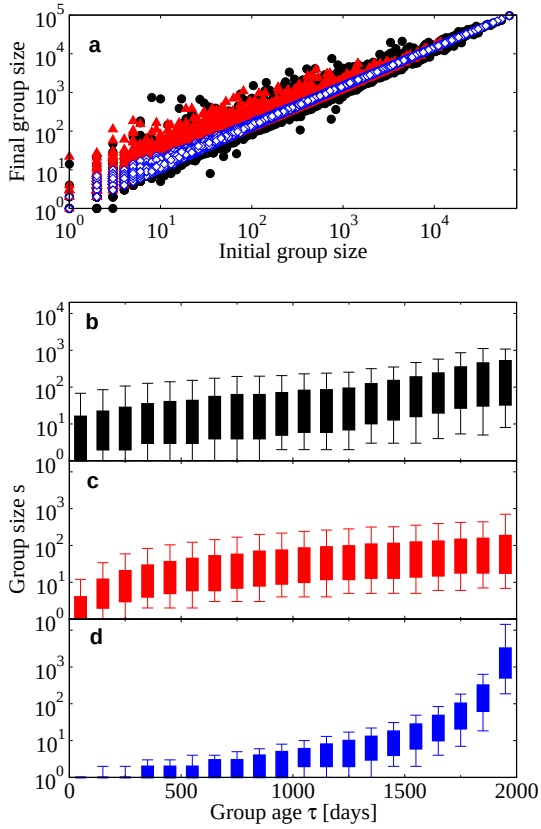


Figure 2.3: Comparison of the Simon and heterogeneous linear growth models vs. real data. (a) Initial and final group sizes over a period of 350 days for the real data (circles), the heterogeneous linear growth models (filled triangles) and the Simon model (diamonds). Each point represents a single group, 9, 503 points are plotted for each set of points. (b-d) Box plots with whiskers at the 9th/91st percentile of the final size of groups as a function of their age at the time of the measurement for 260, 000 groups for (b) the real data, (c) the heterogeneous linear growth model, and (d) the Simon model.

isolate the effect of the heterogeneity. First, in the Simon model, the final size of groups is heavily determined by their initial size measured one year before (Figure 2.3a); thus, there is little heterogeneity among the groups, in contrast to the heterogeneous linear growth model that displays a degree of heterogeneity similar to that of real groups. Second, for the Simon model, the correlation of size and age is strong, while it is weak for real groups and the heterogeneous linear growth model (Figures 2.3b-d).[3] The wide spread of group sizes corresponds to the high heterogeneity of groups, which is not captured by the preferential growth model (as observed in other systems such as, for instance, in the World Wide Web, where the number of links to a page is not strongly correlated with the age of the Web page (see Figure 1.8) (Adamic and Huberman, 2000).

2.6 Conclusions

In this chapter, we have proposed a simple growth model of heterogeneous elements with associated growing counters, based on the findings for a social system in an online community. We found that the model captures many of the features of the real system of online groups, namely the heavy-tailed distribution of group sizes, the average growth proportional to the current size of groups, and the weak correlation between the age and the size of groups.

Furthermore, we made a direct comparison of the heterogeneous linear growth model with a preferential growth model and showed the similarities and the differences between these models. In the heterogeneous linear growth model, the heavy-tailed distribution of the final sizes of elements does not emerge from the growth process itself (e.g., the rich-gets-richer principle) but from the intrinsic heterogeneity of elements that take part in this growth process. This certainly does not answer the question of why some groups grow faster than others, as we do not understand yet what factors influence the fitness of the groups. The simplicity of our approach suggests that the characterization of the heterogeneity may play an important role in understanding the origin of broad distributions and the time evolution of many real systems.

[3] In the heterogeneous linear growth model, the average size of groups of a given age is $\langle s|\tau\rangle = 1 + \tau \exp(\mu + \frac{\sigma^2}{2})$, where μ and σ are parameters of the log-normal distribution. In the Simon model, it is given by $\langle s|\tau\rangle = (\frac{2+mT^2}{2+m(T-\tau)^2})^{1-q}$, where T is the age of the system, m controls the number of new users introduced into the system at each time step (mT), and q is the probability of new group creation within the model (in our case, $T = 1959$, $m = 10$ and $q = 0.014$).

Chapter 3

Strength of intermediary ties in online social networks

In the first chapter, we introduced theories that relate network structure, strength of ties and information diffusion, i.e., Granovetter's theory of strength of weak ties, Burt's theory of structural holes, Aral's diversity-bandwidth theory and the concept of structural folds introduced by Vedres and Stark. Twitter's distinction between different types of interactions allows us to establish a parallelism between online and offline social networks, and to test if the offline theories hold also in the online environment. Here, we demonstrate it by showing that personal interactions are more likely to occur on internal links of groups (the weakness of strong ties), events transmitting information pass preferentially through links connecting different groups (the strength of weak ties), or even more through links connecting to users belonging to several groups that act as brokers (the strength of intermediary ties).

3.1

Introduction

There exists an open discussion on the validity of online interactions as indicators of real social activity (Cummings et al., 2002; Van Dijk, 2006; Watts, 2007; Avnit, 2009; Danon et al., 2005; Vespignani, 2009; Lazer et al., 2009). Most of the OSNs incorporate several types of pairwise interactions that satisfy different user needs and different level of involvement between users (as in Subsection 1.6.3). The cost of establishing the declared links is usually very low. These connections can accumulate and pile up to an extremely large number (as shown in Subsection 1.6.2). If the number of connections increases to the thousands or

the millions, the amount of effort that a user can invest into the relation that each link represents must fall to near zero. Does this mean that declared online connections are irrelevant for understanding social relations or for predicting where higher quality activity (e.g., personal communications, information transmission) is taking place? By analyzing the clusters found in the network of the declared follower links between users of Twitter, we show that even this network bears valuable information on the localization of more personal interactions between users. Furthermore, we show that certain types of users act as brokers of information between groups.

The theory of the strength of weak ties (Granovetter, 1973) deals with the relation between structure, strength of social ties and diffusion of information in offline social networks. It has raised some interest in the last decades (Onnela et al., 2007a; Csermely, 2006; Iribarren and Moro, 2011) and its predictions have been checked in a mobile phone calls dataset (Onnela et al., 2007a). Social networks are usually composed of groups of individuals, connected among them by long range ties known as bridges. Thus, a tie can be internal to a group or a bridge. Granovetter's theory predicts that weak ties act as bridges between groups and are important for the diffusion of new information across the network, while strong ties are usually located at the interior of the groups. Furthermore, Burt's work (Burt, 2005) emphasizes the advantage of connecting different groups to access novel information. More recent works, however, suggest that the people with strong ties in different groups are the most exposed to novel information (Vedres and Stark, 2010) and consider existence of a tradeoff between diversity and bandwidth (Aral and Van Alstyne, 2011) (see Subsection 1.5.3). Due to the different nature of online and offline interactions, it is not clear whether online networks organize following these principles. Our aim in this work is to test if these theories apply also to OSNs.

The study focuses on the two following types of interactions in Twitter. Mentions (tweets containing "@username") are messages that are either directed only to the corresponding user or mentioning the targeted user as relevant to the information expressed to a broader audience. Retweets (tweets containing "RT @username") correspond to content forward with the specified user as the nominal source. In contrast to the normal tweets, mentions usually include personal conversations or references (Honeycutt and Herring, 2009) while retweets are highly relevant for the viral propagation of information (Galuba et al., 2010). Moreover, these interaction types have emerged in Twitter as social conventions (Kooti et al., 2012) and have been later implemented as part of Twitter's sys-

tem.[1] This particular distinction between different types of interactions qualifies Twitter as a perfect system to analyze the relation between topology, strength of social relation and information diffusion in OSNs.

The properties of the follower network have been extensively analyzed especially in relation to its topological structure, propagation of information, homophily, tie formation and decay, etc (Kwak et al., 2010; Mendoza et al., 2010; Ratkiewicz et al., 2011; Asur et al., 2011; Romero and Kleinberg, 2010; Pujol et al., 2010; Borge-Holthoefer et al., 2011). Finding users with thousands or even millions of followers is not exceptional (Avnit, 2009), so the question is whether the structure of the follower network carries any information on where personal relations (mentions) or information transmission events (retweets) take place. To answer this question, we first analyze a sample of the follower network with community detection algorithms and identify a set of groups. Whether the clusters we identify are traces of underlying groups is a question we cannot answer directly due to the lack of ground truth in Twitter. We tackle this problem directly in the next chapter, where we calculate the overlap between detected and declared groups in Flickr. Here, we check the correlation between the location of the personal conversations (mentions) and information diffusion events (retweets) and the structural properties of the links bearing those activities with respect to the detected groups in the network. Note that we consider mentions and retweets to happen always on follower links. This restriction allows us to describe user activity in terms of the detected groups.

3.2 Dataset and preprocessing

Network	Followers	Mentions	Retweets
Users	2,408,534	377,760	26,480
Links	48,776,888	1,224,484	32,169

Table 3.1: Overall characteristics of the follower network and the interactions taking place on it.

The data analyzed in this chapter was collected in a two step process: the first stage corresponds to the collection of the *follower network* (followers and

[1] Blog posts from Twitter developers describing the first implementation to the system of social conventions for mentions/replies and retweets: `http://blog.twitter.com/2009/03/replies-are-now-mentions.html` and `http://blog.twitter.com/2009/08/project-retweet-phase-one.html`.

followees), while the second consists in the retrieval of the user activity from the stream of Twitter (plain tweets, mentions and retweets). In the first stage, the directed unweighted network is obtained from the information on the followers and followees of each user. The data was collected using a breadth-first search technique. Starting from several seeds, followers and followees of the seeds were retrieved. Then, the same procedure was repeated for the newly discovered users obtaining a so-called snowball sampling of the follower network. The procedure is stopped after several steps when the number of newly discovered users in n-th breadth is small compared with the total number of users already discovered in the $(n-1)$-th step. This method tends to detect the users with the highest in or out-degree that belong to the largest connected cluster of the network. The process was run in November 2008, gathering information for a total of $2,408,534$ users.

The second stage consists in searching for all the tweets of the users found in the follower network for a period of time from November 20 to December 11. The activity dataset was constructed from these gathered tweets. The tweets containing usernames with a "@username" functional syntax were used for the mentions. Tweets that were reposted from other users, and which also hold a special format of the form "RT @username", were used to build our retweet dataset. In some cases of mentions and retweets, multiple users were specified. Then, we count only the first user for the purpose of our analysis. It is also worthy to note that at the moment of the data collection mentions and retweets were not yet fully implemented into Twitter system and existed as a social convention. The subset of retweets has been removed from a set of mentions to avoid overlap. In total, we obtained $12,486,784$ tweets from $587,142$ users in the network, what stands for 24% of all users from the follower network. The rest of users either did not posted any tweet in their profile during the period of data collection (80-90% of cases), had a protected profile (5-10% of cases) or removed their profiles (5-10% of cases). Out of these tweets $1,742,956$ where mentions and $46,156$ where retweets. For the purpose of the analysis we have filtered out mentions and retweets which happened without underlying follower relation, in order to avoid inclusion of messages sent to not-known users and also to be able to perform comparisons with our baseline model consisting of the follower network. The resulting set of links with different interactions is summarized in Table 3.1. Note that links with mentions/retweets can have multiple mentions/retweets happening over them.

The dataset is a good representation of what Twitter was at the end of 2008 both in the social network and in the activity of the users. At the time of the data collection Twitter had almost 5 million registered users.[2] Therefore we estimate that our dataset contains information about around 50% of the most active users from that time. Additionally, in order to test if our results are independent on

[2] According to `http://bit.ly/1cLMwxa`.

the sample of the network we repeat our analysis on various subsamples of the collected follower network arriving to the same conclusions (see Appendix B). Other aspects of this dataset related to system scalability and trace generation were studied by (Pujol et al., 2010, 2009; Erramilli et al., 2011).

3.3 Description of the groups

Our first step is to identify the groups in the follower network. Due to the size, density and directness of the follower network, and in order to capture the possible inclusion of users in multiple groups or in none, we have used OSLOM (Lancichinetti et al., 2011, 2010) (see Appendix I of the dissertation). The analysis has also been performed with other clustering techniques (Rosvall and Bergstrom, 2008; McDaid and Hurley, 2010; Raghavan et al., 2007; Blondel et al., 2008)[3], reaching similar conclusions (see Appendix B for a detailed account on these results). We have detected 92,062 groups at the lowest hierarchical level, three of which are graphically depicted in Figure 3.1A with each sphere corresponding to a single user. In general, the links can be classified according to their position with respect to the user groups: internal, between groups, intermediary, and links involving nodes not assigned to any group as shown in Figure 3.1B. Note the correspondence of the intermediary position to the concept of structural fold (compare Figure 3.1B and Figure 1.10, respectively).

The statistics characterizing the groups and links are displayed in Figure 3.2. The group size distribution decays slowly for three orders of magnitude and does not show a characteristic group size (Figure 3.2A). For instance, the largest group contains around 10,000 users. Also the number of groups each user belongs to shows high heterogeneity: 37.4% of the users have not been allocated to any group, while there exists a user belonging to more than 100 groups (see Figure 3.2B). The percentage of links falling in the different types regarding the groups is depicted in Figure 3.2C. Although the non-classified users are 37% of the total, the links connected to them are less than 6% and the percentage is even lower for those with mentions or retweets. The most common type of connections is the between-group links. One may wonder if the algorithm for clusters detection is doing a good job when there is such a large proportion of between-group links. The clustering method is trying to find groups of mutually interconnected nodes that would be extremely rare in a randomized instance of the network, rather than optimizing the ratio between number of between-group and internal links. In Appendix A, this argument is further developed and the capacity of

[3] We also used an implementation of modularity optimization available online at `http://deim.urv.cat/~aarenas/data/welcome.htm`

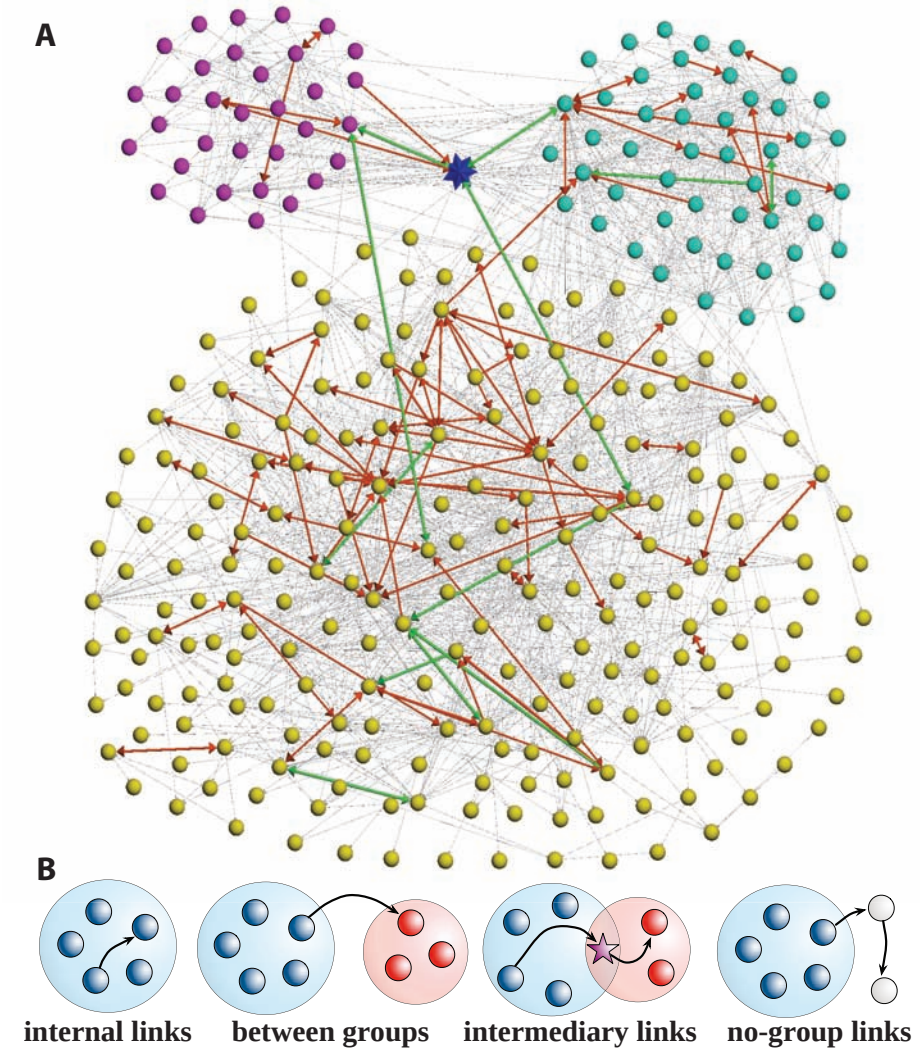


Figure 3.1: Groups and links. (A) Sample of Twitter network: nodes represent users and links, interactions. The follower connections are plotted as gray arrows, mentions in red, and retweets in green. The width of the arrows is proportional to the number of times that the link has been used for mentions. We display three groups (yellow, purple and turquoise) and a user (blue star) belonging to two groups. (B) Different types of links depending on their position with respect to the groups' structure: internal, between groups, intermediary links, and no-group links.

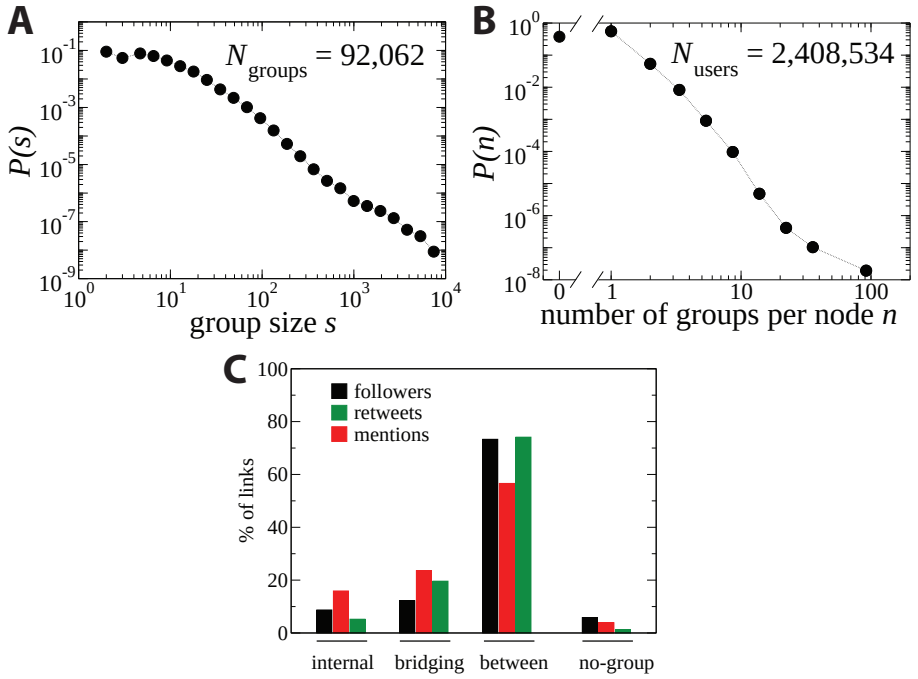


Figure 3.2: Group and link statistics. (A) Size distribution of the group. (B) Distribution of the number of groups to which each user is assigned. (C) Percentage of links of different types, e.g., follower links (black bars), links with mentions (red bars) or retweets (green bars), staying in particular topological localizations in respect to detected groups.

OSLOM to detect planted communities is proved in a benchmark even in situations with a high ratio between the number of between-groups and internal links. Another relevant point to highlight is the different potential of each type of links to carry mentions and retweets. As it can be seen in Figure 3.2C, the red bars for mentions in internal links and intermediary links almost double the abundance of links in the follower network in these categories. The links between groups, on the other hand, attract far less mentions.

3.4

The strength of ties

Besides their location with respect to the groups, the links can be also characterized by their intensity. In Twitter mentions are typically used for personal communication, which establishes a parallelism between links with mentions and strength of social ties. The more mentions has been exchanged between two users, even more so if reciprocated, the stronger we consider the tie between them. We define intensity of a link as the number of mentions interchanged on it. Different predictors have been considered to estimate social tie strength (Marsden and Campbell, 1984) including, for instance, time spent together (Marsden and Campbell, 1984) or the duration of phone calls (Onnela et al., 2007a). We consider the intensity as an approximation to social strength given that writing a mention involves some effort and addresses only single targeted users.

3.5

Internal links

According to Granovetter's theory, one could expect the internal connections inside a group to bear closer relations. Unfortunately, there is no means to measure the closeness of a user-user relation in a sociological sense in our Twitter dataset. However we can verify whether the link has been used for mentions, whether the interchange has been reciprocated or whether it has happened more than once. We define the fraction f_p^i of links with interaction i in position p with respect to the groups of size s as

$$f_p^i(s) = \frac{L_p^i(s)}{L^i}, \tag{3.1}$$

where $L_p^i(s)$ is the number of links with that type of interaction in position p with respect to the groups of size s and L^i in the total number of links with interaction i. The fractions $f_{\text{internal}}^i(s)$ reveal an interesting pattern as function of the group size (Figure 3.3A). Note that the fraction of links in the follower network (black curve) is taken as the reference for comparison. Links with mentions are more abundant as internal links than the baseline follower relations for groups of size up to 150 users. This particular value brings reminiscences of the Dunbar number (Dunbar, 1992), the cognitive limit to the number of close relationships that has recently been discussed in the context of Twitter (Gonçalves et al., 2011). Although we have identified larger groups, the density of mentions is similar to the density of links in the follower network. In addition, the distribution of

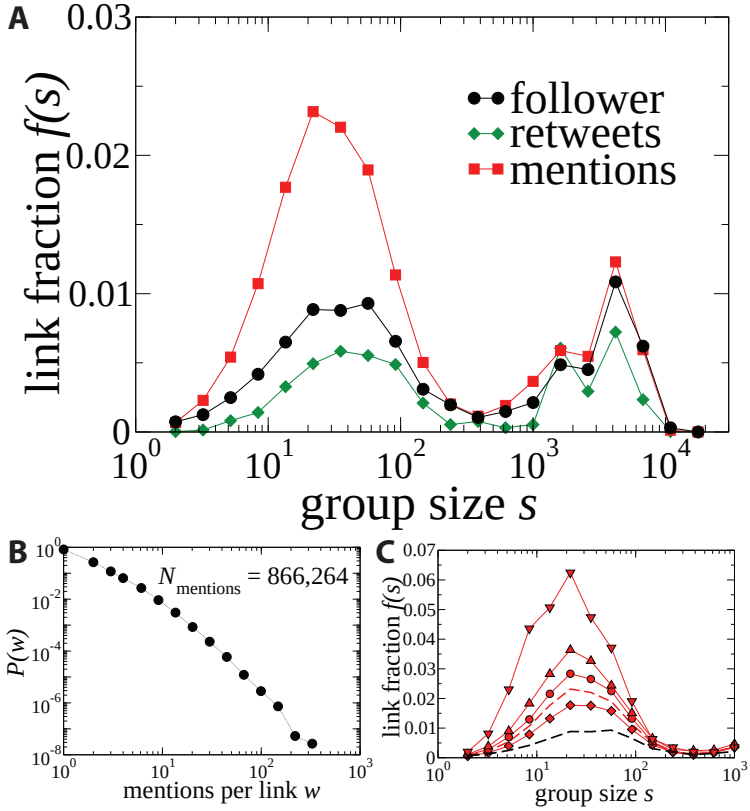


Figure 3.3: Internal activity. (A) Fraction f of internal links as a function of the group size in number of users. The curve for the follower network acts as baseline for mentions and retweets. Note that if mentions/retweets were randomly appearing over follower links then the red/green curve should match the black curve. (B) Distribution of the number of mentions per link. (C) Fraction of links with mentions as a function of their intensity. The dashed curves are the total for the follower network (black) and for the links with mentions (red), while the other curves correspond (from bottom to top) to fractions of links with: 1 non-reciprocated mention (diamonds), 3 mentions (circles), 6 mentions (triangle up) and more than 6 reciprocated mentions (triangle down).

the number of times that a link is used (intensity) for mentions is wide, which allows for a systematic study of the dependence of intensity and position (see Figure 3.3B). The more intense (or reciprocated) a link with mentions is, the more likely it becomes to find this link as internal (Figure 3.3C). This corresponds to Granovetter expectation that the stronger the tie is the higher number of mutual contacts of both parties it has and the higher the chance that the parties belong to the same group. Similar trends are found using other clustering algorithms (see Appendix B [Figure 3.7-3.9]).

3.6 Links between groups

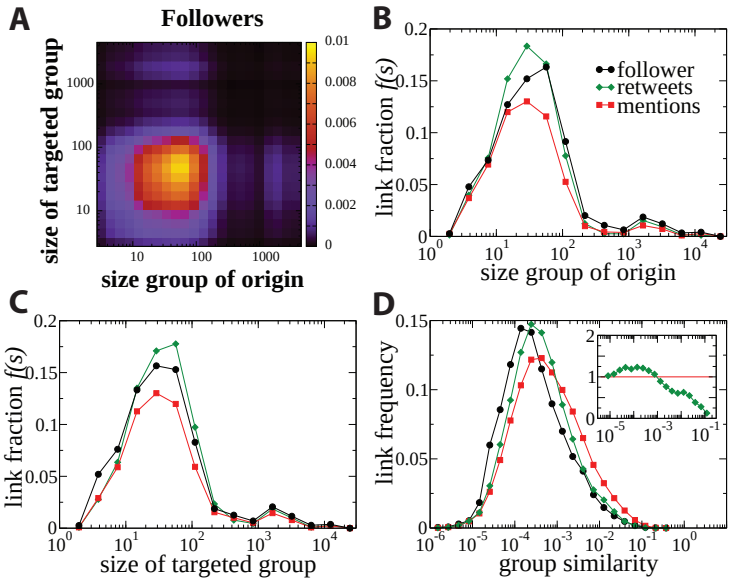


Figure 3.4: Group-group activity. (A) Distribution of the number of links in the follower network between groups as a function of the size of the groups. (B) Fractions f of links of the different types (follower, with mentions and with retweets) as a function of the size of the group at the link origin, and (C) at the targeted group. (D) Frequency of between-group links as a function of the group-group similarity for the different type of links. In the inset, ratio between the frequency of links with retweets and with mentions.

The next question to consider is the characteristics of links between groups. These links occur mainly between groups containing less than 200 users (Figure 3.4A-C). However, their frequency depends on the quality of the links (if they bear mentions or retweets). While links with mentions are less abundant than the baseline, those with retweets are slightly more abundant. According to the strength of weak ties theory (Granovetter, 1973; Onnela et al., 2007a; Iribarren and Moro, 2011; Burt, 2005), weak links are typically connections between persons not sharing neighbors, being important to keep the network connected and for information diffusion. We investigate whether the links between groups play a similar role in the online network as information transmitters. The actions more related to information diffusion are retweets (Galuba et al., 2010) that show a slight preference for occurring on between-group links (Figures 3.4B and 3.4C). This preference is enhanced when the similarity between connected groups is taken into account. We define the similarity between two groups, A and B, in terms of the Jaccard index of their connections: The similarity is the overlap between the groups' connections and it estimates network proximity of the groups. The general pattern is that links with mentions more likely occur between close groups and retweets occur between groups with medium similarity (Figure 3.4D). Mentions as personal messages are typically exchanged between users with similar environments, what is predicted by the strength of weak ties theory. Links with retweets are related to information transfer and the similarity of the groups between which they take place should be small according to the Granovetter's theory. The results show that the most likely to attract retweets are the links connecting groups that are neither too close nor too far. This can be explained with Aral's theory about the trade-off between diversity and bandwidth: if the two groups are too close there is no enough diversity in the information, while if the groups are too far the communication is poor. These trends are not dependant on the size of the considered groups (see Appendix B [Figure 3.10]).

3.7 Intermediary links

The communication between groups can take place in two ways: the information can propagate by means of links between groups or by passing through an intermediary user belonging to more than one group. We have defined as intermediary the links connecting a pair of users sharing a common group and with at least one of the users belonging also to a different group (see Figure 3.1B). These users and their links have a high potential to pass information from one group to another in an efficient way (Csermely, 2006). Several previous works pointed out

to the existence of special users in Twitter regarding the communication in the network (Asur et al., 2011; Wu et al., 2011). In order to estimate the efficiency of the different types of links as attractors of mentions and retweets, we measure a ratio r_p^i for links in position p and for interaction i defined as

$$r_p^i = \frac{L_p^i}{L_p}, \tag{3.2}$$

where, as before, L_p^i is the number of links with the interaction i in position p and L_p is the total number of links in that position. The bar plot with the values of r_p^i is displayed in Figure 3.5A. We compare how well the different types of links attract mentions (red bars) and retweets (green bars). Links internal to the groups attract more mentions and less retweets than links between groups in agreement with the predictions of the strength of weak ties theory. Intermediary links attract mentions as likely as internal links: the fraction of intermediary links with mentions is very close to the fraction of internal links with mentions. This is expected because intermediary links are also internal to the groups. However, the aspect that differentiates more intermediary links from other type of links is the way that they attract retweets. Intermediary links bear retweets with a higher likelihood than either internal or between-groups connections (see Figure 3.5). This fact can be interpreted within the framework of the tradeoff between diversity and bandwidth (Aral and Van Alstyne, 2011): strong ties are expected to be internal to the groups and to have high bandwidth, while ties connecting diverse environments or groups are more likely to propagate new information. High bandwidth links in our case correspond to those with multiple mentions, while links providing large diversity are the ones between groups. Intermediary links exhibit these two features: they are internal to the groups and statistically bear more mentions, and introduce diversity through the intermediary user membership in several groups. Although some theoretical works (Granovetter, 1973; Aral and Van Alstyne, 2011) suggest that ties with high bandwidth and high diversity should be scarce, we find that intermediary links are as abundant as internal links (see Figure 3.2C). Similar results have been found in sociological studies of offline entrepreneurial groups (Vedres and Stark, 2010). The authors recognize the importance of individuals belonging to multiple groups (see Figure 1.10). Their concept of *structural fold* is strikingly similar to our concept of intermediary individual (at the time of our study we were unaware of the Vedres' and Stark's study). Moreover, in line with the theories (Granovetter, 1973; Burt, 2005; Aral and Van Alstyne, 2011), higher diversity increases the chances for a link to bear retweets (Figure 3.5B), which implies a more efficient information flow. The number of non-shared groups assigned to the users connected by the link positively correlates with a higher than expected number of retweets (results with another clustering algorithm are presented in Appendix B [Figure 3.11]).

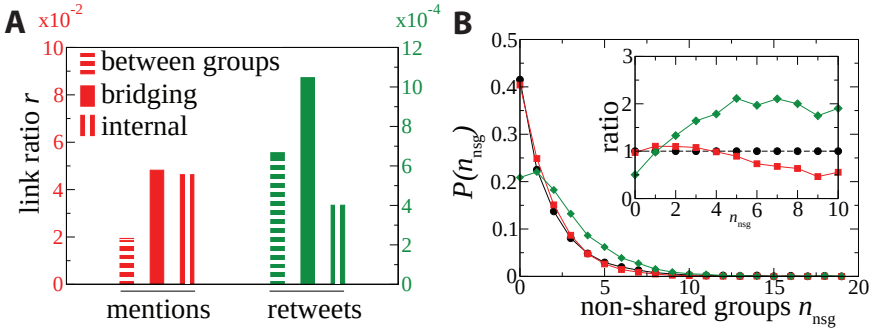


Figure 3.5: Intermediary links. (A) Ratio r between the number of links with mentions or retweets and number of follower links. (B) Distribution of the links in the follower network (black curve), those with mentions (red curve) and retweets (green curve) as a function of the number of non-shared groups of the users connected by the link. Inset, ratios between these distributions and the follower network.

3.8 Conclusions

In summary, we have found groups of users analyzing the follower network of Twitter with clustering techniques. The activity in the network in terms of mentions and retweets clearly correlates with the landscape that the presence of the groups introduces in the network. Mentions, which are supposed to be more personal messages, tend to concentrate inside the groups or on links connecting close groups. This effect is stronger the larger the number of mentions exchanged and if they are reciprocated. Retweets, which are associated to information propagation events, appear with higher probability in links between groups, especially those that connect groups that do not show a high overlap, and more importantly on links connected to users who intermediate between groups. These intermediary users belong to multiple groups and play an important role in the spreading of information. They acquire information in one group and launch retweets targeting the other groups, which they are members of. At the same time, the access to new information can transform them into attractive targets to be retweeted by their followers. Our method provides a way to identify these special users as brokers of information between different groups using as only input the follower network.

From the sociological point of view, the way that the activity localizes with respect to the groups allows us to establish a parallelism with the organization

of offline social networks. In particular, the theory of the strength of weak ties proposed by Granovetter to characterize offline social network applies also to an online network. Furthermore, some of our results can be explained within the framework of Burt's brokerage and closure and Aral's diversity-bandwidth trade-off theories. The specific properties of Twitter offer an opportunity to study directly the importance of the links for personal communications or for information diffusion. According to these theories, the strong social ties tend to appear at the interior of the groups or between close groups as it happens for the links with mentions in Twitter. In addition, the socially weak ties are expected to be more common connecting different groups and to be important for the propagation of information in the network. This is similar to what we observe for the links with retweets that concentrate with high probability in links between dissimilar groups or in intermediary links. Besides the roles assigned by these two theories to the links, we have found that intermediary users and links are also an important component to take into account for understanding information propagation, in accordance with the theory of structural folds of Vedres and Stark. The intermediary links tend to be characterized by high bandwidth and diversity in the context of Aral's study, and exhibit high information diffusion efficiency. Based on all these findings, despite the myth of one million friends and the doubts on the social validity of online links, the simplest connections of the online network bear valuable information on where higher quality interactions take place.

Appendix A: Balance between the number of internal links and links between groups

In this section we discuss in more detail the imbalance between the number of internal and between-group links that is seen in Figure 3.2C. The objective of a clustering algorithm is to detect areas of the network with high number of connections. How is it possible that the overall number of internal links is lower than that of links between groups? The answer is that OSLOM, as many other community detection algorithms, is not attempting to optimize the balance between internal and between-group links in a straightforward manner. The method searches for areas denser in internal connections than the baseline for random graphs obtained by reshuffling the links of the original network while maintaining the node's in and out-degrees (see Appendix I of the dissertation).

To illustrate this idea, we have generated a benchmark graph formed by N_c cliques (i.e., fully connected subgraphs) of size S_c each. We add to it L_{bet} between-group links connecting nodes of different cliques at random. To quan-

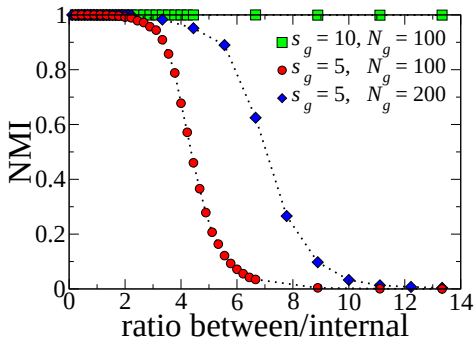


Figure 3.6: Normalized mutual information as a function of the ratio between the number of links between groups and internal links in a benchmark. The benchmark is composed of N_c cliques (fully connected subgraphs) of size S_c each.

tify the level of similarity between the original cliques and the groups detected by OSLOM, we use the normalized mutual information between partitions NMI (Danon et al., 2005; Lancichinetti et al., 2009). This quantity is equal to one when the two partitions of the network in groups, i.e., the original cliques and the groups detected by OSLOM, are identical. It tends to zero when there is no relation between the groups. The value of normalized mutual information is shown in Figure 3.6 as a function of the ratio between the number of links between groups and the number of internal links $L_{\rm int} = N_{\rm c}\, S_{\rm c}\, (S_{\rm c} - 1)/2$. OSLOM is able to detect the planted cliques even for high values of the ratio, higher in any terms than the values seen in the follower network that we study. Note the performance of the method improves with larger groups and with a larger number of cliques.

The connections between groups are introduced at random, without any clear statistical preference for connections between two particular groups. OSLOM detects these random links and ignores them to evaluate which nodes belong to each group despite the high ratio of between-groups link to internal links. Only if a systematic bias existed in the connections between certain groups, OSLOM would detect the groups as a single group.

Appendix B: Results with other clustering algorithms

Here, we check the reliability of the results presented in the main text by repeating the analysis for clustering algorithms different than OSLOM for various network

samples.

The reasons to select OSLOM as the main method are the following (i) the source code is publicly available, (ii) the method is able to analyze the full directed follower network in reasonable amount of time, (iii) it detects the overlapping communities and nodes not belonging to any group, and (iv) the clusters obtained are statistically significant according to a clear null model (as presented in Appendix I of the dissertation) (Lancichinetti et al., 2011)[4]. The other algorithms that we test here meet the following minimal conditions: (i) the methods are available online in the form of software tools; (ii) they are able to deal with relatively large samples of dense graphs in a reasonable amount of time. We have found several methods satisfying these conditions and we show in the remainder results for groups detected by:

1. Infomap (Rosvall and Bergstrom, 2008, 2011)[5],
2. Moses (McDaid and Hurley, 2010)[6],
3. A message-passing algorithm proposed by Raghavan et al (Raghavan et al., 2007; Leung et al., 2009) that we refer to as "Real-time" method,
4. Louvain method for community detection based on modularity optimization (Blondel et al., 2008)[7].

We apply each of the clustering algorithms to three samples of the follower network: the full network, the snowball sample of the network (i.e., all nodes and all directed connections between them within 3 neighbors from a random seed in the symmetrized version of the network) and the full network with all hubs removed (i.e., nodes having more than a thousand of followers).

Internal links

The links with mentions are more abundant than the follower links for groups of size up to 150 users. Larger groups do not behave in the same way and the fraction of links with mentions falls to the baseline. We find a similar signal when the groups are extracted with other clustering algorithms, see the summary in Table 3.2. The results for the full network for 2 out of 3 algorithms tested are in qualitative correspondence with the results of OSLOM (see Figure 3.7). Furthermore, Infomap's results show the signal for all the network samples (see Figures 3.7 to 3.9). OSLOM and Infomap are supposed to be one of

[4] Source code available at `http://www.oslom.org`.

[5] Source code available at `http://www.tp.umu.se/~rosvall/code.html`.

[6] Source code available at `http://clique.ucd.ie/moses`.

[7] Implementations available at `https://sites.google.com/site/findcommunities/`.

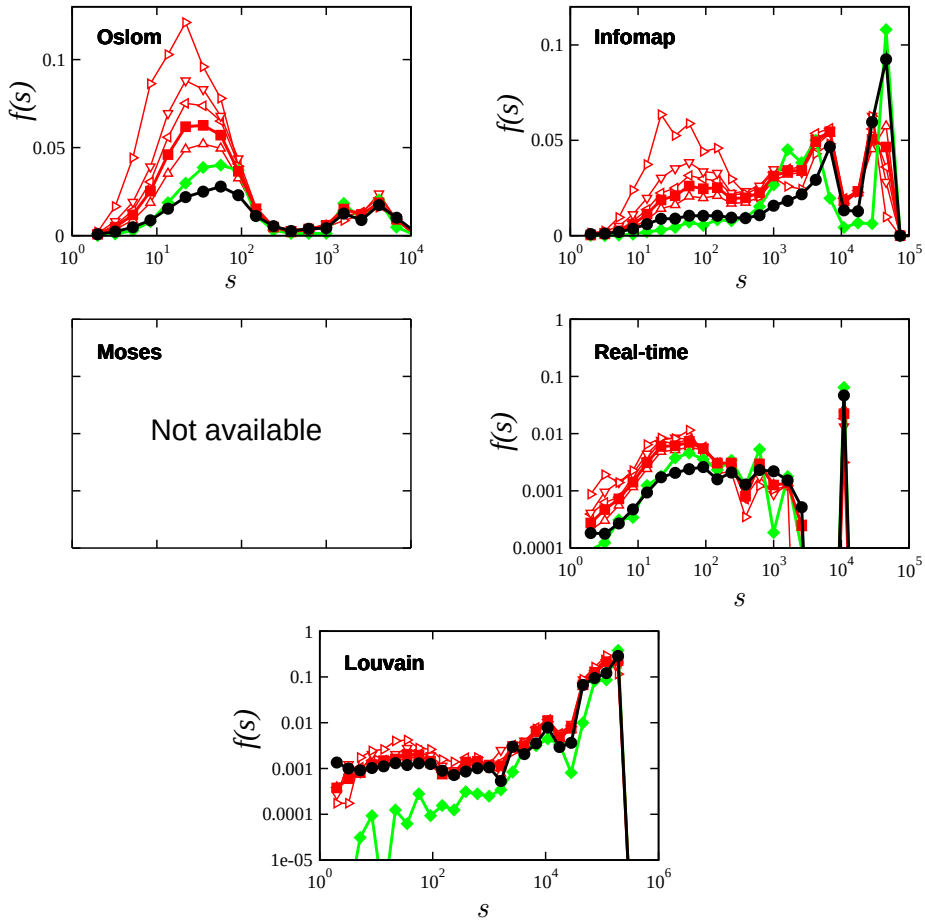


Figure 3.7: Internal activity for different clustering algorithms. Fraction f of internal links as a function of the group size in number of users. The structure of the figure reproduces Figure 3.3. The curve for the follower network (black circles) acts as baseline for mentions (red squares) and retweets (green diamonds). Other curves correspond (from bottom to top) to fractions of links with: 1 non-reciprocated mention (triangles up), 3 mentions (triangles left), 6 mentions (triangle down) and more than 6 reciprocated mentions (triangle right). Note that if mentions/retweets were randomly appearing over follower links then the red/green curves should match the black curve.

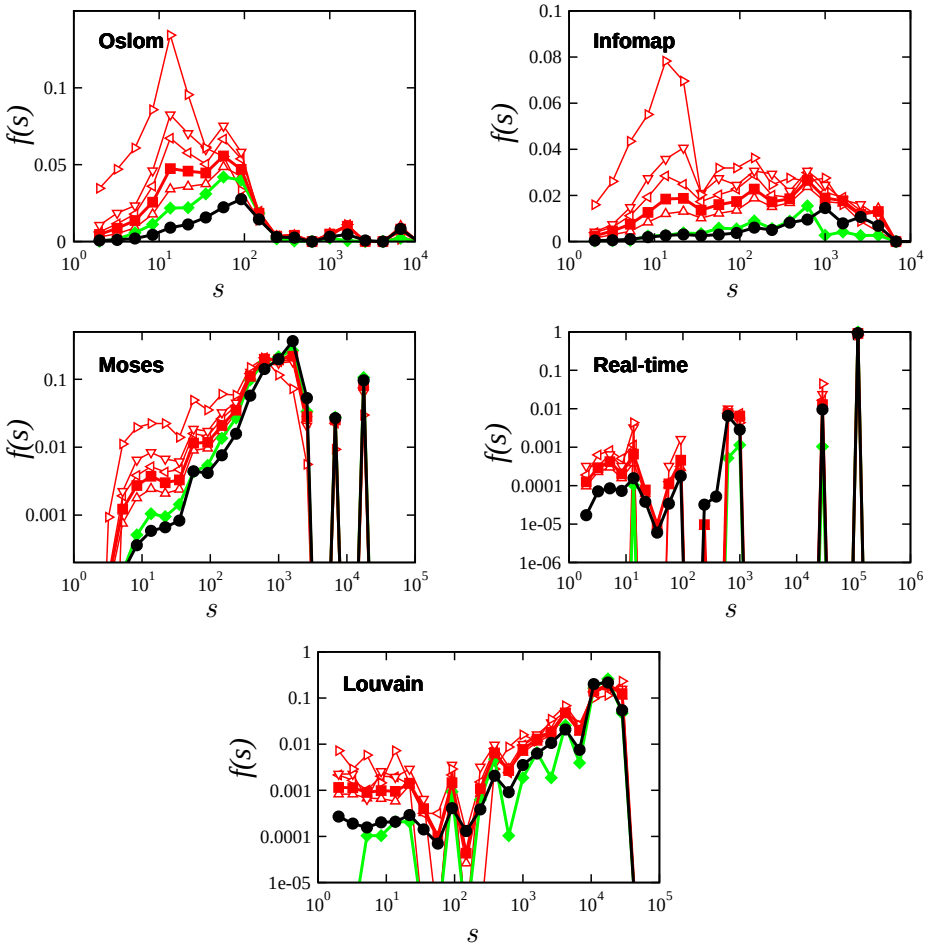


Figure 3.8: Internal activity for different clustering algorithms for the snowball sample of the network. Fraction f of internal links as a function of the group size in number of users. The structure of the figure reproduces Figure 3.7.

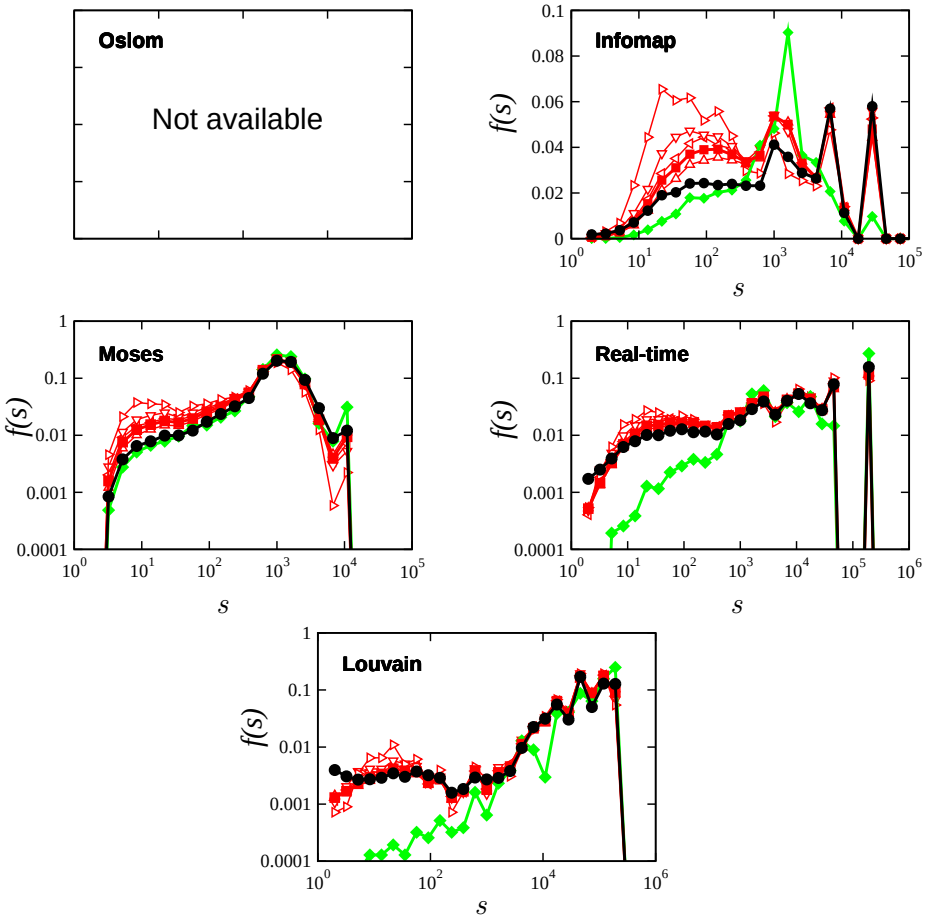


Figure 3.9: Internal activity for different clustering algorithms for the sample of the network without hubs. Fraction f of internal links as a function of the group size in number of users. The structure of the figure reproduces Figure 3.7.

Network sample	Nodes	Edges	OSLOM	Infomap	Moses	Real-time	Louvain	Figure
Whole network	2,408,534	48,776,888	✔	✔	-	✔	✗	3.7
Snowball 3 hops	175,078	10,356,020	✔	✔	✔	✗	✔	3.8
No hubs	2,395,415	23,404,103	-	✔	✔	✗	✗	3.9

Table 3.2: Summary of the results for different clustering algorithms and various samples of the network. We evaluate the trend of links with mentions to concentrate inside groups. A hyphen is inserted if the results are not available, i.e., the clustering algorithm has crushed or has been running for a long time without finishing.

the most trustworthy methods for community detection (Lancichinetti and Fortunato, 2009b; Lancichinetti et al., 2011). The figures reveal that the fraction of links with mentions inside of groups is higher than the fraction of any links inside groups irrespectively of the algorithm and the network sample used. In case of all clustering algorithms, the effect is not visible for groups larger than 100-5,000 users (this number varies for different algorithms). Finally, taking into account the number of mentions and whether they are reciprocated, the results show a remarkably consistent pattern. The more mentions, especially reciprocated, the link has the higher the probability that it is inside of a small group, independently on the community detection algorithm used or the sample of the network considered (see Figures 3.7 to 3.9).

Links between groups

In order to check whether the interaction localization patterns in the links between groups (discussed in Figure 3.4) can be reproduced with groups obtained by other clustering methods, we repeat the analysis of the group-group links using the groups found by Infomap. We present the results in Figure 3.10. Even though shape of some of the curves is different, the main qualitative results confirm the trends observed with OSLOM. Mentions appear less often between groups and tend to concentrate in the links between similar groups, while retweets appear more often between groups and tend to concentrate in the links connecting groups with medium or low similarity. The difference in similarity is observed irrespectively of the size of the groups of origin and destination of links (compare Figures 3.10E and 3.10F)

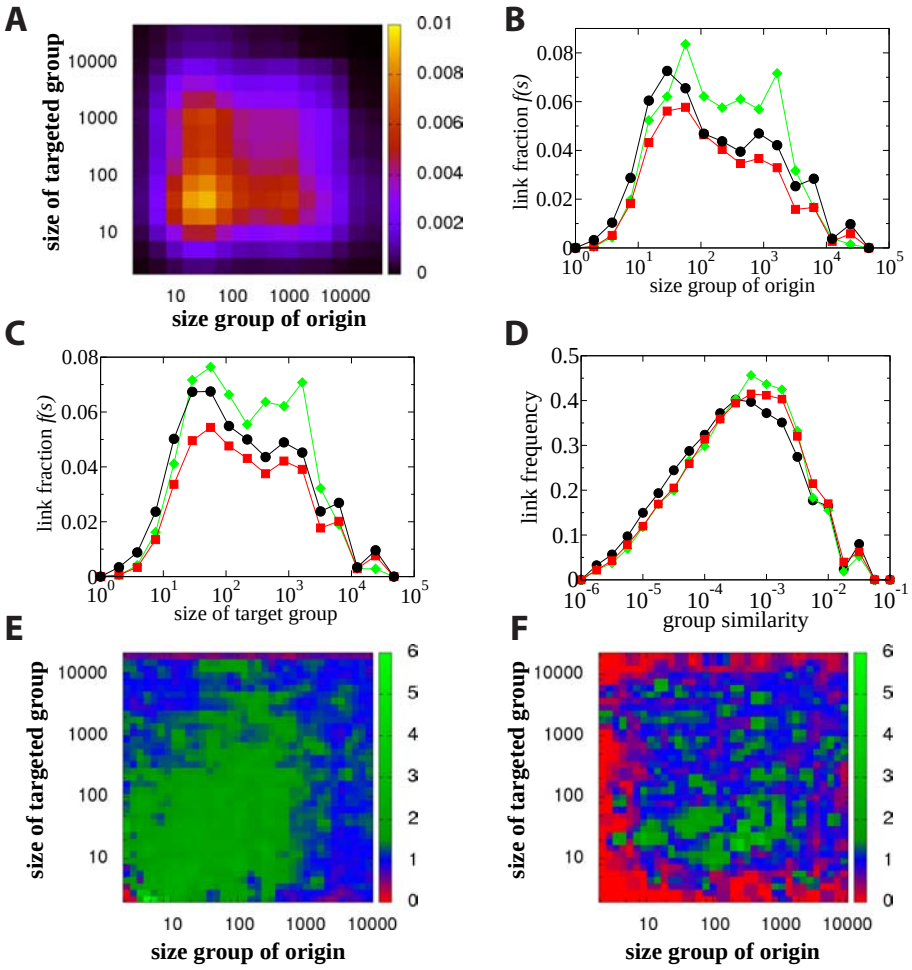


Figure 3.10: Group-group activity for Infomap for the network sample without hubs. The structure of the figure reproduces Figure 3.4. (A) Distribution of the number of links in the follower network between groups as a function of the size of the groups. (B) Fractions f of links of the different types (follower, with mentions and with retweets) as a function of the size of the group at the link origin, and (C) at the targeted group. (D) Frequency of between-group links as a function of the group-group similarity for the different type of links. (E) Ratio between the average group similarity for the between-group links with mentions, or (F) retweets, and the follower network as function of the size of the groups of origin and destination.

Intermediary links

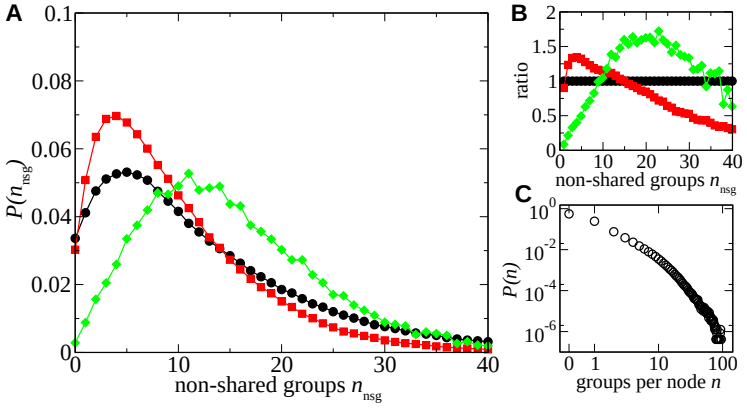


Figure 3.11: Intermediary links for Moses for the network sample without hubs. (A) Distribution of the links in the follower network (black curve), those with mentions (red curve) and retweets (green curve) as a function of the number of non-shared groups of the users connected by the link. (B) Ratios between these distributions and the follower network. (C) Distribution of the number of groups to which each user is assigned

The role of the intermediary users and the intermediary connections can only be investigated with clustering algorithms capable of detecting overlapping communities, and so capable of assigning nodes to more than one group, i.e., with OSLOM and Moses. The distributions of number of groups a user belongs to for the two clustering algorithms are shown in Figures 3.2B and 3.11C, respectively. The results of the analysis of the intermediary links for OSLOM are presented for the full network in Figure 3.5. Here we present the results for the Moses algorithm for the sample of the network with removed hubs (see Figure 3.11). The shape of the curves, especially for the ratios of distributions for different types of links, is consistent for the two clustering algorithms. The probability of having a retweet over an intermediary link steadily grows with the number of non-shared groups by the interacting users until it reaches a maximum and then decreases (see inset of Figure 3.5C and Figure 3.11B).

Appendix C: An alternative procedure to validate our results

The Granovetter's theory of the strength of weak ties has two formulations. The macroscopic formulation predicts strong ties to be inside of communities, whereas weak ties as the bridges between these communities. The microscopic formulation states that strong ties happen between users having many friends in common, and vice versa. Since the main text focuses on the communities here we also show that in fact the microscopic prediction takes place in this social system as well.

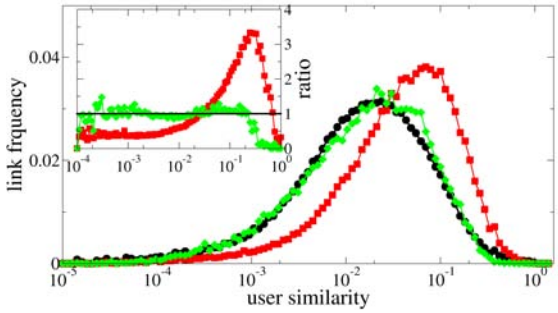


Figure 3.12: Distribution of users' similarity for pairs of users connected by a follower link (black circles), by a link with a mention (red squares) or by a link with retweet (green diamonds). Inset: ratio between these distributions taking the follower network as a baseline.

Jaccard similarity of two users is equal to number of shared followers by the two users divided by total number of unique followers the two users have. In Figure 3.12 we plot the distribution of the similarity for pairs of users connected with a follower link with or without mention and retweet interactions. The distribution of similarity for the links with mentions is shifted to the right, showing that indeed mentions tend to happen between users who share contacts. This result and the finding that mentions are more abundant inside of communities are consistent with the expectations of the Granovetter's theory.

CHAPTER 3. STRENGTH OF INTERMEDIARY TIES IN ONLINE SOCIAL NETWORKS

Chapter 4

Predicting types of groups based on identity and bond theories

In this chapter, we use network theory to develop metrics that quantify some of the characteristics of identity-based and bond-based groups, which was characterized in Subsection 1.5.4. We show that the metrics can be used to make accurate predictions of group type with statistical classifiers. In this study we characterize and compare two different sets of groups, i.e., declared and declared groups in Flickr (see Subsection 1.6.5). We measure their overlap as well as their social and topical properties.

4.1

Introduction

The theories of common identity and common bond (Prentice et al., 1994) about the creation of social communities affirm that people join groups driven by either the interest in the group as a whole or by strong personal ties with other members, respectively. As a result, depending on the prevalent motivation of members, spontaneously generated groups can be categorized as either topical or social. The theories assume that the two types of groups have distinct and well recognizable traits that characterize them (see Subsection 1.5.4).

In recent years, the theories have been widely commented and elaborated by social scientists from a theoretical perspective and through small-scale experiments both in online and offline settings (Sassenberg, 2002; Utz and Sassenberg, 2002; Ren et al., 2007), but a validation over large-scale datasets together with the development of rigorous, automated methodologies to distinguish the group types is missing. Indeed, the availability of big data from OSNs provides the op-

portunity to study the dynamics of the online communities from a data-mining perspective (Mislove et al., 2007; Negoescu and Perez, 2008; Kairam et al., 2012). None of those experiments, however, have been directly aimed at verifying the common identity and common bond theory.

In this chapter, we contribute to fill this gap by proposing a set of general metrics based on the theory. We establish the ground truth in an editorial process by labeling groups as topical or social. We show that the metrics' values computed on a large corpus of groups extracted from Flickr confirm the cardinal points of the theory are indeed good predictors of the group type. In addition, we repeat the same analysis on groups identified by a graph-based community detection algorithm. This allows us to compare the user-generated communities to the automatically detected ones not only from a structural perspective but also along the dimensions of sociality and topicality. Since community detection techniques have been largely employed in recent years to describe the structure of complex social systems (Fortunato, 2010), the need for a clearer assessment of the meaning of the detected clusters has been often expressed from different angles (Lancichinetti and Fortunato, 2009a; Yang and Leskovec, 2012), but never completely satisfied. Our study also contributes to shed light on this matter.

To the best of our knowledge, this is the first attempt of formalization of the common identity and common bond theory, and of its validation over a large and diverse set of user communities. The obtained results open a new perspective on the semantic interpretation of implicit and explicit user groups in OSNs.

Our main contributions are summarized as follows:

- Translation of the common identity and common bond theory into general metrics applicable to social graphs. An insightful characterization of a large group dataset from Flickr is performed using the proposed metrics.

- Comparison between user-declared groups and groups discovered by a community detection algorithm, both in terms of their overlap and their properties of sociality and topicality.

- Design of a method to predict whether a group is social or topical, based on the defined metrics. Prediction on the user-generated groups from the Flickr data yields surprisingly good results.

For the prediction task we perform statistical classification that is introduced to the reader in the following subsection. Related work is described in the subsequent subsection. The description of the metrics, the comparison between declared and detected groups, the detailed description of the prediction method and its results are presented in the next sections.

4.1.1 Statistical classification

In machine learning and statistics, classification is a problem of identifying to what category an item belongs, on the basis of a set of other items. The method that performs the classification task is known as a classifier. A wide variety of methods has been proposed to solve this task (Caruana and Niculescu-Mizil, 2006). Here, we introduce the problem and describe briefly how to evaluate accuracy of results of a classifier.

In classification task each item is represented by a set of features with integer, real or categorical values. These features are known for all the items. The classifier can be understood as a mathematical function that maps features of an item to a category. This function can be learned in a supervised or unsupervised fashion. In supervised learning the function is learned from a training set of items, whose category, also known as class, is given explicitly. In unsupervised learning the function is implied a priori without any training set, usually through a cost function in clustering algorithms. The simplest supervised classifiers are based on linear or logistic regressions. The category is known only for the training set. The classifier learns the prediction model based on the training set and tries to make accurate predictions on the testing set.

		Actual class (measured)	
		P′	N′
Predicted class	P	*TP* (correct results)	*FP* (spurious results)
(expected)	N	*FN* (incorrect absence)	*TN* (correct absence)

Table 4.1: Confusion matrix of a binary classification problem. The variable to be predicted (the dependent variable)) can take two values: positive (P) or negative (N). The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to whether that prediction corresponds to the result of observation.

The number of classification categories, i.e., classes, can vary. In practice binary classification with only two categories is most frequently used. The classes are named as positive and negative, both for the actual and predicted classes. Furthermore, the predictions are called true positive, false positive, false negative, and true negative, to mark the outcome of the prediction, as shown in Table 4.1, where terms "positive" and "negative" refer to the classifier's prediction, and the terms "true" and "false" refer to whether that prediction corresponds to the result of observation. One of the ways of measuring how good the classifier is predicting the classes of items in the testing set is

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\,, \tag{4.1}$$

that is a ratio of the number of correct predictions and the number of all items. However, the accuracy does not provide complete information about the performance of the classifier. If we deal with a strongly unbalanced set of items; e.g., a set of items where there are many more items of negative than positive class; we can artificially inflate the accuracy by predicting the class of all the items as negative. This problem is known as the *accuracy paradox*. For that reason another methods are used to evaluate a classifier, some of which we describe below.

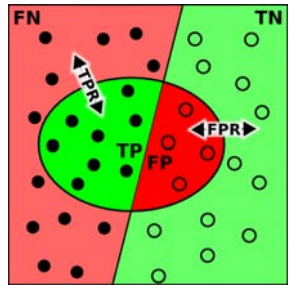


Figure 4.1: An illustration of measures of the performance of a binary classification. The actual class of the items is marked both for the positive class (filled circles) and the negative class (open circles). The ellipse in middle contains the elements predicted as positive. The green regions show true predictions, i.e., correct results, while the red regions represent false predictions, i.e., errors. Additionally, true positive rate (TPR) and false positive rate (FPR) are illustrated.

Typically a binary classifier has a parameter that affects its outcome. For instance, the parameter can be the discrimination threshold of the mapping function above which the prediction is positive and below which it is negative, or vice versa. The accuracy and precision depend on the value of this parameter. Other relevant measures of performance used in such circumstances are true positive rate

$$TPR = TP/P = TP/(TP + FN), \tag{4.2}$$

and false positive rate

$$FPR = FP/N = FP/(FP + TN). \tag{4.3}$$

The TPR measures how many correct positive results occur among all positive items, while the FPR quantifies how many incorrect positive results occur among all negative items, as shown in Figure 4.1. Both TPR and FPR take values from 0 to 1.

One of the most common and established ways of demonstrating the performance of a binary classifier is receiver operating characteristic curve (ROC). The

ROC is a figure that plots TPR (on y-axis) versus FPR (x-axis) as the discrimination threshold of the classifier is varied, depicting relative trade-off between true positive and false positive. An illustration of the ROC is shown in Figure 4.2. Each instance of the confusion matrix for the given threshold represents one point in the ROC space. The best possible prediction would be represented by a point in the upper left corner of the ROC space, corresponding to $FPR = 0$ (no false positives) and $TPR = 1$ (no false negatives). A completely random prediction would be represented by points along the diagonal line from the left bottom to the top right corner. Note that the results of a consistently poor predictor can simply be inverted to obtain a good predictor. Finally, the area under the curve (AUC) in the normalized units is equal to the probability that a classifier will rank a randomly chosen positive item higher than a randomly chosen negative one. The closer it is to 1 the better is the prediction. The AUC is commonly used to compare performance of different classifiers.

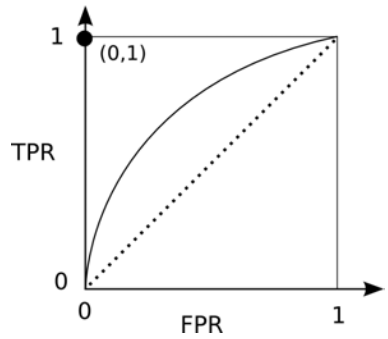


Figure 4.2: An illustration of the receiver operating characteristic curve (ROC). The diagonal corresponds to completely random predictions, while the solid line corresponds to results of an exemplary classifier. The large black dot corresponds to the results of a perfect classifier.

The learning process optimizes classifier's parameters to make the classifier fit the training data as well as possible. However, in general, the classifier will not fit the validation data as well as it fits the training data. This is a result of *overfitting*, which is particularly likely to happen when the size of the set of training items is small or the number of parameters in the classifier is large. The overfitting consists of memorizing training data, rather than generalizing it in order to retrieve the trends. In an extreme case, the classifier can completely memorize the training data and make perfect predictions for the training set, but it will fail drastically when making predictions about new or unseen data.

Cross-validation helps to mitigate the overfitting by repeatedly partitioning

the data into training and testing sets. Each partition is called a fold and is obtained randomly. One performs the classification task for each of the folds and calculates the performance of the classifier as an average over all the folds.

4.1.2 Related work

Social and thematic components of communities have been widely studied in social science. Nevertheless, the principles behind the common identity and common bond theories have never been translated into practical methods to categorize groups, nor tested on large datasets. On the other hand, data-driven studies have investigated social and thematic components separately when characterizing groups (Cox et al., 2011). Preliminary insights on the interweavement between such dimensions have been given in exploratory work on Flickr, where signals of correlation between social density and tag dispersion in groups is shown (Prieur et al., 2008). In this chapter, we go far beyond that point, defining metrics that can be used to predict if a group is social or topical and testing their effectiveness against a reliable ground truth.

Since the emergence of online social media, the global structure, evolution, and dynamics of groups have been investigated over large-scale and heterogeneous datasets (Grabowicz et al., 2013b). Evolution of groups has been characterized as a broad phenomenon (Mislove et al., 2007; Cox et al., 2011) that is dependent on the nature of the group (Cummings et al., 2002), its intrinsic fitness (Grabowicz and Eguíluz, 2012) and on the density of social ties connecting its members (Backstrom et al., 2006). Dependency of activity and connectivity on group size has been studied in several platforms (Grabowicz et al., 2012; Kairam et al., 2012; Gonçalves et al., 2011), showing relations to Dunbar's theory on the upper bound of around 150 stable social relationships for an average human (Dunbar, 1998). Besides activity, similarity between users is an important dimension in modeling individual users in groups (Tang et al., 2011), particularly given that, to a large extent, users of OSNs tend to aggregate following the homophily principle (Aiello et al., 2012). Nevertheless, similarity is not necessarily the best indicator for group activity and longevity, as diversity of content shared between group members is a relevant factor to keep alive the interest of members (Ludford et al., 2004).

At a finer scale, social communities can be described in terms of user engagement. From a quantitative perspective, the amount of participation of members in activities related to the group is varied and dependent on group size (Backstrom et al., 2008). Intra-group activity has been characterized in terms of propensity of people to reply to questions of other members (Welser et al., 2007), coherence of discussion topics (Gloor and Zhao, 2006), or item sharing practices (Negoescu and Perez, 2008). Modeling inner activity of groups has helped in finding effec-

tive strategies to predict future group growth or activity (Kairam et al., 2012), recommend group affiliation, or enhance the search experience on social platforms (Negoescu et al., 2009).

Besides the analysis of user-declared groups, communities detected with the graph-based algorithms are supposed to represent meaningful aggregations of people where social interactions take place among members (as shown in Chapter 3). Nevertheless, even if synthetic methods to verify the quality of clusters have been proposed (Lancichinetti and Fortunato, 2009a), the question of whether such artificial groups capture some notion of community perceived by the users remains open. Although the computation of cluster-goodness metrics over user-created groups gives useful hints about their structural cohesion (Yang and Leskovec, 2012), a direct comparison between user-created groups and detected communities is still missing, particularly in terms of the amount of sociality or topicality they embed.

4.2 From theory to metrics

Based on the theoretical principles of common identity and bond theories (see Section 1.5.4), it is possible to construct metrics to differentiate between the two types of groups. In particular, it is possible to quantify the reciprocity of interactions, and the topicality of the information exchanged between group members. For the representation of the system under study we adopt a generic *multidigraph* model that fits most of the current OSN platforms. Members are represented as nodes, and each distinct *interaction* between any two members is represented by a directed arc. Nodes can belong to multiple *groups* and we associate, with each group, a bag of user-generated *terms* (e.g., tags, group posts).

Next, we describe: i) *reciprocity* metrics, used to quantify group sociality; ii) *entropy* of terms, to determine how broad is the topic of discussion within a group; and iii) *activity* metrics, to measure the liveliness of the group.

4.2.1 Reciprocity

Reciprocity occurs whenever a user interacts with another user and that user responds her at any time later with the same type of interaction. We define *intra-reciprocity* of a group g as:

$$R_g^{\text{int}} = \frac{L_g^{\text{int,rec}}/2}{L_g^{\text{int,rec}}/2 + L_g^{\text{int,nrec}}}, \tag{4.4}$$

where $L_g^{\text{int,rec}}$ and $L_g^{\text{int,nrec}}$ are, respectively, the number of reciprocated and non-reciprocated links internal to the group g. Correspondingly, the *inter-reciprocity* at the border of the group is defined by R_g^{ext}, accounting for the reciprocity between members and non-members.

We normalize the intra-reciprocity score using the average reciprocity value $\langle R_g^{\text{int}} \rangle$ over all groups:

$$t_g = \frac{R_g^{\text{int}}}{\langle R_g^{\text{int}} \rangle}. \tag{4.5}$$

The larger the intra-reciprocity, the higher the probability that the group is social. Alternatively, to compensate for the effect of the correlation between reciprocity and the number of internal interactions, and to account for local effects, the intra-reciprocity can be normalized by the inter-reciprocity:

$$u_g = \frac{R_g^{\text{int}} + 1}{R_g^{\text{ext}} + 1}. \tag{4.6}$$

We add 1 to both numerator and denominator to reduce the fluctuations of u_g at low values of R_g^{ext}. This relative reciprocity compares the reciprocity between the members with their reciprocity toward people not belonging to the group. It reflects how sociality of group members distinguishes itself from the environment.

4.2.2 Topicality

The set of terms $T(g)$ associated with a group indicates the topical diversity of the group. Thus we measure the Shannon entropy of the group as

$$H(g) = -\sum_{t \in T(g)} p(t) \cdot \log_2 p(t), \tag{4.7}$$

where $p(t)$ is the probability of occurrence of the term t in the set $T(g)$. The higher the entropy, the greater is the variety of terms and, according to the theory, the more social the group is. Conversely, the lower the entropy, the more topical the group is. In addition, since not all groups have the same number of terms and the entropy value grows with the total number of terms, we introduce the *normalized entropy* h_g, which is normalized by the average value of entropy for the groups with the same number of terms:

$$h_g = \frac{H(g)}{\langle H(f) \rangle_{|T(g)|=|T(f)|}}. \tag{4.8}$$

4.2.3 Activity

Even if, for the considered theory, activity is not a discriminating factor between social and topical groups, it is useful to characterize the liveliness of a community. Activity is quantified in terms of the number of internal interactions normalized by the expected number of internal interactions for a set of nodes with the same degree sequence:

$$a_g = \frac{L_g^{\text{int}}}{(K_g^{\text{in}} K_g^{\text{out}})/E}. \tag{4.9}$$

K_g^{out} and K_g^{in} are total numbers of interactions originated by members of the group g or being targeted to members of this group, where E is the total number of interactions in the network. If this property has a value higher than 1, then the number of internal interactions is higher than the number of interactions expected in a random scenario with the same group activity volume.

Another way of measuring activity of a group is by comparing density of its internal interactions with the density of its external interactions:

$$b_g = \frac{L_g^{\text{int}}/(s_g(s_g-1))}{L_g^{\text{ext}}/(2(N-s_g)s_g)}, \tag{4.10}$$

where s_g is the cardinality of group g and N is total number of nodes in the network. Values of b_g greater than 1 indicate a density of internal interactions higher than interactions between the group and the rest of the network. This metric effectively compares intensity of interactions between members of the groups with the intensity of their interactions with the entire network.

4.3 Dataset and preprocessing

The wide variety of user groups and the richness of interaction types make Flickr an ideal platform for our study. We use only public, anonymous data retrievable via the Flickr public API, until the end of 2008. Table 4.2 summarizes the data described below.

comments	favorites	contacts	declared groups	detected groups
238M	112M	71M	504K	646K

Table 4.2: Total number of interactions and groups.

4.3.1 User interactions

We collected three types of pairwise, directed interactions:

Comments. User u comments on a photo of user v. This interaction is *mediated* through the photo. We filter out the comments of users on their own photos, obtaining a total of 238M comments.

Favorites. User u marks one of user v's photos as a *favorite*. The interaction is mediated through the favorited photo. We extract 112M favorite interactions.

Contacts. User u adds user v among his contacts. Social contacts in Flickr are directed and may be reciprocated. One person can choose another person as his contact only once and the relation remains in the same state until the contact is removed. There are 71M contacts in our dataset.

4.3.2 Groups

Users of Flickr can create, moderate and administer their own groups. Most groups are open, so users can join without an invitation. Others are only by invitation and joining requires the administrator's permission. There are over 500K groups in our Flickr dataset.

In addition to user-defined groups (we refer to them as declared), we analyze the sociality and topicality properties of groups that are not created by users but are instead found by community detection algorithms (we name these detected groups). We applied the OSLOM community detection algorithm (see Appendix I) over the entire network of social contacts in our dataset. We choose OSLOM because of similar reasons as the ones stated in the previous chapter, i.e., it detects overlapping communities, which is a natural feature of real groups, it performs well in recent community detection benchmarks (Lancichinetti and Fortunato, 2009a), and it outperformed other algorithms that we tested. OSLOM detected 646K groups.

4.3.3 Tags

We use *tags* of the photos as terms for our model. The primary set of photos from which we extract tags is the *photo pool* of the group (i.e., the photos uploaded to the group by its members). Photo pools are available for declared groups only. In addition, in both declared and detected groups, the interactions between members of the group that are mediated through photos (i.e., comments, favorites) result in two additional photo sets from which tags are extracted. We process the three tag sets separately (pool, comments, favorites), and for each of them we compute the normalized entropy (h_g^{pool}, h_g^{com}, h_g^{fav}).

4.4 Group labeling

To determine whether the defined metrics correctly capture the sociality and topicality of groups, we compare them against a reliable ground truth. We asked human editors to label groups based on well-defined guidelines extracted directly from the common identity and common bond theory (Ren et al., 2007). For the labeling we randomly selected groups meeting the following requirements: i) more than 5 members, ii) more than 100 internal comments, iii) relative activities a_g^{com} and b_g^{com} higher than 10^2. The third requirement ensured us that the selected groups were active well above the expected values in a random case. After this selection we obtained over 34K declared groups and over 33K detected groups. We describe the labeling process of such groups in detail in the following subsections.

4.4.1 Information provided to editors

The labeling is based on the human capability of processing the semantics and sentiment behind text and photos. The labeling was performed to generate a ground truth of social and topical groups. The editors were asked to make judgments in this respect and were presented with the following information for each group:

Group profile. The Flickr group profile consists of the group name, description by the creator of the group, discussion board, photo pool, and map of places where photos uploaded to the group pool were taken. This information is available only for declared groups.

Comments. We provide text of all comments that happen between the members. Comments are shown in chronological order and are grouped by thread, if they appear under the same photo. Additionally a link to the photo is also included.

Tags. Editors are shown the list of the 5 most frequent tags attached to the photos that mediate the internal comments to the group. The list is sorted alphabetically.

4.4.2 Labeling guidelines

Human labelers were shown the information described above and asked to categorize groups as either "social", "topical" or "unknown". The last case is reserved for groups for which text is written in a language unknown to the labeler, making the task impossible to accomplish. Intentionally, no "unsure" category was allowed to keep the categorization strictly binary, as the theory does. Some groups can be both topical and social, and therefore, difficult to categorize, but for the

sake of clarity and conformity with the theory we kept the categorization as a binary task. Editors were provided with specific instructions on how to recognize social and topical groups, and on how to perform the categorization. The guidelines are summarized as follows:

Comments and photos. By examining comments and photos, find traces of people who know each other or who have a personal relationship. Knowing each other's real names, spending time together, co-appearing in photos, sharing common past experiences, referencing mutually known places, and disclosing personal information are all signals of the presence of a social relationship (Collins and Miller, 1994). The predominance of friendly and colloquial comments (e.g., jokes, laughter) is another element distinguishing social groups from topical groups. In topical groups, the atmosphere is more formal and comments tend to be more impersonal (Sassenberg, 2002). Examples of impersonal comments include expressing appreciation for photos, praising the photographers, thanking them for their work, or commenting on any particular topic in a neutral way. As a rule of thumb, if many personal comments are detected, then the sociality of the group should be considered high. If such comments are not many (e.g., just between small subsets of members), but the overall atmosphere of the interaction is rather personal and friendly, then we consider the sociality of this group as fairly present. If, on the other hand, comments are mainly impersonal and neutral, sociality has to be considered low, in favor of higher topicality.

Tags and description. Read the tags and the profile description of the group. If the tags are semantically consistent, then the topicality of the group should be considered high, and even higher if the name and description of the group corresponds to the content of the tags. In some cases, tags or group descriptions can contain words indicating personal relations or events (e.g., "wedding", "grandpa", names, etc.), indicating a higher sociality of the group. Tags can also contain names of specific locations. Geo-characterized tags can be reasserted by looking at the map of places where photos were taken. Such tags are a good indication that the sociality of the group is present, but that has to be confirmed through the inspection of comments.

The editors labeled the groups after judging the two aspects above. If both tags and comments are highly social or topical, then the choice of label is straightforward. If the tags are highly topical and the comments are not social, then the group is labeled as topical, and vice versa. If the tags are a bit topical and comments highly social, then the group is labeled as social. The labelers were asked to read as many comments as needed to arrive to a fairly clear decision.

4.4.3 Group examples

To provide a sense of how the defined guidelines were applied in practice, we describe two examples. The first one is a group titled "Airlines Austrian", tagged with labels "aircraft", "airport", and "spotting." Photos are from different countries in Europe and most of them depict airplanes. Members are active in commenting and writing comments related on the aircraft theme (e.g., "I just love this airplane, the TU-154M is just a plane Boeing or Airbus could never design"). In this case, all of the features are aligned with the concept of topical group defined in the guidelines. The second group is named "Camp Baby 2008" and it is described in the main page as a collection of photos of a two-day event for young mothers taking place at a specific location. Photos depict people attending the event and interacting with each other with a friendly attitude. Tags and comments often contain names of individuals and references to past common experiences (e.g., "I love Mindy and can not wait to see her again!!"). Although the group has a specific topic, its social component is very strong.

In practice, more ambiguous cases can occur and, ultimately, the decision of the labeler has an arbitrary component, as in every complex annotation process. Nevertheless, the defined guidelines gave the labelers precise instructions and, as described next, we recurred to multiple independent editors to assess the quality of the extracted ground truth.

4.4.4 Labeling outcome

A total of 101 declared groups and 69 detected groups were labeled by 3 people: two of the authors and an independent labeler who was not aware of the type of study nor of the purpose of the labeling. The inter-labeler agreement, measured as Fleiss' Kappa, is 0.60 for the declared groups, meaning that there exists good agreement between labelers.

In order to assess the quality of the labels, we also counted the number of messages exchanged between group members. The counting was done anonymously in aggregate and the content of the messages was not accessed. Groups labeled as social contain around twice as many messages between their members compared to topical groups of similar size. Even if this does not constitute a proof of higher sociality, intuitively people who get in touch via one-on-one communication are more likely to have a more intimate social relationship.

The Kappa value for detected groups is around 0.44, revealing lower agreement. A factor that partially determined such result is the lack of information about the group's profile, since it is not available for detected groups. Another cause of the disagreement is a higher variability in the comments. This can happen because we use a network of contacts for the purpose of finding clusters and defining detected groups, which may not be the best proxy of personal relations.

In total we label 565 distinct declared groups and 126 distinct detected groups. We characterize them in the following section.

4.5 Characterization of groups

We begin the analysis with a direct comparison of the overlap between the declared and detected groups. Then, we characterize the two sets of groups in terms of the metrics that we introduced in Section 4.2. Finally, we study the relation between the labels of the declared groups annotated by the editors and the values of the metrics. Additionally, we report the ratio of groups labeled as social and topical among both declared and detected groups.

4.5.1 Membership overlap of declared and detected groups

The groups from the two sets share typical properties of groups found in on-line social networks. The distribution of sizes of groups in both cases is heavy-tailed and close to power-laws (not shown). Declared groups tend to be much bigger, having on average 61 members versus 7 members in detected groups.

To test if the groups from the two sets overlap, and to what extent, we measure the Jaccard similarity between their sets of members. Similarity is computed for all declared-detected group pairs and for each detected group we select the declared one with the highest similarity value as the best match. We plot the average similarity of the best matches as a function of the size of groups in Figure 4.3a. Zero values of similarity are not taken into account for these averages. For the purpose of comparison with a null model, in Figure 4.3b we draw the same plot after randomly reshuffling the members of detected groups, while preserving their sizes. We observe that the two plots differ in values significantly along the diagonal, and that the difference between them is substantial, as shown in Figure 4.3c, meaning that indeed detected groups are, to some extent, similar to the declared ones. Further insights are shown in Figure 4.3d, where we depict the distribution of similarities of pairs of groups extracted from a small sector of the diagonal, having between 32 and 64 members. The figure shows that there exist multiple detected groups that overlap significantly with declared groups, and that randomized groups do not show this pattern. This holds for groups of all sizes, as shown in Figure 4.3e, in which we plot the 91th and 99th percentiles of the best match similarity for detected groups of various sizes (e.g., 1% of detected groups of size 20 have similarity with declared groups higher than 0.75, while for the randomized case 1% of the groups have similarity higher than just 0.05). Therefore, in some cases, the community detection algorithm finds groups that

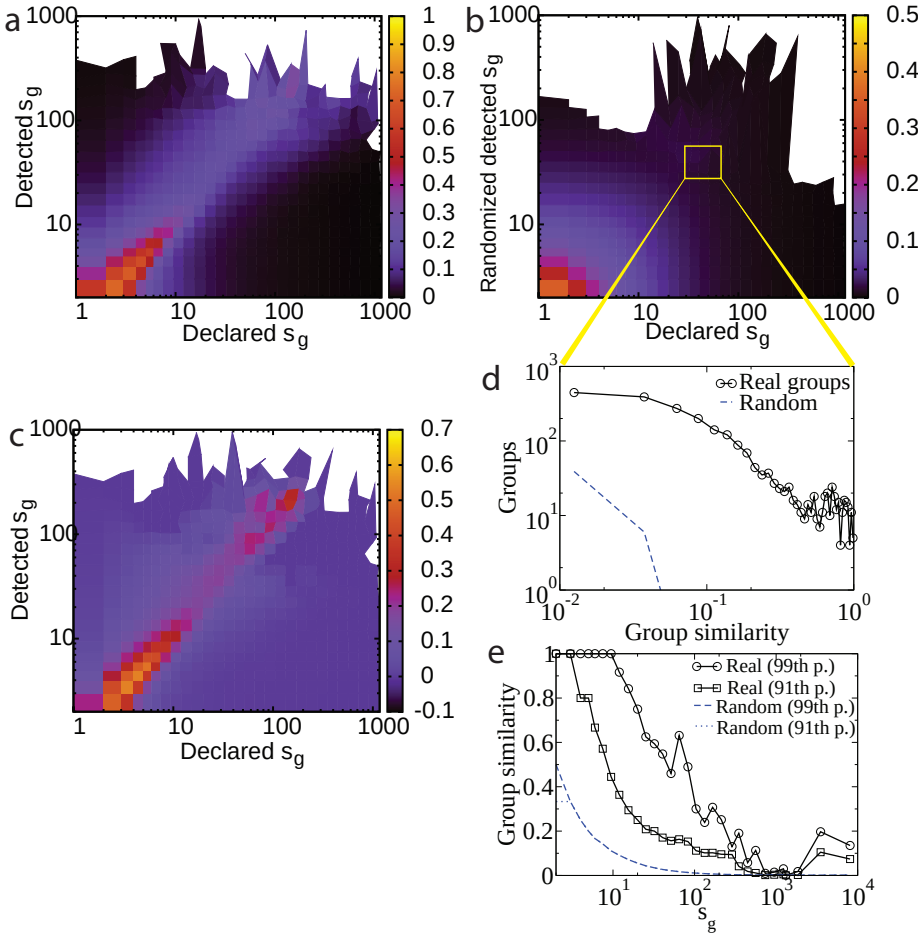


Figure 4.3: (a) Jaccard similarity between declared and detected groups as a function of their sizes. Diagonal shows an interesting pattern, which (b) is not reproduced for randomized detected groups. We subtract (b) from (a) and plot (c) the result. (d) Histogram of the similarity between declared and detected groups for a sample of groups lying at the diagonal, for both real detected groups and the randomized ones. (e) 91th and 99th percentiles of the similarity between declared and detected groups.

are close to the ones defined by users (i.e., declared groups). We present evidences that this does not occur by chance through the comparison with the randomized case. Nevertheless, a substantial overlap is found for just a small percentage of groups. Most of the group pairs have similarity close to 0. Consequently, the similarity of detected groups to the best-matching declared groups is 0.082, while for the randomized detected groups it is not much lower, yielding 0.058.

4.5.2 Statistical properties of metrics

Besides directly comparing membership overlap, we study the variation of the metrics defined in Section 4.2 with the group size. Reciprocity and normalized entropy have a wide local maximum for groups of sizes between 50 and 100 members, both for declared and detected groups, as shown in Figures 4.4a-d. This holds for all interactions and all sets of tags, with the exception of contacts, for which the curves are relatively flat. We have found a similar local maximum for pairwise interactions in Twitter using various community detection algorithms in Section 3.5. We perform a randomization of photos between groups, keeping the number of photos per group fixed. The normalized entropy calculated for the shuffled photos stays close to 1, as expected, and the maximum disappears. A possible interpretation of the existence of the maximum is that these sizes tend to correspond to social groups, while bigger groups are more frequently topical. Further findings to support this interpretation are presented in the next subsection.

Strong correspondence of the maxima for normalized entropy and reciprocity suggests that these properties are correlated, as shown in Figure 4.5. Whereas it seems straightforward to explain the correlation between reciprocity of comments and normalized entropy based on commented photos, it is not clear why there is also a positive correlation with normalized entropy based on other sets of photos. Apparently, high intra-reciprocity correlates with wider variety of topics covered inside of that group, and vice versa. The theory predicts that social groups have high intra-reciprocity and wide variety of topics, while topical groups have both properties low, which implies the correlation between the two properties. Therefore, this result is the first signal that predictions of the theory are valid to some extent.

The values of relative activity both in declared and detected groups are very high, as presented in Figures 4.4e,f. As expected, activity of randomized groups exhibits values around 1 for all group sizes. For real groups instead, the value of relative activity decreases with the size of groups and gets close to 1 for very large ones. This is caused by the fact that larger groups cannot be as engaging to the users as smaller groups and the social commitment of their members towards other members of the group drops. Additionally, we observe a sharp decay in

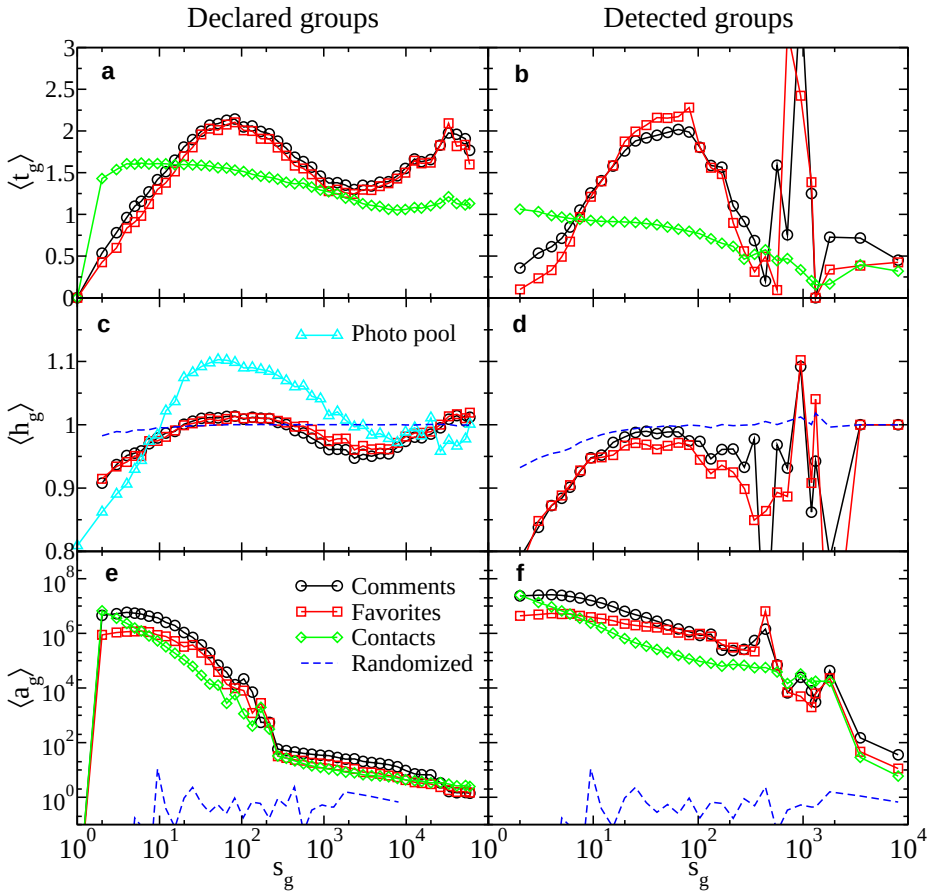


Figure 4.4: (a-b) Dependency of normalized reciprocity, (c-d) normalized entropy and (e-f) relative activity on size of groups for comments, favorites, contacts, and photo pools, for declared and detected groups. Blue dashed line is for (c-d) randomized photos and (e-f) randomized groups.

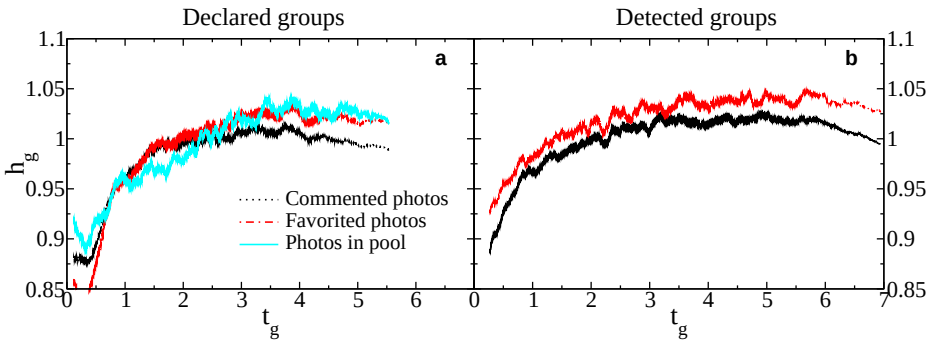


Figure 4.5: Correlation between reciprocity of comments inside a group and entropy of photos commented or favorited between its members, or belonging to the photo pool of this group, for declared and detected groups.

the activity for groups of size around 200, in agreement with Dunbar's theory on the upper bound of the number of stable relationships manageable by a human. We have found a corresponding drop in the fraction of links with mentions in respect to the baseline in Twitter in Section 3.5. The drop in the activity for detected groups is continuous and more moderate (Figure 4.4f), since community detection algorithms tend by design to output node clusters with high numbers of connections between them.

4.5.3 Relation between metrics and group label

Here we analyze properties and values of the metrics for groups labeled through the editorial process. First, the ratio of groups labeled as social differs between declared and detected groups. In declared groups we find around 48% social groups, whereas among detected groups almost 69% are labeled as social. Additionally, we picked 50 detected groups among the ones that are the most similar to declared groups. Specifically, we selected them randomly from the 99th percentile shown in Figure 4.3. These groups have significant overlap with declared groups and should share similar properties. Indeed, the ratio of groups labeled as social among them is closer to that of declared groups and equal to 53%. We conclude that detected groups are more likely to be social than declared ones. It is a somewhat expected result, since clustering algorithms detect dense parts of a network, and so they are inclined to detect areas with more reciprocal connections. Note that the theory envisions more reciprocal relations in social groups. Thus, community detection algorithms are more likely to find social groups, however, determining to what extent it happens is not trivial.

One of the expectations is that bond-based groups should not be very large, as the human capacity for stable relationships is limited. As pointed in Subsection 4.5.2, the Dunbar number is considered as a possible cap for the size of such groups, while topical groups do no yield such a restriction. In line with this expectation, we find that declared groups labeled as social have on average 35 members, whereas groups labeled as topical have on average around 172 members.

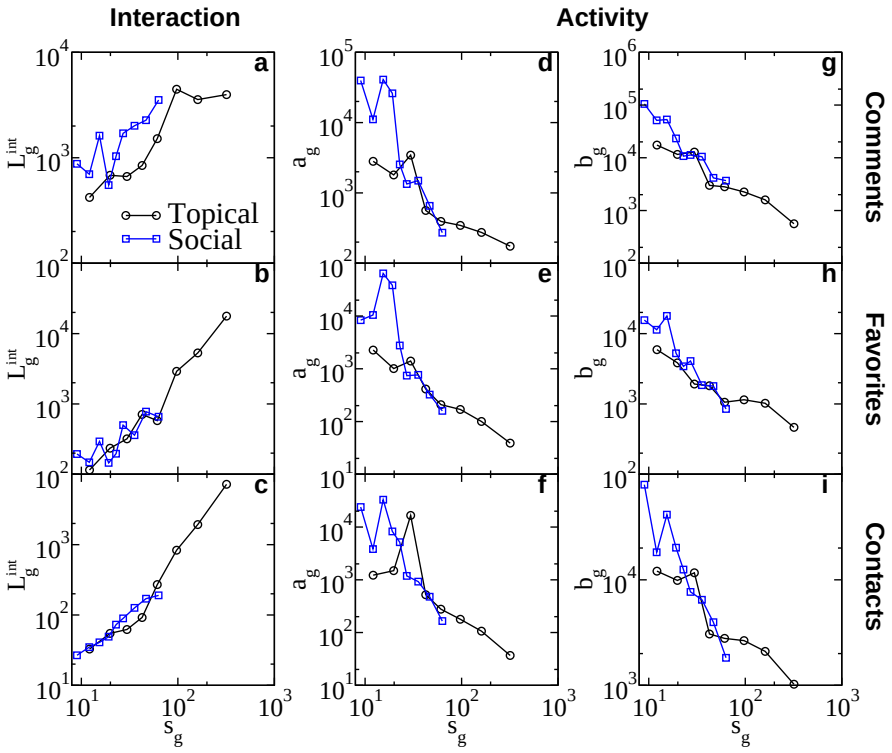


Figure 4.6: Averages of various properties of topical (black circles) and social (blue squares) groups as a function of their size. Each point corresponds to 30 groups.

We find insightful differences and similarities in various properties, which we explore in detail in Figures 4.6 to 4.8. We plot them as a function of the size of groups as they vary drastically with it, and one needs to compare groups of similar sizes in order to draw unbiased conclusions.

First, there are almost no differences in the number of photos (not shown),

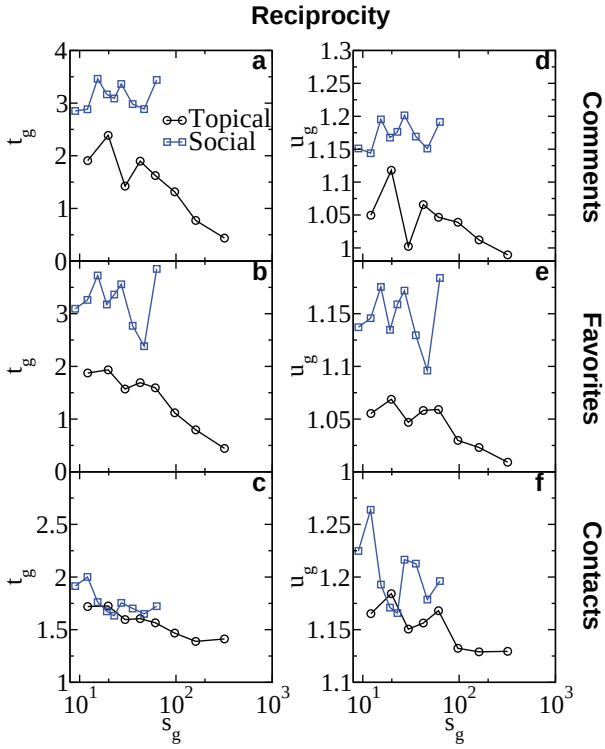


Figure 4.7: Averages of various properties of topical (black circles) and social (blue squares) groups as a function of their size. Each point corresponds to 30 groups.

favorites, and contacts (as in Figures 4.6b,c) inside social and topical groups. The number of comments is, however, around 2 times higher in social groups than in topical groups of similar size (Figure 4.6a). More differences can be found when looking at relative activity (Figures 4.6d-i), which compares the interaction internal to the group with the overall activity level of users belonging to groups. In all three types of interaction the relative activity metrics for social groups yield values from 2 to over 10 times higher than for topical groups. The activity metric b_g compares the density of internal interactions with the density of external interactions. Therefore this result reflects a stronger focus or even an isolation of members belonging to social groups from the rest of people they interact with.

Most importantly, we observe large differences in values of reciprocity and

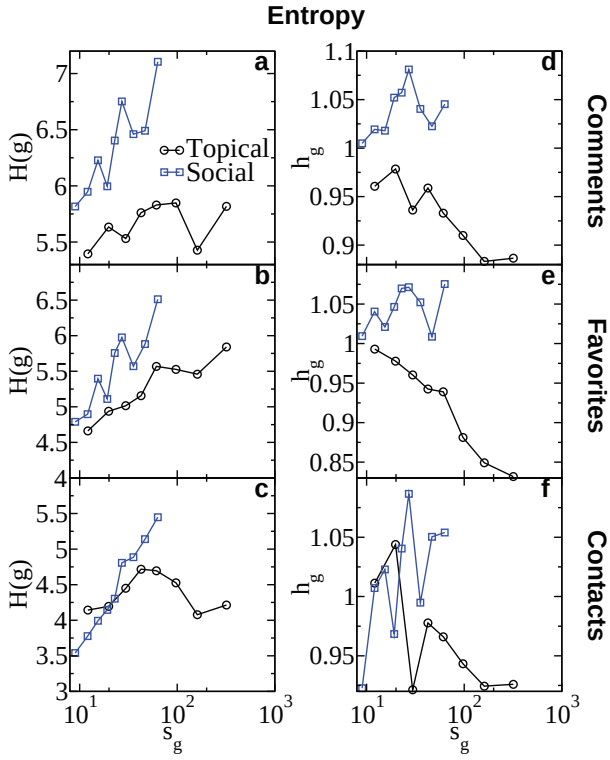


Figure 4.8: Averages of various properties of topical (black circles) and social (blue squares) groups as a function of their size. Each point corresponds to 30 groups.

relative reciprocity of comments and favorites. Social groups exhibit significantly higher reciprocity than topical groups (Figures 4.7a-f), in line with common identity and common bond theory. There is no difference in reciprocity of contacts. A plausible interpretation is that contacts do not reflect personal relations between connected users, as users often add people they do not know and do not interact with as contacts in order to follow their content. Finally, we observe significantly higher values of entropy and normalized entropy in social groups than in topical ones (Figures 4.8a,b,d,e). This holds for the tags extracted from photos commented and favorited between members. Assuming that tags of photos represent topics of interactions, the result is consistent with bond attachment. It is expected for members of bond-based groups to cover many different topics and areas in their interactions, whereas members of identity-based groups focus

their interactions on specific topics. However, this effect is weaker for the tags extracted from photo pool of the group (Figures 4.8c,f). Apparently, the content of the photo pool does not always reflect well the interactions and relations between members of the group.

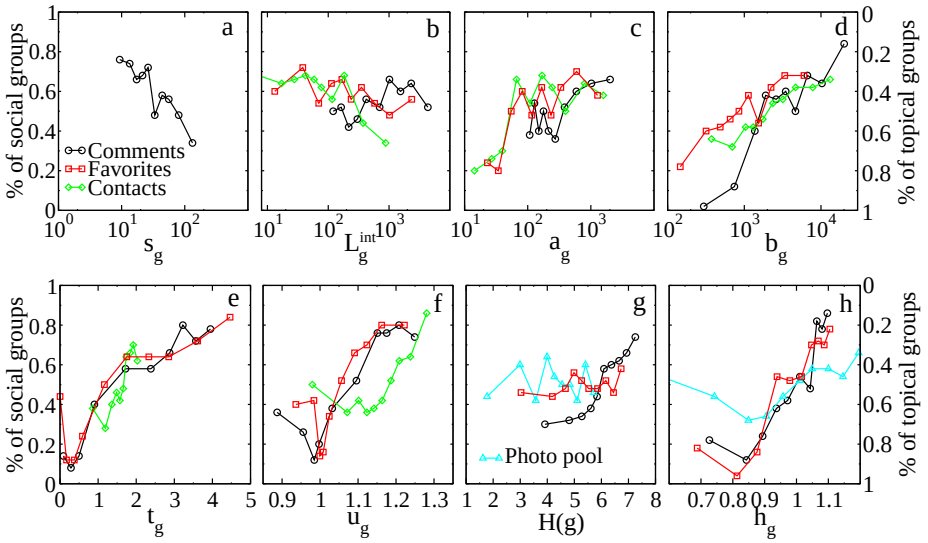


Figure 4.9: Dependence of fraction f of groups labeled as social on various metrics: based on comments, favorites, contacts, and photo pools. The remaining $(1 - f)$ groups are topical. Each point corresponds to 50 groups.

Additionally, we plot the fraction of groups labeled as social with respect to group size, activity, reciprocity, and entropy (Figure 4.9). The fraction correlates negatively with group size, as expected (Figure 4.9a). The correlations with the number of interactions and relative activity a_g are rather weak (Figures 4.9b,c), whereas, surprisingly, there is a strong dependency on relative activity b_g (Figure 4.9d). For the lowest values of b_g^{com}, 95% of the groups are topical, while for the highest, 80% of the groups are social. High values of b_g can mean stronger focus on the group, or even an isolation of the group members from the rest of people they interact with. This result is related to the observation that it is hard to enter bond-based groups due to strong relations existing between their members and because high investment is required to create such relations with them (Ren et al., 2007). Direct reciprocity of interactions, with the exception of contacts, correlates strongly with the fraction of social groups (Figures 4.9e,f). This result is expected based on common bond theory. Furthermore, we found

that the entropy of tags correlates with social groups, but entropy based on other sources does not (Figure 4.9g). However, we find that our normalized entropy performs much better than this, and a strong correlation is found both for tags extracted from comments and from favorites (Figure 4.9h). This shows that the normalized entropy of tags is a more appropriate method of measuring topical diversity of communications of a set of people.

4.6 Group type detection

The properties of labeled social and topical groups tend to confirm the validity of the principles identified by the common identity and common bond theory. A further confirmation comes from the ability of the defined metrics to predict the tendency of a group towards sociality or topicality. To this end, we propose and compare two methods to predict the group type and we test their accuracy over the corpus of the labeled groups.

4.6.1 Prediction methodology

The first approach we use is a linear combination of the metrics. To this end, we select the features that are the most related to the sociological theory and for which we built specific metrics, i.e., t_g, u_g and h_g. Each of them is applied to the 3 different interaction types and bags of tags, which produces a total of 9 values. We transform the values of the metrics into their t-statistics by subtracting the average value and dividing them by the standard deviation of the distribution. Then, we weight the normalized scores evenly by dividing them by the total number of metrics considered and we finally sum them up to obtain a single *sociality score* S_g. All of the components are supposed to score high for social groups. Therefore, the higher the value of the score, the higher the chance that the group is social rather than topical. To convert the score into a binary label, a fixed threshold above which groups are predicted to be social must be selected. Using this approach, we aim at testing if those metrics, based on the theory, can be successful in predicting the type of group (social or topical).

The second approach relies on machine-learning supervised methods that use the metrics' values as features. Features are combined in a classifier that is first trained on a sample of labeled data to learn a prediction model. The trained classifier then outputs a binary prediction for any new group instance defined in the same feature space. Due to the limited size of our corpus of labeled groups, we estimate the classifier performance using 10-fold cross validation. We report results on a Rotation Forest classifier, which performed best in comparison to

several algorithms implemented in WEKA (Hall et al., 2009). For the classifier we used a wider set of features than for the linear combination approach, namely: group size s_g and L_g^{int}, a_g, b_g, t_g, u_g, $H(g)$, h_g, each applied to the 3 different interaction types and bags of tags. This results in a total of 22 features. Such a wide set of features to test if indeed the metrics proposed to distinguish between the social and topical groups are the best ones for the task. The relative predictive power of the features is measured through a feature selection algorithm.

4.6.2 Prediction results

The ratio of groups labeled as social increases quickly with the score S_g, as shown in Figure 4.10a. This summarizes the findings of the previous sections, suggesting that the features embedded in the score are able to capture well the nature of the groups. The higher the score, the higher the probability that the group is social; the lower, the more topical. If we fix the threshold for the S_g value in order to perform a binary group classification, it is clear that several misclassifications will occur, especially around the threshold value. An example for threshold at 0 is shown in Figure 4.10a. Conversely, the classifier performs significantly better and achieves the ratio that adheres much more to the actual ratio of social and topical groups.

Both methods, however, fail more frequently for groups with mixed social and topical features. When the score is around zero, groups can be either social or topical, or both, and the decision about the nature of the group is more difficult. The prediction accuracies of the classifier and of the score-based predictions have an evident drop of performance around 0 (Figure 4.10b). The accuracy at the extreme values of the score is close to 0.95, while it falls below 0.6 for groups with a score close to 0. On the other hand, this drop appears also in the agreement between two of the human labelers, measured as a ratio of groups that have been given the same label. Apparently, this is a shortcoming of the binary classification itself, as opposed to multi-label classification.

The overall performance of the two approaches can be compared fairly through ROC curves (Figure 4.10c), which astray from the selection of a fixed threshold. The curve for the classifier (computed for the 10-fold cross validation) always performs better, and this is reflected in the considerably higher AUC value and accuracy, as shown in Table 4.3.

In addition, to determine the most predictive features, we rank the features using chi-square feature selection. The top 5 features are, in decreasing order of importance: h_g^{com}, t_g^{com}, u_g^{com}, h_g^{fav}, and b_g^{com}. The selected set is the optimal for the prediction performance: retraining the classifier on such restricted set of features results in stable performance, as shown in Table 4.3. The top 4 most predictive features correspond directly to the expectations of the theory

and results of the analysis from Section 4.5. Reciprocity-based metrics and normalized entropy are significantly more predictive than other features. The high position of relative activity b_g^{com} is rather unexpected. However, its importance and interpretation is discussed in Section 4.5.

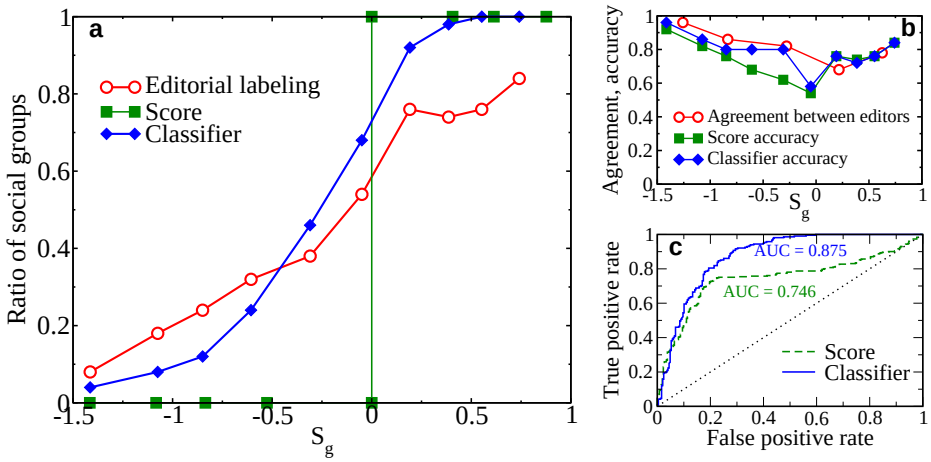


Figure 4.10: Comparison between the prediction methods. (a) Dependence of the ratio of the groups labeled as social on the score for the score-based (threshold at 0) and the supervised classifiers. (b) The accuracy of prediction of the two techniques and agreement between two of the labelers against the score values. (c) ROC curves for the prediction with the two techniques.

Method	Accuracy	AUC
Score	0.763	0.749
Classifier	0.801	0.879
Classifier χ^2_{top5}	0.803	0.872

Table 4.3: Group type prediction performance using i) the score with threshold at 0, ii) 10-fold cross validation on a Rotation Forest classifier trained on all the features, or iii) the same classifier trained on the set of top 5 predictive features, according to the chi-square feature selection.

4.7 Conclusions

Common identity and common bond theory indicates a high-level characterization of topical and social groups. We propose metrics capturing reciprocity of interactions and entropy of user-generated terms, to realize the concepts discussed in the theory and to measure sociality and topicality of groups. We label a set of groups from Flickr as either topical or social based on semantics and sentiment behind text and photos. We leverage this ground truth to show that the metrics, combined with a machine-learning approach, predict the group type with high accuracy. Moreover, we note that the degree of isolation of the group activity from the rest of the social network, measured in terms of the density of interactions, is a good predictor of the group type, in addition to the elements identified in the theory. Besides the main prediction results, the supporting analysis of the group properties in terms of the identified dimensions confirms the theory from different angles and highlights other interesting findings. In particular, dependencies of the metrics with the group size confirm previous observations about the effective size of social communities, peaking around rather small sizes and being limited by a cap of 100-200 members.

The study is complemented with a comparison of the structure and sociality and topicality traits between declared groups and groups from a community detection algorithm. Declared groups do not overlap much with detected groups on average, but they match each other significantly more than the random case for groups of comparable sizes. Furthermore, detected groups are more often social than the declared ones. A natural question is if this result holds also for declared groups from other OSNs and for groups detected with other clustering algorithms. For instance, one can expect that system-wide privacy settings of groups may affect the way the groups are used. This is an interesting question for the future research.

Extensions to the study include a more exhaustive extraction of detected groups using a different network than the network of contacts e.g., we find mutual comments to carry more social traits than the contacts do. Another interesting extension could be multi-label classification of groups, in order to better categorize groups with mixed social and topical components. Furthermore, relations between tags could be taken into account, e.g., identifying synonyms between different tags would improve the accuracy of the prediction.

Finally, development of the method for the group type detection can contribute to the design of OSNs. The service offered to group members could be tailored depending on the nature of a group. Such contextualized services could range from the change of the user interface (e.g., highlight more the content shared in a topical group or the members and their activity in a social group),

or the type of advertisement shown (e.g., ads related to the topic of conversation for topical groups, while viral-marketing campaigns for social groups).

CHAPTER 4. PREDICTING TYPES OF GROUPS BASED ON IDENTITY AND BOND THEORIES

Chapter 5

A model coupling link formation and mobility

Individuals tend to be friends with the people they spend time with and they choose to spend time with their friends, inextricably entangling physical location and social relationships (see Subsection 1.2.7). As a result, it is possible to predict not only someone's location from their friends' locations but also friendship from spatial and temporal co-occurrence. Although several models have been developed to separately describe mobility and the evolution of social networks, there is a lack of studies coupling social interactions and mobility. In this chapter, we introduce a new model that bridges this gap by explicitly considering the feedback of mobility on the formation of social ties. For validation we use data coming from three OSNs (Twitter, Gowalla and Brightkite).

5.1 Introduction

People tend to interact and maintain relations with geographically close peers, a tendency reflected by the decay of the probability to interact with physical distance, described in Subsection 1.2.7. Furthermore, it has been shown that online (Crandall et al., 2010) and offline (González et al., 2006) social links can be inferred from user co-occurrences in space and time (see Figure 5.1A) and, likewise, that the location of a person can be predicted from the geographic positions of his or her online friends (Backstrom et al., 2010). Some further aspects of the relation between geography and online social contacts have been studied such as the probability that a link at a given distance closes a triangle (Liben-Nowell et al., 2005; Lambiotte et al., 2008; Scellato et al., 2011), the connections between

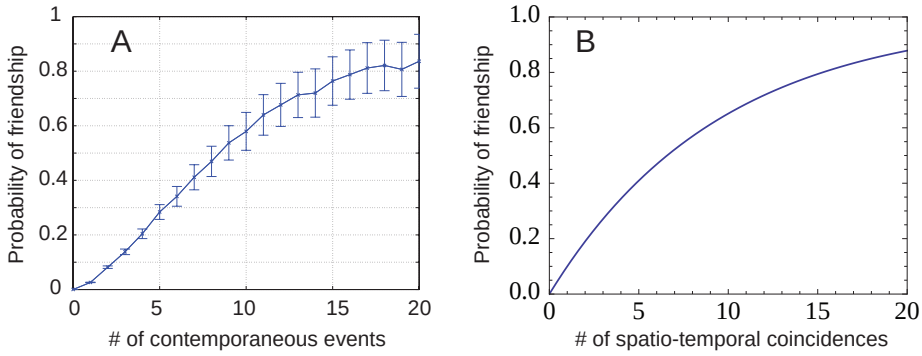


Figure 5.1: The probability of a link as a function of the number of spatio-temporal coincidences: (A) measured in Flickr for a spatial box size of 0.001° and a time window of one day, adapted from (Crandall et al., 2010); (B) assumed in our TF model, for the same size of spatial box and assuming that one time step of the simulation corresponds roughly to one day.

users in different countries (Takhteyev et al., 2012), the social interactions and mobility in emergency situations (Lu et al., 2012) or the overlap between users' ego networks and how it decays with the distance (Volkovich et al., 2012). Multi-parametric inference methods have been applied to empirical data with the aim of predicting link presence and users' locations (Wang et al., 2011; Cho et al., 2011; Sadilek et al., 2012). These works show that the accuracy of link prediction is considerably improved by taking into account the geographical information, and that the accuracy of location prediction is enhanced when the online social links are provided.

The wide availability of geo-localized data has allowed for a detailed exploration of human mobility (Brockmann et al., 2006; González et al., 2008; Balcan et al., 2009; Wang et al., 2009; Brockmann, 2010; Phithakkitnukoon et al., 2012; Simini et al., 2012). The length of displacements between consecutive locations of a person was found to follow a broad distribution, well fitted by a power-law decaying function (Brockmann et al., 2006; González et al., 2008). Mobility models are introduced in more detail in the next subsection. However, despite the supporting evidence (Phithakkitnukoon et al., 2012), most of these models lack a connection between mobility and social interactions (Giannotti et al., 2012).

In this study, we lay a bridge between these two worlds by introducing a model coupling social tie formation and spatial mobility. The model simulates the movement of individuals and creates links between them when they are physically close mimicking the effect of face-to-face interactions. We study the model both

numerically and analytically and confront its results with empirical data obtained from three OSNs.

5.1.1 Models of mobility

Diffusion plays an important role in physics and various other sciences. Brownian motion is a mathematical model used first to describe movements of a particle suspended in a fluid that result from frequent bombardments by fast atoms or molecules. It is known that under such conditions the movement of the particle is described by a normal distribution (Mörters et al., 2010), namely that the jump lengths between consecutive locations in equal time intervals are normally distributed. Note that the normal distribution decays faster than exponentially. It follows that the average distance from the starting positions of the particle, known as the radius of gyration r_g, grows as a square root of time $r_g \propto t^{0.5}$. Brownian motion is broadly found in nature understood as material world.

Considerably different mobility patterns are found in movements of animals and human beings. In these cases, usually the distribution of jumps between consecutive locations is heavy-tailed. Therefore, movements of animals such as monkeys, marine predators, and humans, are approximated better with Lévy flight than the Brownian motion (Giannotti et al., 2012). Here, we focus on the human mobility. It has been reported that the probability of the displacements for humans follows a power-law (González et al., 2008; Song et al., 2010a; Brockmann et al., 2006)

$$P(\Delta r) \sim \Delta r^{-\beta}. \tag{5.1}$$

Naturally, various studies provide different exact values of the exponent $1.5 < \beta < 2.0$, which depends on the specific system under study. Note that such distributions allow occasionally very far travels that correspond to the heavy tail of the distribution. As a consequence, the radius of gyration r_g diverges for $\beta < 2$. Random walks having this property are called super-diffusive. However, it has been measured that the radius of gyration for human grows in time slower than for the Brownian motion, even slower than logarithmic growth. In fact, human tend to come back home and to known places, what leads to an ultra-slow diffusion (Song et al., 2010a). Such properties can be obtained by assuming a cutoff in the distribution of jump lengths and adding memory effects to mobility model. The specific form of the distribution and the cutoff depends mainly on the time interval used for the measurements and on the system under study.

Several more advanced models have been introduced. For instance, the asymmetry of the travels was studied by considering ellipsoidal boundaries to the average individual displacements and analyzing the scaling of the radius of gyration (González et al., 2008). Memory effects in the individual displacements were also analyzed, finding that individuals' home and workplace have a considerable

	TOTAL($\times 10^3$)		US($\times 10^3$)		UK($\times 10^3$)		DE($\times 10^3$)	
	N	L	N	L	N	L	N	L
Twitter	714	15,000	132	1,100	28	117	3.8	8.5
Gowalla	196	950	46	350	5.2	20	5.2	30
Brightkite	58	214	27	167	3.1	10	1.3	7.2

Table 5.1: **Datasets.** Number of users (nodes) N and of links L of the networks obtained from the different geo-localized datasets for the United States (US), the United Kingdom (UK) and Germany (DE).

impact on their mobility patterns (Song et al., 2010a). These results motivated the introduction of several mobility models with the aim of explaining the features observed in the data (Song et al., 2010a; Simini et al., 2012; Jia et al., 2012; Szell et al., 2012; Hasan et al., 2012). For the purpose of this study, we couple the basic mobility models with the network growth models.

5.2 Datasets

We have collected data from OSNs containing both social links and information about the users' physical positions. The first dataset was obtained from Twitter by means of its API.[1] We identify over 714,000 single users, who tweeted using a GPS enabled mobile device during the month of August 2011 (Ratkiewicz et al., 2011). If those users reported various locations in different tweets, the most recent one is taken for the purpose of the study. The other two datasets contain information referring to the users' location check-ins and the social networks of Gowalla and Brightkite (Cho et al., 2011). Both were location-based OSNs, in which users can check-in at their current locations and receive information about services in the area as well as about their friends' positions. Gowalla and Brightkite are no longer active but the data is available online.[2] The main statistical features of our three datasets are displayed in Table 5.1.

Social interactions across country borders have particular properties and are affected by political, linguistic or cultural factors. We overcome this difficulty by restricting our analysis to the networks within each country. Intra-country mobility and social contacts account for the large majority of user activity (State et al., 2013; Ugander et al., 2011). For simplicity, we focus on the three major

[1] See Twitter API at `https://dev.twitter.com`

[2] Data available at the Stanford large network dataset collection, `http://snap.stanford.edu/data`

countries with more than one thousand users in each of our datasets: the United States (US), the United Kingdom (UK) and Germany (DE). Similar results are found for other countries.

5.3

The TF model

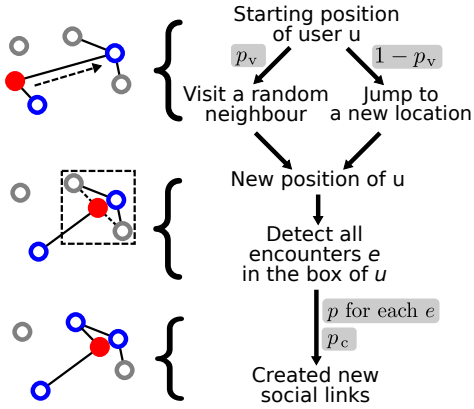


Figure 5.2: **Schematic of the TF model.** The central node is the filled red circle and its neighbors are marked in blue. Directionality of links is neglected in this schematic to maintain simplicity.

The model structure is illustrated in Figure 5.2. The initial condition is a set of individuals located in the last known positions of the online network users as extracted from the data. At each step of the model, a randomly chosen agent performs two actions:

1. Travel
   (a) Visit a randomly selected friend at his current location with probability p_v.
   (b) Otherwise, travel to a new location. The distance of travel is obtained from a distribution of jump lengths, while the direction is chosen proportionally to the population density at the target distance.
2. Friendship
   (a) With probability p, create directed links to agents within a neighborhood of size $\delta \times \delta$.

(b) With probability p_c, create a directed connection to a randomly chosen agent anywhere in the system.

The model is iterated until the number of created connections is equal to the number of links measured in the empirical networks. Despite its simplicity, the model incorporates several major features of human behavior. The *travel* component accounts for both recurring visits to the same location and exploration of new places and the *friendship* component generates both face-to-face contacts and online acquaintances independent of the geography. In the reminder of this chapter we refer to this model as the TF model.

The model has four parameter p_v, p and p_c and δ, and the distribution of jump lengths. Following Ref. (Song et al., 2010a) we consider a power-law distribution of jump lengths with an exponent of -1.55. The values of the probability $p = 0.1$ and the box size $\delta = 0.001^\circ$ are chosen to reproduce the dependence of the probability of friendship link on the number of daily spatio-temporal coincidences of users measured in Flickr (Crandall et al., 2010). To this end, we assume that one time step of the model corresponds roughly to one day and we obtain a good agreement,[3] as shown in Figure 5.1. We have tested different parameter values and we did not observe any strong differences in the results; different shapes of the jump distributions are discussed in Section 5.7. The other two parameters, p_v and p_c, are explored in the remainder, since, as it will be shown, they are essential for the final model result. The effect of each of the underlying assumptions is systematically explored through analysis of model variants in Section 5.9.

5.4 Geo-social properties of the networks

In this section we report six different geographic and social properties of spatial networks and measure them in Twitter, Gowalla and Brightkite, separately for the different countries. The results for the United States are presented in Figure 5.3, while the results for the United Kingdom and Germany are shown in Figures 5.4.

We start by comparing the empirical networks with those generated by the TF model using a set of metrics. First, we measure the probability of two users to have a link at a certain distance $P_\mathrm{l}(d)$. Defined as the ratio between the number of existing links at distance d and the total number of users pairs separated by d, $P_\mathrm{l}(d)$ it is constrained to always lie in the interval $[0, 1]$. It decays slowly with the distance as a power-law with exponent -0.7, which is followed by a plateau for very large distances (see Figure 5.3A). This functional shape remains identical

[3] Most of our simulations finish in less than a $1,000$ time steps, corresponding to a few years, which is of the order of magnitude of users' lifetime.

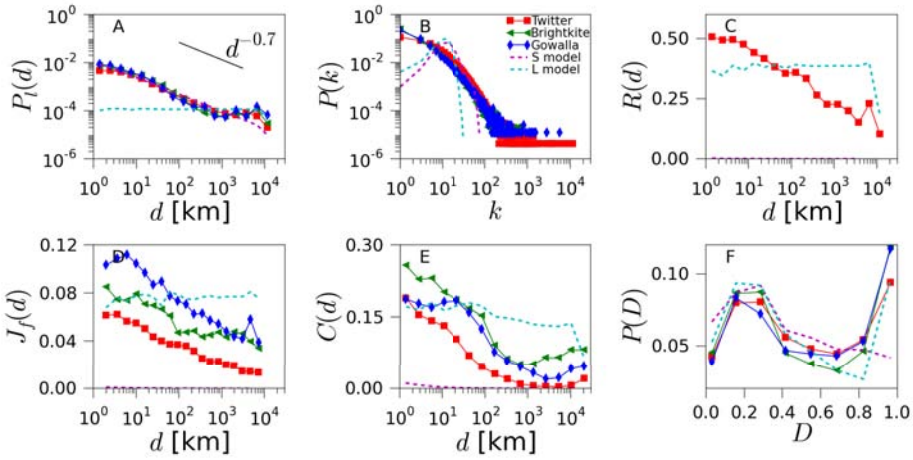


Figure 5.3: **Network geo-social properties.** Various statistical network properties are plotted for the data obtained from Twitter (red squares), Gowalla (blue diamonds), Brightkite (green triangles) and the null models (dashed lines), for the US (see Figures 5.4 for the UK and Germany). The spatial model (magenta), based on geography, matches well the data in $P_l(d)$, but yields near-zero values for $R(d)$, $J_f(d)$ and $C(d)$. The linking model (cyan), based on triadic closure, produces enough clustering, but it does not reproduce the distance dependencies of $P_l(d)$, $R(d)$, $J_f(d)$ and $C(d)$.

for all the countries and all the datasets considered (Figures 5.4A) and matches the behavior reported in the literature for online social systems (Liben-Nowell et al., 2005; Scellato et al., 2011).

A second metric that we consider is the degree distribution of the social networks (Figure 5.3B). For Twitter, which has a directed network, we consider the degrees of its symmetrized version. The distribution $P(k)$ displays heavy tail in all the datasets, even though there are slight differences between them.

Connections in Twitter typically are not reciprocal (Kwak et al., 2010). Reciprocated connections indicate mutual interest between the two users and a closer type of social relation (Gonçalves et al., 2011; Grabowicz et al., 2013a). To assess how geography and reciprocity correlate, we measure the fraction of reciprocated connections as a function of distance $R(d)$ (Figure 5.3C). We find that the reciprocity decreases with the distance in all the countries analyzed. This trend is consistent with the idea that stronger relations occur close to where users spend most of their time, with some longer connections composed of friends who moved, former residences, online acquaintances, etc. Furthermore, long not-reciprocated

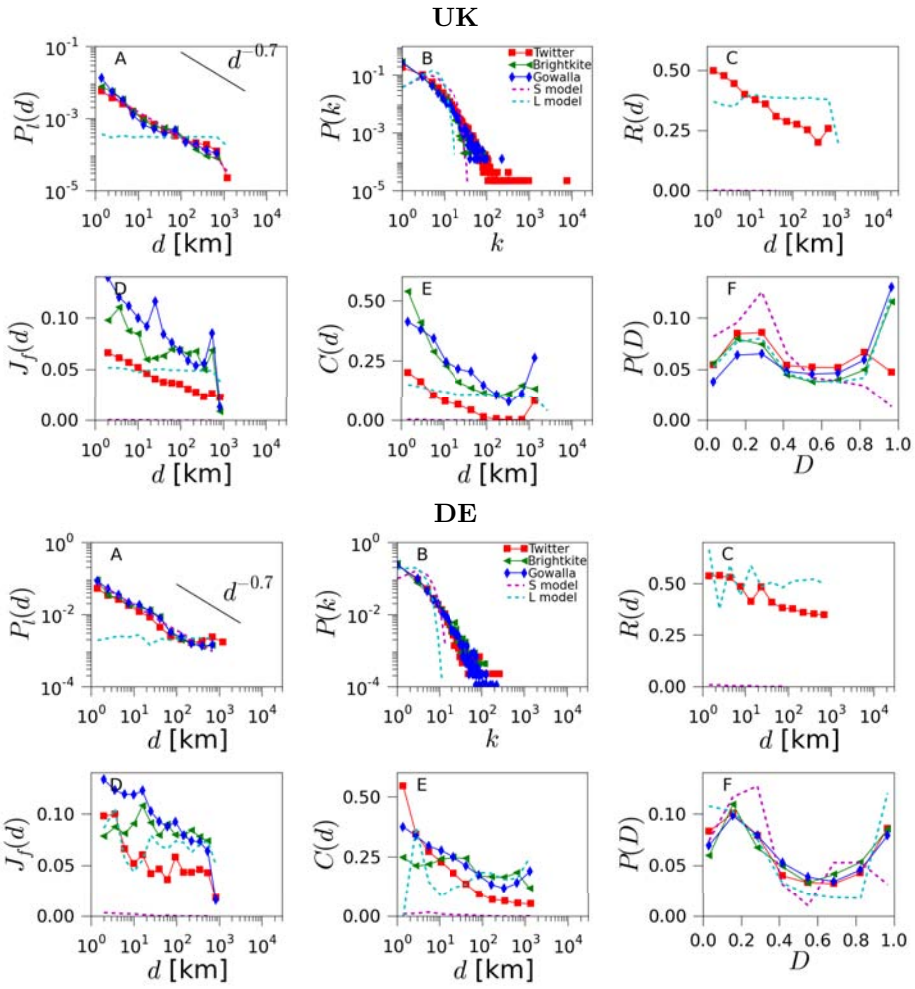


Figure 5.4: **Networks geo-social properties.** Various statistical network properties are plotted for the data obtained from Twitter (red squares), Gowalla (blue diamonds), Brightkite (green triangles) and the null models (dashed lines), for the UK and Germany.

connections may include users following public figures or celebrities.

With the aim of quantifying social closeness between users, we define the social overlap J_{f} of two connected users i and j as

$$J_{\mathrm{f}} = \frac{|\mathcal{K}_i \cap \mathcal{K}_j|}{|\mathcal{K}_i \cup \mathcal{K}_j| - 2} \tag{5.2}$$

where $\mathcal{K}_i$ represents the set of friends of user i. This property is a slightly modified Jaccard index that accounts for the fact that a node i does not have itself among its neighbors, while the node j has i among its neighbors, and vice versa. Without this modification the index could not reach its maximum value of 1. In Figure 5.3D, the average of the social overlap $J_{\mathrm{f}}(d)$ over all pairs of connected users is plotted as a function of the distance between them. The social overlap decreases with the distance. The functional shape of the curves is similar for all the datasets, even though the overlap level is different for each of them. For Twitter, we use the symmetrized version of the network to study social overlap and clustering.

Another well known phenomenon in social networks is triadic closure. As one individual has a close relation with other two persons, there are high chances that these two individuals end up creating a social relation between themselves. In network analysis, a magnitude that quantifies this effect is the average clustering coefficient C. The effect of the distance can be incorporated by measuring the distances from each central node i to two neighbors j and k forming a triad, $d = d_{ij} + d_{ik}$, and calculating the network clustering restricted to triads with distance d. This new function $C(d)$ is the probability of closing a triangle given the distance d in a triad

$$C(d) = \frac{\Delta(d)}{\Lambda(d)}, \tag{5.3}$$

where $\Lambda(d)$ and $\Delta(d)$ are the numbers of triads and closed triads for the distance d, respectively. The value of the global clustering coefficient C can be recovered by averaging $C(d)$ over d. In the datasets, we observe a drop in $C(d)$ followed by a plateau, which is best visible for the US networks (Figure 5.3E).

Given a triangle, several configurations are possible if there is diversity in the edge lengths. The triangle can be equilateral, if all the edges have the same length, isosceles, if two have the same length and the other is smaller, etc. We estimate the dominant shapes of the triangles in the network by measuring the disparity D defined as:

$$D = 6\left[\frac{d_1^2 + d_2^2 + d_3^2}{(d_1 + d_2 + d_3)^2} - \frac{1}{3}\right], \tag{5.4}$$

where d_1, d_2 and d_3 are the geographical distances between the locations of the users forming the triangle. The disparity takes values between 0 and 1 as the

shape of the triangle passes from equilateral to isosceles, where one edge is much smaller than the other two. D shows a distribution with two maxima in the OSNs (Figure 5.3F), for low and high values. The two most common geometries of the triangles are: i) all 3 users are at a similar distance, ii) 2 users are close to each other, while the third one is distant. Since most edges correspond to small distances, this means that most triangles are constituted by three users who are all close to each other geographically. However, the stretched isosceles configuration is also relatively common.

Summarizing, we have defined the following metrics in order to characterize the networks structure and its relation to geographical distance:

- $P_{\mathbf{l}}(d)$: Probability of linking at a distance d (Figure 5.3A).
- $P(k)$: Degree distribution (Figure 5.3B).
- $R(d)$: Reciprocity as a function of the distance (Figure 5.3C).
- $J_{\mathbf{f}}(d)$: Average overlap as a function of the distance (Figure 5.3D).
- $C(d)$: Clustering coefficient as a function of the triad distance (Figure 5.3E).
- $P(D)$: Distribution of distance disparity for the triangles' edges (Figure 5.3F).

We will use these metrics in the coming sections to estimate the ability of model to reproduce social networks comparable with those obtained from the empirical datasets.

5.5 Model fitting

Next, we will find a compromise between the different metrics and search for the parameter values for which a given model best fits simultaneously the various statistical properties. To do so, we define an overall error E to quantify the difference between the networks generated with the model and the empirical ones. The parameters of the model are then explored to find the values that minimize E. We measure the error $E[X]$ for each property X and take the average over all the properties

$$E = \frac{1}{8}\Big\{E[P_{\mathbf{l}}(d)] + E[P(k)] + E[R(d)] + E[J_{\mathbf{f}}(d)] + E[C(d)] + E[P(D)] + E[N_{\mathbf{c}}] + E[C_{\mathrm{avg}}]\Big\}, \tag{5.5}$$

where N_c is number of nodes in all connected components of the network and C_{avg} is the undirected local clustering coefficient averaged over the N_c connected nodes. The properties X integrating E can be scalars, functions or distributions and encompass different orders of magnitude. We define the error of a property X as

$$E[X] = \frac{\sum_{i=1}^{n} \left| y_i^X - f_i^X \right|}{\sum_{i=1}^{n} \left| y_i^X \right|}, \tag{5.6}$$

where y_i^X is the i-th observed value of the property X, f_i^X is the corresponding i-th value of the property obtained by the model. In the case of a distribution, i runs over the n measured bins, while for a scalar (such as the number of nodes or the clustering coefficient) the sum has only one term.

5.5.1 Parameter estimation

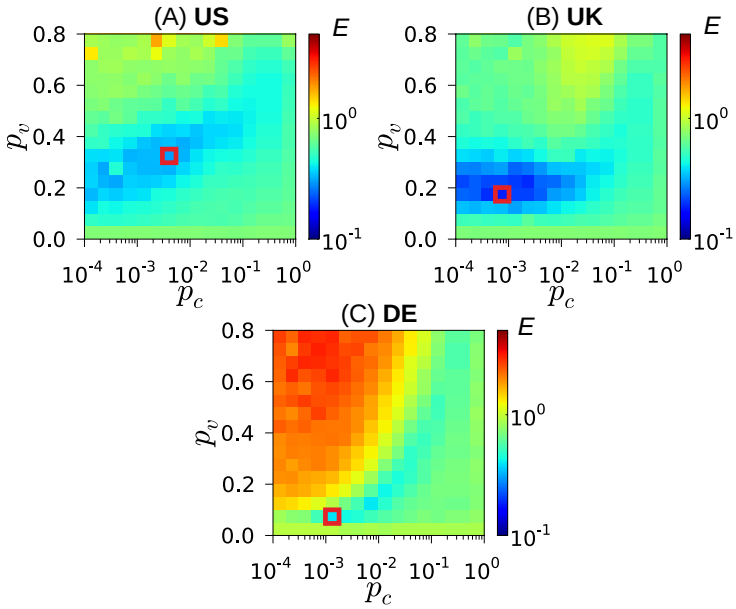


Figure 5.5: **Fitting the TF model.** Values of the error E when p_v and p_c are changed. The minimum error for each of the plots is marked with a red rectangle.

We perform a Latin square sampling of the parameter space of p_v and p_c as shown in Figure 5.5 in order to find the minimum value of E. The parameter

space is covered uniformly in a linear scale for p_v and in a logarithmic one for p_c. For all the countries, the minimum value of the error is obtained for p_v in the interval $(0.05, 0.3)$ and p_c in the range $(10^{-3}, 10^{-2})$. The values of E found at the minimum are 0.30 for the US, 0.18 for the UK and 0.39 for Germany. For simplicity, we focus on the Twitter networks only, although similar results are obtained for the other datasets.

5.5.2 Simulations for the optimal parameters

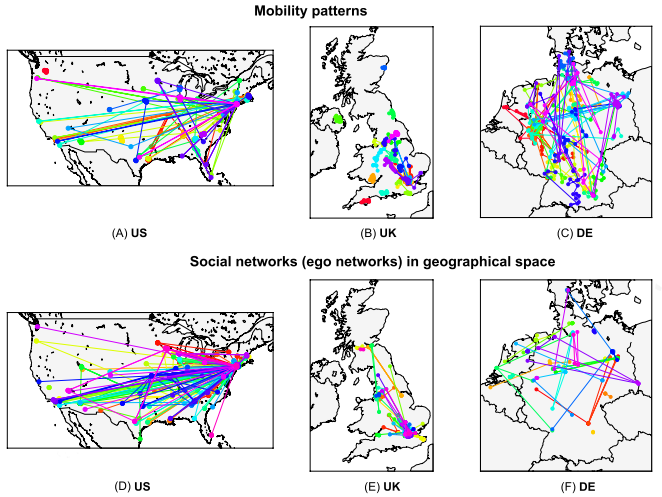


Figure 5.6: **Simulation results: mobility and social networks.** Mobility (upper row) and ego networks (lower row) of 20 random users (different colors) for the instances of the TF model yielding the lowest error E (see Figure 5.5).

An example with the displacements between the consecutive locations and the ego networks for a sample of individuals, as generated by the TF model, are displayed in Figure 5.6. The parameters of the model are set to the ones that correspond to the minimum of the error E. As shown, the agents tend to stay close to their original positions. Occasional long jumps occur due to far friend visits. In this range of parameters and simulation times, the main mechanism for initiating long distance connections is random linking (controlled by p_c). Agents typically return back to their original positions because this is where most of their contacts live. The frequency of the long distance jumps and connections varies for the three countries due to the different spatial distribution of the user

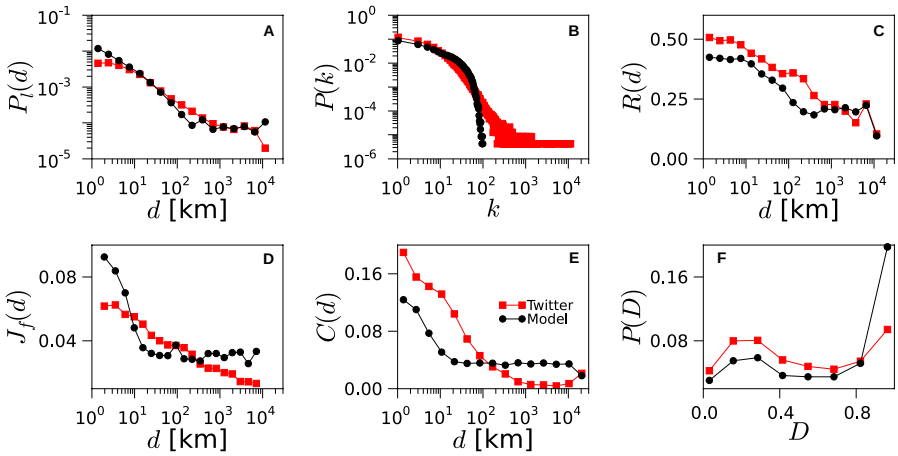


Figure 5.7: **Geo-social properties of the model networks.** Various statistical properties are plotted for the networks obtained from Twitter data (red squares) and from simulation of the TF model (black circles) for the US (for the UK and DE, see Figures 5.8).

populations. In the ego networks, the presence of multiple triangles with long distance edges can be observed.

The geo-social properties of the networks generated by the TF model are shown in Figure 5.7 for the US and in Figures 5.8 for the UK and Germany, respectively. The model is able to reproduce the trends in the probability $P_{\mathrm{l}}(d)$, the reciprocity $R(d)$, the social overlap $J_{\mathrm{f}}(d)$ and the disparity distribution $P(D)$ with good accuracy. The difficulties encountered with the degree distribution $P(k)$ and the clustering as a function of the distance $C(d)$ are not unexpected since the model does not incorporate mechanisms to explicitly enhance the heterogeneity in the agents' contacts nor favor any specific dependence of the clustering on the distance.

5.6 Insights of the TF model

In this section, we explore two null models to help us interpret the mechanisms acting in the TF model. The first null model, the spatial model (S model), is based solely on the geography and consists of randomly connecting pair of users with a probability depending on the distance, but does not take network structure

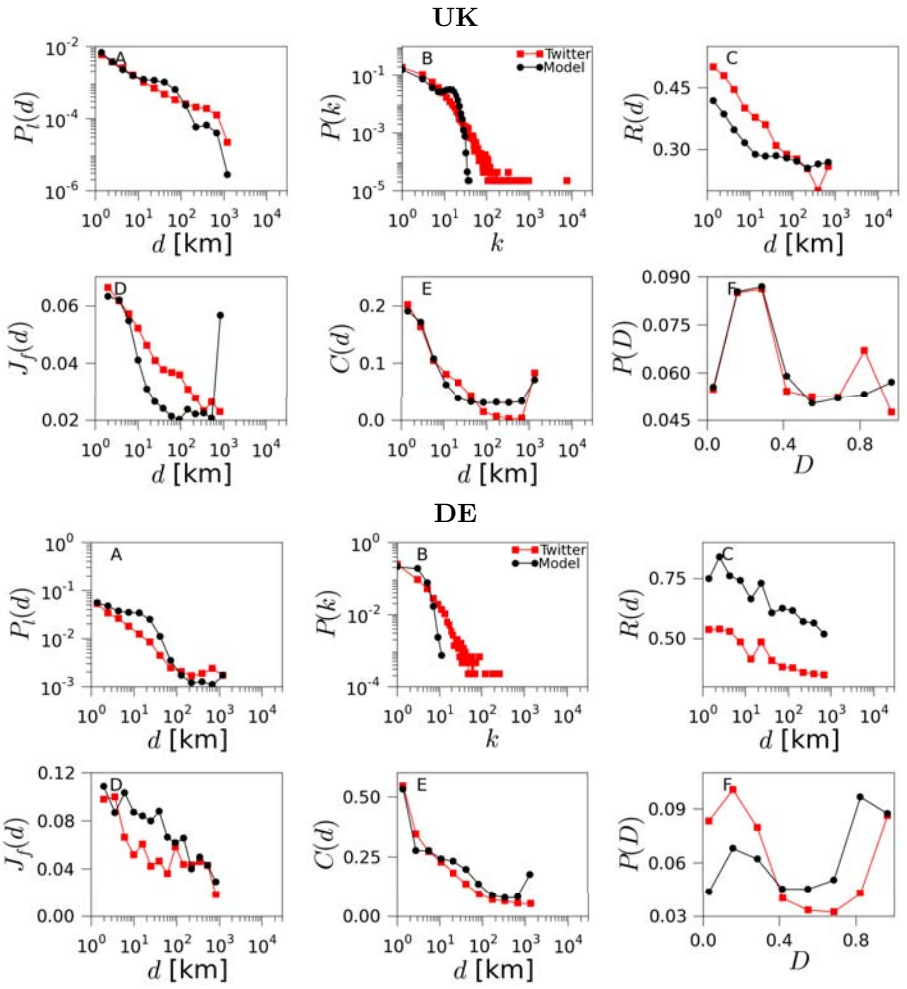


Figure 5.8: **Geo-social properties of the model networks.** Various statistical properties are plotted for the networks obtained from Twitter data (red squares) and from simulation of the TF model (black circles) for the UK and Germany.

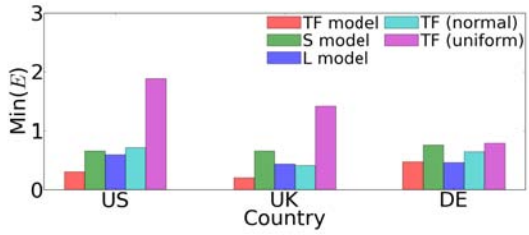


Figure 5.9: **Comparison of different models.** The minimal values of the error E for the TF model, the two null models: spatial (S model) or linking (L model), and the TF model with normally or uniformly distributed travel distances.

into account. The second null model, the linking model (L model), in contrast, is based only on random linking and triadic closure, and it is equivalent to the TF model without the traveling to new locations. We consider the two uncoupled null models and compare them with the TF model. Finally, we demonstrate the importance of the coupling through a realistic mobility mechanism.

The spatial model (S model) consists of randomly connecting pair of users with a probability that decays as power-law of the distance between them (suggested in (Butts et al., 2012)). The exponent of the power-law is fixed at 0.7 following Figure 5.3A. The results of the S model are shown in Figure 5.3. Although the model is set to match $P_{\mathrm{l}}(d)$, other properties such as $P(k)$, $R(d)$, $J_{\mathrm{f}}(d)$, $C(d)$, or $P(D)$ are not well reproduced. The S model fails to account for the high level of clustering and reciprocity in the empirical networks and for their dependence on the distance. The error E of this null model is between 0.66-0.76 for the three countries, around twice the error of the TF model (see Figure 5.9).

The linking model (L model) is a simplified version of the TF model, without random mobility and the box size $\delta \to 0$. Agents move to visit their contacts with probability p_{v}, whereas with probability $1 - p_{\mathrm{v}}$ they do not perform any action. In this version of the model, users connect only by random connections or when directly visiting each other, what leads to triadic closure. These two processes do not depend on the distances between the users. A thorough description can be obtained with a mean-field approach (see Section 5.8). The results of the L model are shown in Figure 5.3. Due to the triangle closing mechanism this null model creates networks with a considerable level of clustering. However, it does not reproduce the distance dependencies of $P_{\mathrm{l}}(d)$, $R(d)$, $J_{\mathrm{f}}(d)$ and $C(d)$. The error E of the L model also tends to be twice higher than the error of the TF model (see Figure 5.9).

The geography and the structure are coupled in the TF model through the

random mobility. Changes in the underlying mobility mechanism affect the quality of the results. The lowest E values are obtained with the power-law distribution in the jump lengths. The normal or uniformly distributed jumps yield worse results, increasing the error E by a value from 0.2 to 1.5 (Figure 5.9). These cases are described in more detail in Section 5.9.

In summary, simplified models that neglect either geography or network structure perform considerably worse than the TF model in reproducing the properties of real networks. Likewise, non-realistic assumptions on human mobility mechanism yield worse results than the default TF model. To conclude, the coupling of geography and structure through a realistic mobility mechanism produces networks with significantly more realistic geographic and structural properties.

5.7 Sensitivity of the TF model to the parameters and its modifications

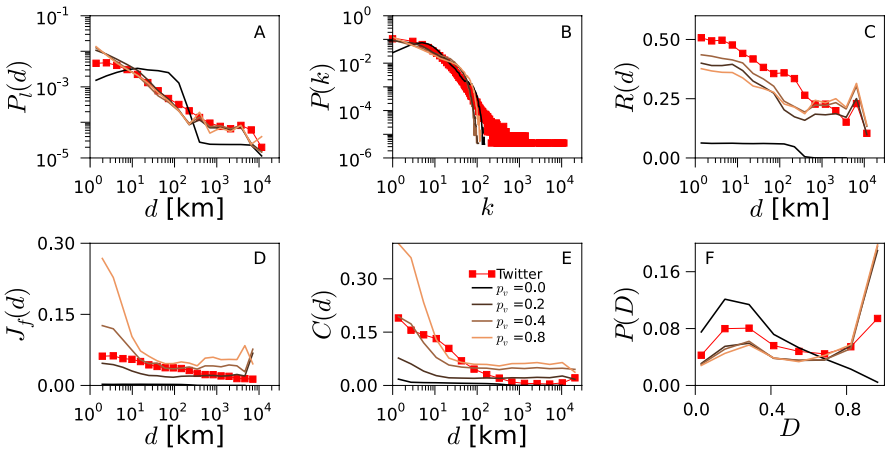


Figure 5.10: **Impact of p_v on the TF model.** We change the value of p_v while keeping p_c fixed to the optimal value, for the US network (for the UK and Germany see Figure 5.12). Note that this corresponds to an exploration of the parameter space along the vertical line crossing the minimum of E as plotted in Figure 5.5 for the US.

The results presented so far have been obtained at the optimal values of p_v

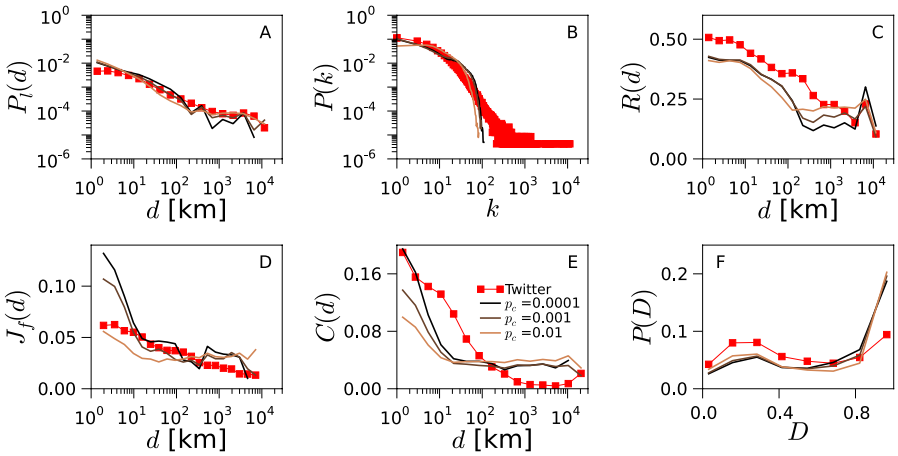


Figure 5.11: **Impact of p_c on the TF model.** We change the value of p_c while keeping p_v fixed to its optimal value, for the US network (for the UK and Germany see Figure 5.13). Note that this corresponds to an exploration of the parameter space along the horizontal line crossing the minimum of E as plotted in Figure 5.5 for the US.

and p_c. The question remains, however, of how robust these results are to changes in the values of the parameters.

In Figure 5.10, we report the effect of varying p_v while p_c is maintained constant in its optimal value. The linking probability $P_l(d)$ loses its power-law shape for very low values of p_v, marking the limit in which random mobility is the main mechanism for the agents' traveling in detriment of friend visits. In this case, most of the links are created due to encounters occurring in nearby locations or are random connections, and so the distribution of triangles disparity $P(D)$ loses its bimodal shape. Furthermore, the friend visits provide opportunities to reciprocate the connections. This is why for extremely low values of p_v, the reciprocity $R(d)$ is close to zero. Towards the other limit, i.e., $p_v \to 1$ the social overlap $J_f(d)$ and the triangle-closing probability $C(d)$ steadily increase from near-zero values. In this limit, the linking probability $P_l(d)$, the reciprocity $R(d)$ and the distribution of triangles disparity $P(D)$ recuperate their shapes of the optimum.

In Figure 5.11, we explore the impact of varying p_c while p_v is fixed to its optimal value. The effect of p_c on $J_f(d)$ and $C(d)$ is the opposite to that of p_v: these metrics decrease at all distances with increasing p_c. The reason for this is that visits to friends are the main forces behind the creation of new triads and

the subsequent closure of triangles. Note that the more connections are created randomly (higher p_c), the less links will be a result of friend visits. We will expose and describe in detail the interplay between these two mechanisms in the mean-field calculations.

A possible variation of the TF model consists of eliminating friend visits or random connections (i.e., setting p_v or p_c to 0). This prevents the model from producing networks with characteristics comparable to the real ones in all the cases, leading to increase in E of around 0.5 (see Section 5.9).

5.8 Mean field approach

In this section we consider the L model, introduced in Section 5.6, to gain analytical insight. Although this model is a simplified version of the TF model, the results of the simulations yield a relatively low value of E (shown in Figures 5.9, 5.15 and 5.16). We write the equations for the time evolution of the properties of the network and solve them numerically. Among all the properties, we focus on the average clustering coefficient C, the overall reciprocity R and the degree distribution $P(k)$.

The clustering coefficient is defined as a ratio of all closed triads to all triads existing in the network, i.e., $C = \Delta/T$. The number of triads Λ can be calculated knowing the degree distribution. The number of closed triads Δ in the L model grows with time mostly due to the friend visits mechanism. A triangle is formed every time two friends of the same hosting agent meet in the host's place and decide to connect. Note that an undirected triangle corresponds to 3 undirected closed triads. Assuming that the contribution of random links is negligible, the time evolution of the number of closed triads is described by

$$\frac{d\Delta}{dt} = 3\,N\,(k>0)\,\left(1-(1-p)^2\right)(1-C)\,M\,S,, \tag{5.7}$$

where $k = \left(k^{\text{in}} + k^{\text{out}}\right)/2$, meaning that we do not distinguish between in-degree and out-degree; $N\,(k>0)$ represents the number of nodes with the degree higher than 0, i.e., the number of potential hosts; M is an estimate of the lower bound for the number of triangles closed by one closing link $M = 1 + C^2\,\left(\frac{2}{1+R}\,k - 2\right)$; and S is the expected number of encounters per host, which can be calculated as

$$S = \sum_{k=2}^{\infty} \frac{N\,(k)}{N} \sum_{i=2}^{k} \left(\frac{p_v}{\langle k\rangle}\right)^i \left(1-\frac{p_v}{\langle k\rangle}\right)^{k-i} \binom{k}{i}\binom{i}{2}, \tag{5.8}$$

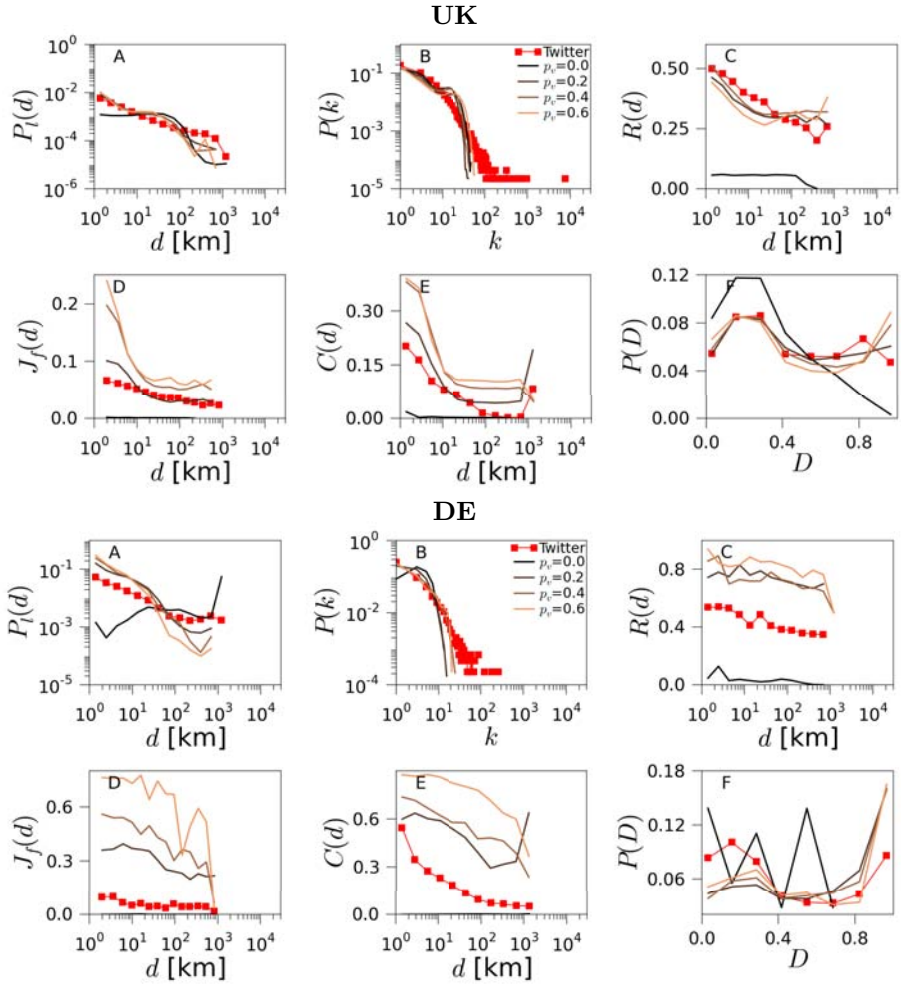


Figure 5.12: **Impact of p_v on the TF model.** We change the value of p_v while keeping p_c fixed to the optimal value, for the UK and Germany.

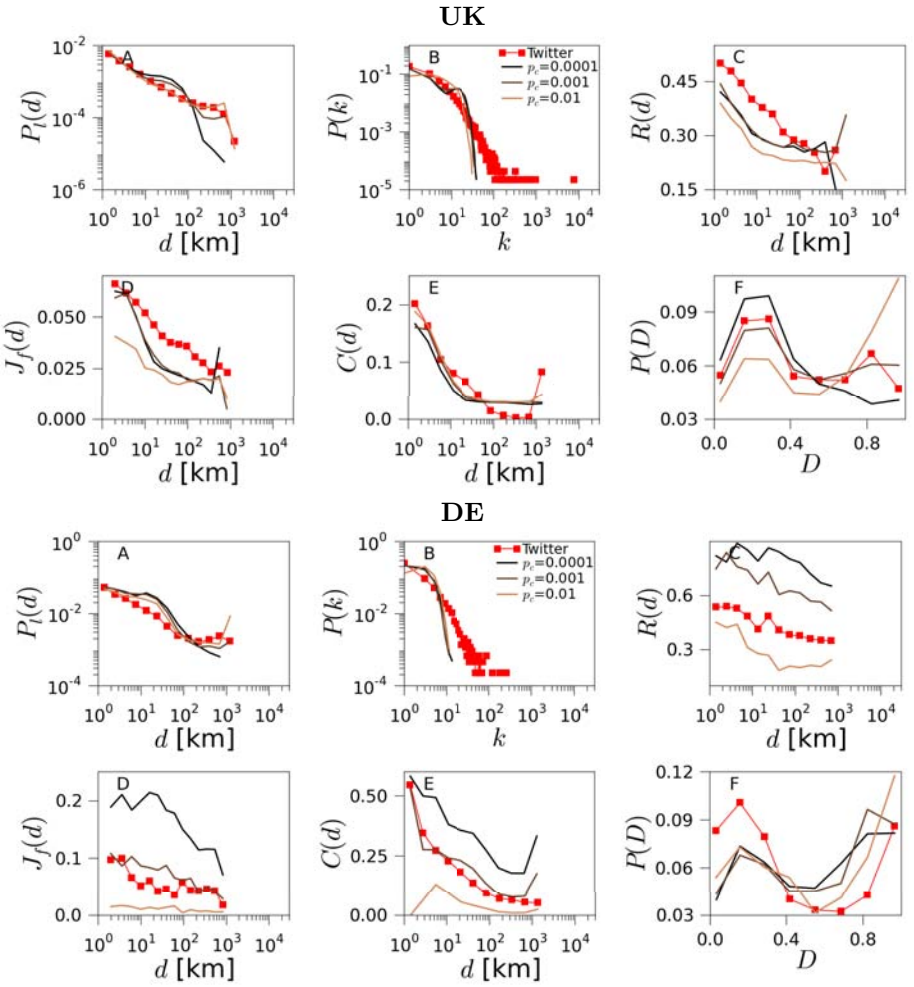


Figure 5.13: **Impact of p_c on the TF model for the UK and Germany.** We manipulate the value of p_c, while keeping p_v fixed to its value from the best fit, for the UK and Germany

where $N(k)$ is the number of nodes with a given degree k in the network. Finally, note that the above definition of degree and the one obtained from symmetrizing directed networks (used in previous sections) are related by a proportionality factor $k = k_{\mathrm{sym}}(1+R)/2$.

The reciprocity of connections R can be expressed as $R = L_{\mathrm{rec}}/(L_{\mathrm{rec}}+2L_{\mathrm{nrec}})$, where L_{rec} is the number of reciprocated links, L_{nrec} is the number of non-reciprocated links and the total number of links $L = L_{\mathrm{nrec}} + L_{\mathrm{rec}}$. The numbers of links evolve as

$$\begin{aligned} \frac{dL_{\mathrm{rec}}}{dt} &= 2\,N(k>0)\,\{p_{\mathrm{rec}} + p^2\,(1-C)\,S + p\,(1-R)\,C\,S\}, \\ \frac{dL_{\mathrm{nrec}}}{dt} &= p_{\mathrm{c}}\,N + \frac{1}{3M}\,\frac{d\Delta}{dt} - \frac{1}{2}\,\frac{dL_{\mathrm{rec}}}{dt}, \end{aligned} \tag{5.9}$$

where $p_{\mathrm{rec}} = p\,p_{\mathrm{v}}\,(1-p_{\mathrm{v}})\,(1-R)$ corresponds to the probability that an agent visiting a neighbor gets her connection reciprocated (their connection is initially single directional). As can be seen, Δ, L_{rec} and L_{nrec} are mutually dependent.

To calculate the degree distribution $P(k)$, we estimate the probability p_{con} of a node to increase its degree by one unit in the current time step due to multiple encounters with friends of her friends

$$p_{\mathrm{con}} = \sum_{k'=2}^{\infty} \frac{k'\,N(k')}{\langle k\rangle N} \binom{k'-1}{2} p_{\mathrm{c}}^2\,(1-p_{\mathrm{c}})^{k'-2}, \tag{5.10}$$

where $p_{\mathrm{c}} = p\,p_{\mathrm{v}}/\langle k\rangle\,(1-(1+R)\,/2C)$. In the L model, however, every node can increase its degree by multiple links at each time step. For simplicity, we neglect higher order terms induced by the possibility of creating multiple links. Moreover, we note that Equation (5.10) is a good estimate if there is not a strong correlation between node degrees. The number of nodes of certain degree k is given by

$$\begin{aligned} k>1: \quad \frac{dN(k)}{dt} &= p_{\mathrm{inc}}\,(N(k-1) - N(k)), \\ \frac{dN(1)}{dt} &= p_{\mathrm{c}}\,N(0) - p_{\mathrm{inc}}\,N(1) + p_{\mathrm{rec}}\,N_{\mathrm{s}}(0), \\ \frac{dN(0)}{dt} &= -p_{\mathrm{c}}\,N(0) - p_{\mathrm{rec}}\,N_{\mathrm{s}}(0), \end{aligned} \tag{5.11}$$

where $p_{\mathrm{inc}} = p_{\mathrm{c}} + p_{\mathrm{rec}}/2 + p_{\mathrm{v}}\,p_{\mathrm{con}}$ is an estimate of the probability that the node degree increases, $N_{\mathrm{s}}(0)$ is the number of nodes with 0 out-degree and non-zero in-degree. Such nodes are important because their connection can be easily reciprocated as a result of a friend visit. However, these nodes are not counted directly into $N(1)$, and so a correction is needed to account for them explicitly,

as in Equation (5.11). The number of such nodes can be calculated as

$$\frac{dN_{\mathrm{s}}(0)}{dt} = p_{\mathrm{c}} N\,(0) - p_{\mathrm{rec}} N_{\mathrm{s}}\,(0)\,. \tag{5.12}$$

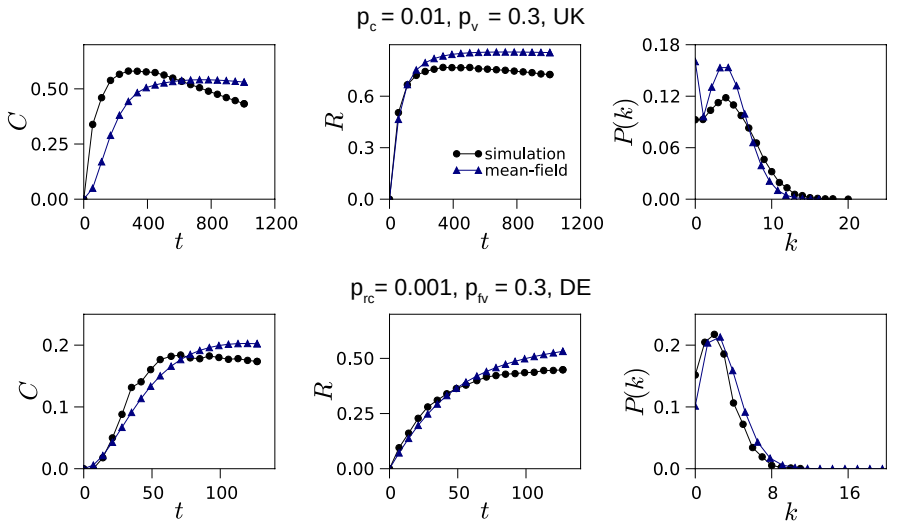


Figure 5.14: **Mean field approximation.** Predictions of the analysis versus results of the simulation of the L model for the clustering coefficient C, the reciprocity R and the degree distribution $P(k)$. In this case, we are taking the users from the UK and Germany because their lower numbers facilitate the numerical integration of the Equations 5.7, 5.9, 5.11 and 5.12

The numerical solution of this set of equations describing the evolution of the L model is shown in Figure 5.14. The equations accurately predict the dynamics of the clustering coefficient C, the reciprocity R and the degree distribution $P(k)$ for certain values of the parameters (i.e., for medium and high values of p_{c}, as in Figure 5.14A). The approximation yields slightly worse results when the number of random connections is small in comparison with the number of connections created due to friend visits (i.e., for low values of p_{c}, as in Figure 5.14B). In the latter case, neither the degree distribution is well approximated, probably due to the degree-degree correlations introduced through the friend visit mechanism.

5.9

Variants of the TF model

In this section, we consider several variants of the TF model and the L model and evaluate their results. We describe a total of 36 variants marked with different colors in the tables in Figures 5.15 and 5.16. For each variant we explore the space of the parameters p_v and p_c. We run the models for p_v from the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and p_c from $\{0, 0.0003, 0.001, 0.003, 0.01, 0.03\}$, yielding in total 48 parameter combinations. For each of the model variants, we find the parameters that minimize the fitting error E. We plot its value in Figures 5.15 and 5.16. In the following paragraphs we describe in detail each of the variants and its results.

First, we modify the jump size distribution to understand its impact on the geo-social properties. We consider the following cases: the default power-law jumps with exponent 1.55, a minimal jump length of 1 km and a cutoff at 100 km, as in (Song et al., 2010a) (the TF model), uniformly distributed random jumps up to 100 km (TF-uniform), normally distributed jumps (TF-normal) with standard deviation of 1 km, and no jumps to new locations at all (the L model). We plot the minimal fitting error of these cases in Figures 5.15 and 5.16 using different colors of the curves. The default power-law jumps show the best results with the lowest error for most of the variants. The wiener distribution and the L model tend to perform considerably worse. The highly unrealistic uniform jumps understandingly provide the worst results and the highest error values for almost all variants.

To assess the role of friend visits and random connections we turn on and off these two components by setting to zero the corresponding parameters p_v and p_c. We plot the results in Figures 5.15 and 5.16 with dashed and dotted lines. We observe significantly higher error values whenever one of these two components is turned off, for most of the model's variants, what demonstrates their importance for the TF model.

To prevent users from spreading into inhabited regions, we include in the TF model an angular preference for the jumps. Namely, the direction of each jump is chosen randomly with a probability proportional to the number of inhabitants present at the destination. Note that this does not affect in any way the length of the jumps, which is drawn independently beforehand. To estimate the population of the target area, we use the gridded population of the world[4]. To test how this angular preference impacts the results, we consider a variation of the models without it and compare the results. The two variants are included in the lowest

[4] The Gridded Population of the World and The Global Rural-Urban Mapping Projects, Socioeconomic Data and Applications Center of Columbia University, http://sedac.ciesin.columbia.edu/gpw.

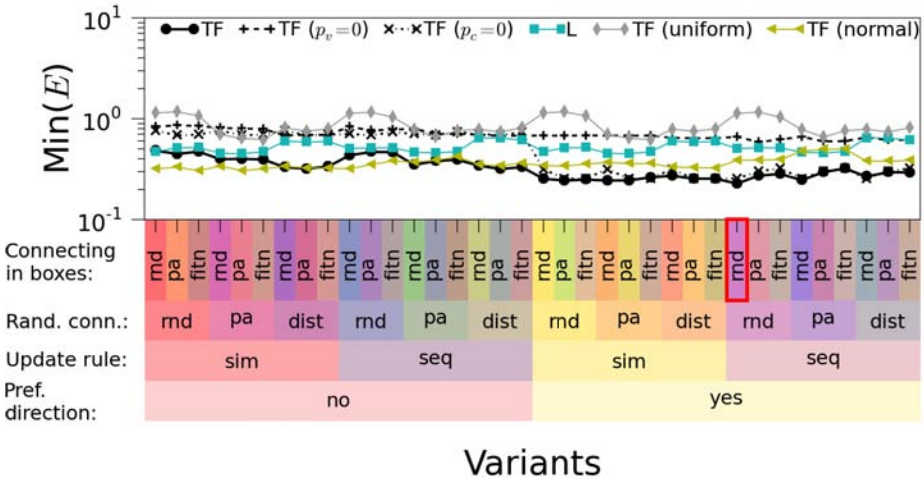


Figure 5.15: **The model variants.** Values of the fitting error E for the UK for the variants of the following models: the TF model, the TF model with $p_{\mathrm{v}} = 0$, the TF model with $p_{\mathrm{c}} = 0$, the L model and the TF model with uniformly or normally distributed jumps. The default variant described in Section 5.3 is marked with the red rectangle.

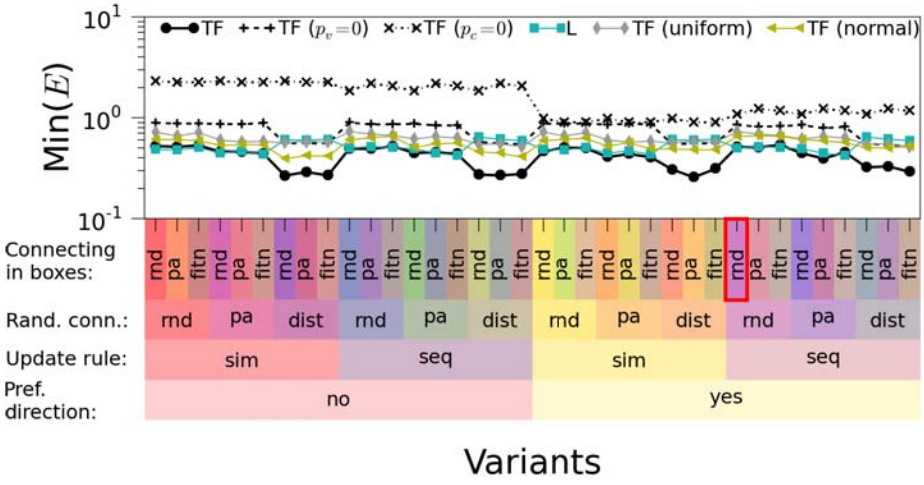


Figure 5.16: **The model variants.** Values of the fitting error E for Germany for the variants of the models as in Figure 5.15.

row of the table in Figures 5.15 and 5.16. They show almost no difference in the error values for the Germany, although a systematic difference exists for the UK; the error of the variant with direction preference is consistently lower in the case of that country. The presence of the sea around the UK introduces a distorting factor for the TF model. Without the directional preference the agents freely spread over the sea independently on the geographical shape of the country, leading to unrealistic results.

Agents' traveling and link creation can be realized in the simulation in various update orders. By default, at each time step, each agent first moves, next connects to other random agents, and then connects locally; the following agent performs the same actions in the same order, etc. We call this method sequential ("seq"). In an alternative update rule, which we call simultaneous ("sim"), first all the agents move, then all of them create random connections, and finally all of them create connections locally. The two update rules are included in the second row, counting from the bottom, of the table in Figures 5.15 and 5.16. The update rules have little impact on the final networks resulting from the simulation.

In the TF and L models, the agents create with probability p_c random connections. These links can be created in different ways; we consider three variants. First, each agent chooses another agent uniformly at random, what constitutes the default mechanism ("rnd"). Second, each agent randomly picks another node with probability proportional to the current degree of the node, which corresponds to the preferential attachment mechanism ("pa"). Third, the agent draws another node with probability decaying as a power-law of the distance between the two agents ("dist"), with its exponent equal to 1.4 and the minimal distance of 0.1 km. The type of random connecting mechanism used is listed in the third row, counting from the bottom, of the table in Figures 5.15 and 5.16. In some cases, e.g., for Germany, the distant-dependent probability of creating a random link provides better results than the other variants.

We consider similar variants for the connections formed inside spatial boxes, which are created with the probability p. The agents can connect uniformly at random ("rnd"), with a preference for high-degree nodes ("pa"), or a preference toward the nodes with high intrinsic fitness ("fitn"). The fitness of the nodes is drawn from a power-law distribution with an exponent of 1.5, which roughly corresponds to the distribution of the growth rates reported in Chapter 2. These variants are implemented in the following way. First, we note that the number of connections created by the agent is a result of a binomial process with probability p and the number of trials equal to the number of agents that currently stay in the given spatial box. The expected number of links created in such binomial process is known, therefore, an equivalent number of connections can be created with one of the two mentioned preferential processes. The type of connecting mechanism applied in the spatial boxes is listed in the top row of the table in Figures 5.15

and 5.16. There is no consistent difference in the error values between these variants. Thus, the connecting mechanism applied in the spatial boxes has little impact on the results.

We conclude that the main components of the TF models are crucial to reproduce the structure and geography of the social networks. These components include the mobility model, friend visits and random connections. The power-law mobility model tends to produce the best results. The angular preference of travels is important for countries whose geography is strongly restrained, e.g., by sea. Other modifications to the model have low or no consistent impact on the results, with the exception of the distance dependent random connections, which in certain cases consistently influence the results.

5.10 Conclusions

We introduce a model that couples human mobility and link creation in social networks. The aim is to characterize the relation between network topology and geography observed in empirical online networks. The model has two free parameters p_c and p_v but, despite its simplicity, it is able to reproduce a good number of geo-social features observed in real data at a country level. Due to comparisons with null models we find that the coupling of geography and structure through a realistic mobility mechanism produces significantly more realistic social networks than the uncoupled models.

Social links in our model are formed mostly with relational (due to triadic closure), and proximity (through spatio-temporal coincidences) (Rivera et al., 2010) mechanisms. Visiting friends helps to reinforce the existing relations and favors the closure of triads with particular properties regarding the distance balance of their edges. Random link creation accounts for online acquaintances or for historical face-to-face encounters as individuals move their residence from one city to another. Finally, individual random mobility allows the agents to explore new locations. Our results show that by establishing an appropriate balance between friend visits and random link creation, the model can reproduce the main features of OSNs, e.g., we show that $10\% - 30\%$ of the mobility has to be directed towards existing friends. We demonstrate that these are the fundamental mechanisms at play in the model.

The model is generic and functional for different datasets. Human mobility driven by social ties has impact on the modeling of disease spreading, and may improve its predictions. The model can also be used in simulations of processes that involve social networks and geography, e.g., simulations of opinion formation, language evolution, or responses of a population to extreme events. Moreover,

it can also be helpful to design network benchmarks with realistic geo-social properties to test, for instance, the scalability of technical solutions in social online networks related to geography of its physical infrastructure.

Chapter 6

Discussion and outlook

In this dissertation, we described quantitatively and modeled several aspects of online social systems. Using concepts and methods of network theory, statistical physics, and data mining we characterized the growth of groups and the structure, activity, and geography of several OSNs. We unveiled various statistical patterns and confronted them with different sociological theories finding a good correspondence between the online and offline worlds. Our studies contribute to the emerging field of computational social science through the introduction of models, metrics, and methods for complex social systems. We learn that heterogeneity plays important role in the growth of groups, that groups are correlated with the interactions of users of OSNs accordingly to the predictions of sociological theories, that the type of groups can be predicted based on metrics evaluating characteristics of group interactions, and finally that spatio-temporal coincidences generate realistic social networks.

The studies presented here were based on large datasets from several OSNs, which include information about social links, pairwise interactions, groups, content tags and geo-localizing tags. On the one hand, the level of completeness of the datasets allowed us to develop quantitative methods suited specifically for studies and predictions of social systems, e.g., we propose several metrics to describe group sociality and topicality. On the other hand, the large volume of information available in the datasets permitted us to derive statistics per links of different types and to find refined statistical patterns, e.g., we show that links internal to communities attract different interactions than links between communities. Finally, a dataset with high temporal resolution combined with a dataset of low temporal resolution but spanning all the elements in the system allowed us to describe the dynamics of the elements and its impact on system-wide properties, e.g., we show that groups grow linearly in time, whereas heterogeneity shapes statistical properties of the whole system of groups. To conclude, the

recent availability of fairly complete and large datasets capturing the temporal dynamics of social systems empowers researchers to perform detailed quantitative studies of such systems and leads to emergence of computational social science.

6.1 Heterogeneity and temporal dynamics in social systems

We showed that the heavy-tailed distribution of group sizes in an OSN is the result of the growth process based on heterogeneity. We contrasted the heterogeneity approach with a preferential growth model to discuss the shortcomings of the latter approach. Our study shows that a simple model based solely on heterogeneity explains the heavy-tailed distribution and other properties of the system better than a model based solely on preferential growth. In general, in real systems both mechanisms are coupled (Salganik et al., 2006; Wang et al., 2013). However, there is no consensus on what kind of model could be used in order to describe such interplay. The existing contributions in this direction, which couple preferential growth and heterogeneity, are still lacking and require further work (Bianconi and Barabási, 2001b,a). How do these two components influence each other? How could we measure the contribution of each of them in shaping statistical properties of real systems? These questions are still not answered and motivate the need for further research.

Furthermore, it is unclear whether the heterogeneity-based model of growth of declared groups in Flickr generalizes to groups from other systems. First, it depends on the type of groups, i.e., whether the groups tend to be social or topical. Social groups are limited in size due to human cognitive limits (Dunbar, 1992). Topical groups are not limited in size. Thus, the growth process of the two types of groups must differ. Second, the properties of the growth of groups depend on the growth of the system as a whole, e.g., whether the number of users in an OSN grows or is stable during the period of the observation. These matters can be addressed by future research.

The availability of data on time evolution of social systems has improved significantly in recent years. Large temporal datasets from social and communication networks have already proved that the dynamics of human behavior is bursty and complex (Barabási, 2005; Miritello, 2013) and that individuals adopt various social strategies (Miritello et al., 2013). Temporal networks have become one of the focal interests in network science (Holme and Saramäki, 2012). Such methods are promising for the studies of social systems.

6.2 Information diffusion and groups

News spread in OSNs in a way that depends on their topic (Romero et al., 2011) and the social network of its spreaders (Rodrigues et al., 2011), but it is hard to predict how will a given piece of content spread among users (Bakshy et al., 2011; Kooti et al., 2012). On the one hand, social contagion needs a better understanding. Two competing ideas suggest that social reinforcement is crucial for spreading of controversial concepts (Centola and Macy, 2007) or that structural diversity is important for adoption of services (Ugander et al., 2012). On the other hand, community structure influences the way news spread in social networks. This dissertation has shown that information diffusion tends to happen more often between groups. Furthermore, the way the news spread between groups affect the virality of the news (Weng et al., 2013). Therefore, the studies of groups are vital for understanding spreading processes in social networks. This thesis has contributed to the quantitative description and classification of groups. Given that we know how to distinguish between different types of groups, i.e., topical and social groups, we can study how these types of groups influence information diffusion and other processes happening in social networks.

In the study of common identity and common bond groups we show that high reciprocity correlates with high normalized entropy, and that both properties correlate with the perception of social group conceived by human labelers. This perception is based on higher cognitive abilities to understand semantics, such as friendly and personal sentiment or coherent topical alignment. Naturally, there exist more group characteristics that are quantifiable and correlate with the metrics that we introduced. Finding other metrics describing sociality and topicality of groups will allow better understanding of the two types of groups.

Given that we distinguish between topical and social groups, one could build community detection algorithms specifically aimed at detection of these two types of groups. Methods for finding topical groups could combine network science and machine learning approaches by clustering users into topics of their interests based on the social graph and their profiles. The methods should allow overlapping clusters, to reflect multiple interests per user, and should be fast enough to cluster millions of users. An early approach to solve this problem has already been suggested (Bhattacharya et al., 2014).

Finally, in our studies we described various types of interactions by representing them as separate networks. Advancements in the theory of multiplex networks can provide additional methods and tools to analyze social systems in which interactions of various types happen interchangeably.

6.3 Geography of social networks and modeling

In the last chapter of this dissertation, we show that triadic closure can be achieved by means of spatio-temporal co-occurrences with friends. Since high clustering in networks is reported to cause community structure (Foster et al., 2011), it would be interesting to find if our model creates a network with a community structure corresponding to the real one. In general, the relation between geography and groups has not been investigated in OSNs due to the lack of proper data and problems with the detection of viable communities. Furthermore, one can expect that topical and social groups exhibit different geographic properties, i.e., that social groups tend to be more often geo-localized. These questions may be tackled in the future with improved data availability and community detection techniques.

Finally, our model coupling mobility and tie formation is a generic model aimed at explaining geographic and structural properties of OSNs. While it reproduces the statistical features of the real networks, it is not valid for detailed predictions of each user movements. Our model has only two free parameters, in contrast to multi-parametric interference methods (Wang et al., 2011; Cho et al., 2011; Sadilek et al., 2012). Although there exists an open discussion on the validity of hypothesis testing in the wake of statistical inference based on correlations (Anderson, 2008), our modeling efforts can inform future studies leading to more accurate models and better understanding of social systems.

6.4 Outlook for computational social science

The recent availability of complete and large datasets capturing the dynamics of social systems has lead to the emergence of computational social science. Researchers start to perform studies of social systems on such a scale for the first time in the history, promising the development of quantitative theories for social systems. This thesis contributes to the development of such quantitative statistical description of social systems. The digital traces of human behavior not only allow the advancements of our knowledge about such systems but also making predictions for social systems based on this knowledge. Various studies showed that social systems are highly predictable, but the predictability has certain limits (Song et al., 2010b; Krumme et al., 2013). These estimates of the limits of predictability, however, are based on the datasets available at the moments of the studies. Given that the amount and the completeness of data on social systems keep growing, it is unclear what is the potential of computational social science

and what are the limits of predictability of social systems. The future research of social systems can clarify these matters.

Appendix I: Order statistics local optimization method of community detection

Here, we describe in more detail the clustering algorithm that is used in a couple of studies described in this dissertation, presented in Chapters 3 and 4. It is based on order statistics local optimization method, due to which it is called OSLOM. In fact, it is one of the most accurate algorithms to detect communities (Lancichinetti et al., 2011; Lancichinetti and Fortunato, 2009a). It uses statistical significance as a measure of quality of clusters (Lancichinetti et al., 2010), which is defined as a probability of finding the cluster in a random null model, namely the configuration model described in Section 1.3.1.

Imagine a graph $\mathcal{G}$ with N vertices and L directed edges. The goal is to assess the significance of a cluster $\mathcal{C}$. We consider a node i that belongs to $\mathcal{G} \setminus \mathcal{C}$, as shown in Figure 1. The total degree of the subgraph C is $K_{\mathcal{C}} = \sum_{j \in C} k_j$, the degree of node is k_i and the total degree of the rest of the network is $K_{\mathcal{G}\setminus\mathcal{C}}$. Each of these degrees can be split into the part that connects to $\mathcal{C}$ and to the rest of the network, i.e., k_i^{int}, k_i^{ext}, $K_{\mathcal{C}}^{\text{int}}$, $K_{\mathcal{C}}^{\text{ext}}$, $K_{\mathcal{G}\setminus\mathcal{C}}^{\text{int}}$, $K_{\mathcal{G}\setminus\mathcal{C}}^{\text{ext}}$, respectively. Keeping these degrees (both internal and external to $\mathcal{C}$) fixed and assuming that all the edges are drawn randomly, we calculate the probability that the node i has k_i^{int} neighbors in $\mathcal{C}$

$$p(k_i^{\text{int}}|i,\mathcal{C},\mathcal{G}) = A\frac{2^{-k_i^{\text{int}}}}{k_i^{\text{ext}}!\,k_i^{\text{int}}!\,(K_{\mathcal{C}}^{\text{ext}} - k_i^{\text{int}})!\,(K_{\mathcal{G}\setminus\mathcal{C}}^{\text{int}}/2)!}. \tag{1}$$

This equation is derived by enumerating the possible configurations of the graph assuming the fixed degrees. The factorials in the denominator express the respec-

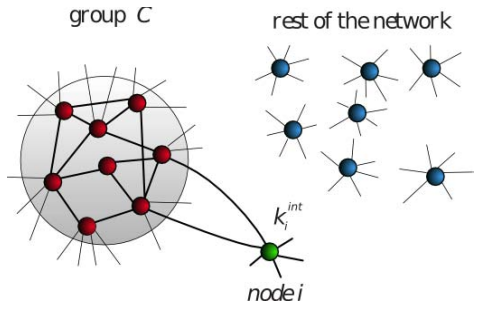


Figure 1: An illustration of the configuration model in a scenario with a group having fixed degree, which is considered in OSLOM to assess the statistical significance of the group. The probability of a node i to have k_i^{int} links connecting to the group is considered. Illustration adapted from (Lancichinetti et al., 2010).

tive connection combinations, while the power of 2 in the numerator corresponds to permutations of directionality of edges connecting i and $\mathcal{C}$. Finally, A is a normalization factor. The higher is the probability $p(k_i^{\text{int}}|i,\mathcal{C},\mathcal{G})$ the more likely it is that the node i has k_i^{int} links in a random scenario. The lower is the probability the less likely it is that the internal links are due to random factors. The nodes with the lowest $p(k_i^{\text{int}}|i,\mathcal{C},\mathcal{G})$ will be considered for inclusion in $\mathcal{C}$.

The ranking of nodes considered for addition to $\mathcal{C}$ is prepared as follows. First, one must calculate the cumulative probability $r(k_i^{\text{int}}) = \sum_{k=k_i^{\text{int}}}^{k_i} p(k|i,\mathcal{C},\mathcal{G})$ of having k_i^{int} or more internal connections to the group. Next, one ranks the cumulative probabilities from lowest to highest. The first candidate node to be included to the cluster has the lowest cumulative probabilities r_1. In fact, the variable r is a uniformly distributed random variable between zero and one for vertices of the null model. It follows that it is fairly easy to compute its order statistic distributions. The cumulative distribution of r_q in the null model is

$$\Omega_q(r) = P(r_q < r) = \sum_{i=q}^{N} \binom{N}{i} x^i (1-x)^{N-i}, \tag{2}$$

where $n_{\mathcal{C}}$ is the number of vertices in $\mathcal{C}$. One cannot assume that nodes inside the cluster follow the null model, because of the correlations between nodes belonging to the cluster that the method looks for. However, we can assume that the nodes that do not belong to the cluster follow the null model. For these nodes the statistics can be calculated. The values of $\Omega_q(r)$ inform us if the nodes external to the community follow the null model. To evaluate the quality of the cluster we calculate $c_m = \min_q(\Omega_q(r_q))$. Finally, the score of the cluster is defined as

the cumulative distribution $P(c_m^* < c_m) = \phi(c_m, N - n_{\mathcal{C}})$.

The algorithm optimizes the score by partitioning the network into a set of clusters. The initial clusters can be provided with another method at the start. The algorithm optimizes the scores of the clusters as follows. First, the algorithm considers node additions to each cluster by calculating the values of $\phi(c_m, N - n_{\mathcal{C}})$ for nodes external to the cluster that are connected to the cluster. Second, the method considers removals of the nodes by calculating the same score for nodes internal to the cluster, by treating them as if they were external. Finally, the algorithm merges communities if they become too similar. Only modifications and clusters whose score $\phi(c_m, N - n_{\mathcal{C}})$ is smaller than the statistical significance P are accepted. The statistical significance is an input parameter to the algorithm.

OSLOM takes directionality of links into account and detects overlapping communities, and it is one of the best performing methods in benchmarks (Lancichinetti and Fortunato, 2009a; Lancichinetti et al., 2011). Furthermore, this method can decide to leave a node without a group assignment in case it does not find any statistically significant community containing the node. It has been shown that nodes with random connections in a graph with bona fide group structure are detected by the algorithm as no-group nodes (Lancichinetti et al., 2011). For these reasons this method is useful in studies of groups in OSNs.

Bibliography

Adamic, L. A. and Huberman, B. A. (2000). Power-Law Distribution of the World Wide Web. *Science*, 287(5461):2115.

Ahn, Y. Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *Proc. 16th Int. Conf. World Wide Web - WWW '07*, pages 835–844, Banff, Alberta, Canada. ACM.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9:1–9:33.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Mag.*, 16:27–28.

Aral, S. and Van Alstyne, M. (2011). The Diversity-Bandwidth Trade-off. *Am. J. Sociol.*, 117(1):90–171.

Asur, S., Huberman, B. A., Szabo, G., and Wang, C. (2011). Trends in Social Media: Persistence and Decay. *SSRN Electron. J.*

Avnit, A. (2009). The Million Follower Fallacy. *http://tinyurl.com/nshcjg*.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks. In *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '06*, page 44, New York, New York, USA. ACM.

Backstrom, L., Kumar, R., Marlow, C., Novak, J., and Tomkins, A. (2008). Preferential behavior in online groups. In *Proc. Int. Conf. Web search web data Min. - WSDM '08*, pages 117–128, Palo Alto, California, USA. ACM.

Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. 19th Int. Conf. World wide web - WWW '10*, page 61, New York, New York, USA. ACM.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer. In *Proc. fourth ACM Int. Conf. Web search data Min. - WSDM '11*, page 65, New York, New York, USA. ACM.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proc. 21st Int. Conf. World Wide Web - WWW '12*, page 519, New York, New York, USA. ACM.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci.*, 106(51):21484–9.

Barabási, A. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–11.

Barabási, A.-L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Phys. A Stat. Mech. its Appl.*, 272(1-2):173–187.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Phys. A Stat. Mech. its Appl.*, 311(3-4):590–614.

Barrat, A., Barthelemy, M., and Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press, Cambridge.

Barthélemy, M. (2011). Spatial networks. *Phys. Rep.*, 499(1-3):1–101.

Bender, E. A. and Canfield, E. (1978). The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theory, Ser. A*, 24(3):296–307.

Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M. B., Ganguly, N., and Gummadi, K. P. (2014). Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale. In *Proc. ACM 2014 Conf. Comput. Support. Coop. Work - CSCW '14*.

Bianconi, G. and Barabási, A.-L. (2001a). Bose-Einstein Condensation in Complex Networks. *Phys. Rev. Lett.*, 86(24):5632–5635.

Bianconi, G. and Barabási, A.-L. (2001b). Competition and multiscaling in evolving networks. *Europhys. Lett.*, 54(4):436–442.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, 2008(10):P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Phys. Rep.*, 424(4-5):175–308.

Boguñá, M. and Pastor-Satorras, R. (2003). Class of correlated random networks with hidden variables. *Phys. Rev. E*, 68(3):036112.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.*, 99 Suppl 3:7280–7.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–8.

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892–5.

Borge-Holthoefer, J., Rivero, A., García, I. n., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iñiguez, D., Pérez, M. P., Ruiz, G., Sanz, F., Serrano, F., Viñas, C., Tarancón, A., and Moreno, Y. (2011). Structural and dynamical patterns on online social networks: the Spanish May 15th movement as a case study. *PLoS One*, 6(8):e23883.

Bornholdt, S. and Ebel, H. (2001). World Wide Web scaling exponent from Simon's 1955 model. *Phys. Rev. E*, 64(3):035104.

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

Brockmann, D. (2010). Statistical mechanics: The physics of where to go. *Nat. Phys.*, 6(10):720–721.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–5.

Brown, W. S., Pierce, J. R., and Traub, J. F. (1967). The Future of Scientific Journals: A computer-based system will enable a subscriber to receive a personalized stream of papers. *Science*, 158(3805):1153–1159.

Buchanan, M. (2009). Meltdown modelling. *Nature*, 460(7256):680–2.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10(3):186–98.

Burt, R. (2005). *Brokerage and closure: An introduction to social capital*. Oxford University Press, USA.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, 325(5939):414–6.

Butts, C. T., Acton, R. M., Hipp, J. R., and Nagle, N. N. (2012). Geographical variability and network structure. *Soc. Networks*, 34(1):82–100.

Byrne, D. E. (1971). *The attraction paradigm*. Personality and psychopathology. Academic Press.

Caldarelli, G., Capocci, A., De Los Rios, P., and Muñoz, M. A. (2002). Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Phys. Rev. Lett.*, 89(25):258702.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proc. 23rd Int. Conf. Mach. Learn. - ICML '06*, pages 161–168.

Cattuto, C., Benz, D., Hotho, A., and Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. In Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., and Thirunarayan, K., editors, *ISWC '08 Proc. 7th Int. Conf. Semant. Web*, volume 5318 of *Lecture Notes in Computer Science*, pages 615 – 631, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cattuto, C., Loreto, V., and Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci.*, 104(5):1461–4.

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996):1194–1197.

Centola, D., Eguíluz, V. M., and Macy, M. W. (2007). Cascade dynamics of complex propagation. *Phys. A Stat. Mech. its Appl.*, 374(1):449–456.

Centola, D. and Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *Am. J. Sociol.*, 113(3):702–734.

Chen, B. L., Hall, D. H., and Chklovskii, D. B. (2006). Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci.*, 103(12):4723–8.

Cho, A. (2009). Ourselves and our interactions: the ultimate physics problem? *Science*, 325(5939):406–8.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11*, number ii in KDD '11, page 1082, New York, New York, USA. ACM.

Christakis, N. A. and Fowler, J. H. (2009). *Connected: The Amazing Power of Social Networks and How They Shape Our Lives*. Hachette Digital, Inc.

Chun, H., Kwak, H., Eom, Y. H., Ahn, Y. Y., Moon, S., and Jeong, H. (2008). Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proc. 8th ACM SIGCOMM Conf. Internet Meas. Conf. - IMC '08*, pages 57–70, Vouliagmeni, Greece. ACM.

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proc. Biol. Sci.*, 280(1755):20122863.

Coleman, J. (1988). Social capital in the creation of human capital. *Am. J. Sociol.*, 94(1988):S95.

Collins, N. L. and Miller, L. C. (1994). Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin*, 166(3):457–475.

Conover, M. D., Gonçalves, B., Flammini, A., and Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Sci.*, 1(1):6.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J. P., Sanchez, a., Nowak, a., Flache, a., San Miguel, M., and Helbing, D. (2012). Manifesto of computational social science. *Eur. Phys. J. Spec. Top.*, 214(1):325–346.

Cox, A., Clough, P., and Siersdorfer, S. (2011). Developing metrics to characterize Flickr groups. *J. Am. Soc. Inf. Sci. Technol.*, 62:493–506.

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci.*, 107(52):22436–41.

Csermely, P. (2006). *Weak links: Stabilizers of complex systems from proteins to social networks*. Springer.

Cummings, J. N., Butler, B., and Kraut, R. (2002). The quality of online social relationships. *Commun. ACM*, 45(7):103–108.

Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *J. Stat. Mech. Theory Exp.*, 2005(09):P09008–P09008.

De Masi, G., Iori, G., and Caldarelli, G. (2006). Fitness model for the Italian interbank money market. *Phys. Rev. E*, 74(6):066112.

Dorogovtsev, S., Goltsev, A., and Mendes, J. (2002). Pseudofractal scale-free web. *Phys. Rev. E*, 65(6):066122.

Dorogovtsev, S. N. and Mendes, J. F. F. (2003). *Evolution of Networks: From Biological Nets to the Internet and {WWW}*. Oxford University Press.

Dorogovtsev, S. N., Mendes, J. F. F., and Samukhin, A. N. (2000). Structure of Growing Networks with Preferential Linking. *Phys. Rev. Lett.*, 85(21):4633–4636.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *J. Hum. Evol.*, 22(6):469–493.

Dunbar, R. I. (1998). The social brain hypothesis. *Evol. Anthropol. Issues, News, Rev.*, 6(5):178–190.

Dunne, J. a., Williams, R. J., and Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecol. Lett.*, 5(4):558–567.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci.*, 106(36):15274–8.

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-Free Brain Functional Networks. *Phys. Rev. Lett.*, 94(1):018102.

Emerson, R. (1976). Social Exchange Theory. *Annu. Rev. Sociol.*, 2(1976):335–362.

Epstein, J. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.

Erdos, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61.

Erramilli, V., Yang, X., and Rodriguez, P. (2011). Explore what-if scenarios with SONG: Social Network Write Generator. *arXiv Prepr. arXiv1102.0699*.

Ferrara, E. (2012). A large-scale community structure analysis in Facebook. *EPJ Data Sci.*, 1(1):9.

Forsyth, D. (2009). *Group Dynamics*. Wadsworth/Cengage.

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci.*, 104(1):36–41.

Fortunato, S., Flammini, A., and Menczer, F. (2006). Scale-Free Network Growth by Ranking. *Phys. Rev. Lett.*, 96(21):218701.

Foster, D., Foster, J., Grassberger, P., and Paczuski, M. (2011). Clustering drives assortativity and community structure in ensembles of networks. *Phys. Rev. E*, 84(6):066117.

Freeman, L. (2004). *The development of social network analysis*. Empirical Press, Vancouver.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc. Networks*, 1(3):215–239.

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *Q. J. Econ.*, 114(3):739–767.

Gallos, L. K., Rybski, D., Liljeros, F., Havlin, S., and Makse, H. a. (2012). How People Interact in Evolving Online Affiliation Networks. *Phys. Rev. X*, 2(3):031014.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., and Kellerer, W. (2010). Outtweeting the twitterers-predicting information cascades in microblogs. In *Proc. 3rd Conf. Online Soc. networks*. USENIX Association.

Gantz, J. and Reinsel, D. (2012). The Digital Universe in 2020 : Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Technical report, IDC.

Garlaschelli, D. and Loffredo, M. (2004). Fitness-Dependent Topological Properties of the World Trade Web. *Phys. Rev. Lett.*, 93(18):188701.

Giannotti, F., Pappalardo, L., Pedreschi, D., and Wang, D. (2012). A complexity science perspective on human mobility. In *http://www.dashunwang.com/pdf/2012-mobilityBook.pdf*.

Gibrat, R. (1931). *Les Inégalités Économiques*. Librairie du Recueil Sirey, Paris.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. *Proc. 27th Int. Conf. Hum. factors Comput. Syst. - CHI 09*, page 211.

Giles, J. (2012). Computational social science: Making the links. *Nature*, pages 8–10.

Gloor, P. A. and Zhao, Y. (2006). Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis. In *Proceedings of the conference on Information Visualization*, IV '06, pages 130–135, Washington, DC, USA. IEEE Computer Society.

Gómez, S., Díaz-Guilera, A., Gómez-Gardeñes, J., Pérez-Vicente, C. J., Moreno, Y., and Arenas, A. (2013). Diffusion Dynamics on Multiplex Networks. *Phys. Rev. Lett.*, 110(2):028701.

Gómez-Gardeñes, J., Reinares, I., Arenas, A., and Floría, L. M. (2012). Evolution of cooperation in multiplex networks. *Sci. Rep.*, 2:620.

Gonçalves, B., Perra, N., and Vespignani, A. (2011). Modeling users' activity on twitter networks: validation of Dunbar's number. *PLoS One*, 6(8):e22656.

González, M., Lind, P., and Herrmann, H. (2006). System of Mobile Agents to Model Social Networks. *Phys. Rev. Lett.*, 96(8):088702.

González, M. C., Hidalgo, C. a., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–82.

Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81(4):046106.

Grabowicz, P. A., Aiello, L. M., Eguiluz, V. M., and Jaimes, A. (2013a). Distinguishing topical and social groups based on common identity and bond theory. In *Proc. sixth ACM Int. Conf. Web search data Min. - WSDM '13*, page 627, New York, New York, USA. ACM.

Grabowicz, P. A. and Eguíluz, V. M. (2012). Heterogeneity shapes groups growth in social online communities. *Europhys. Lett.*, 97(2):28002.

Grabowicz, P. A., Ramasco, J. J., and Eguiluz, V. M. (2013b). Dynamics in online social networks. In Mukherjee, A., Choudhury, M., Peruani, F., Ganguly, N., and Mitra, B., editors, *Dyn. Complex Networks, Vol. 2*, Modeling and Simulation in Science, Engineering and Technology, chapter Dynamics i. Springer New York, New York, NY.

Grabowicz, P. A., Ramasco, J. J., Goncalves, B., and Eguiluz, V. M. (2013c). Entangling mobility and interactions in social media.

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

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociol. theory*, 1(1):201–233.

Granovetter, M. S. (1973). The strength of weak ties. *Am. J. Sociol.*, 78(6):1360–1380.

Gruzd, A., Wellman, B., and Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. *Am. Behav. Sci.*, 55(10):1294–1318.

Guimerà, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(2):025101.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explor. Newsl.*, 11(1):10.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., and González, M. C. (2012). Spatiotemporal Patterns of Urban Human Mobility. *J. Stat. Phys.*, 151(1-2):304–318.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Holme, P. and Kim, B. (2002). Growing scale-free networks with tunable clustering. *Phys. Rev. E*, 65(2):026107.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Phys. Rep.*, 519(3):97–125.

Honeycutt, C. and Herring, S. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In Fielding, N., Lee, R. M., and Blank, G., editors, *42st Hawaii Int. Int. Conf. Syst. Sci.*, pages 1–10, Waikoloa, Big Island, HI, USA. IEEE.

Huberman, B. A. and Adamic, L. A. (1999). Internet: Growth dynamics of the World-Wide Web. *Nature*, 401(6749):131.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci.*, 101(42):15124–9.

Iribarren, J. L. and Moro, E. (2011). Affinity Paths and information diffusion in social networks. *Soc. Networks*, 33(2):134–142.

Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2010). What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.*, 271:166–180.

Jackson, M. O. (2010). *Social and Economic Networks*. Number March. Princeton University Press.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. 9th WebKDD 1st SNA-KDD 2007 Work. Web Min. Soc. Netw. Anal. - WebKDD/SNA-KDD '07*, pages 56–65, New York, New York, USA. ACM.

Jia, T., Jiang, B., Carling, K., Bolin, M., and Ban, Y. (2012). An empirical study on human mobility and its agent-based modeling. *J. Stat. Mech. Theory Exp.*, 2012(11):P11024.

Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., and Fowler, J. H. (2013). Inferring tie strength from online directed behavior. *PLoS One*, 8(1):e52168.

Kairam, S. R., Wang, D. J., and Leskovec, J. (2012). The life and death of online groups. In *Proc. fifth ACM Int. Conf. Web search data Min. - WSDM '12*, page 673, New York, New York, USA. ACM.

Katz, N., Lazer, D., Arrow, H., and Contractor, N. (2004). Network Theory and Small Groups. *Small Gr. Res.*, 35(3):307–332.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '03*, page 137, New York, New York, USA. ACM.

Klemm, K. and Eguíluz, V. (2002a). Growing scale-free networks with small-world behavior. *Phys. Rev. E*, 65(5):057102.

Klemm, K. and Eguíluz, V. (2002b). Highly clustered scale-free networks. *Phys. Rev. E*, 65(3):036123.

Klemm, K., Serrano, M. A., Eguíluz, V. M., and Miguel, M. S. (2012). A measure of individual role in collective dynamics. *Sci. Rep.*, 2:292.

Kong, J. S., Sarshar, N., and Roychowdhury, V. P. (2008). Experience versus talent shapes the structure of the Web. *Proc. Natl. Acad. Sci.*, 105(37):13724–9.

Kooti, F., Yang, H., Cha, M., Gummadi, K., and Mason, W. (2012). The emergence of conventions in online social networks. In *Proc. ICWSM*, pages 194–201.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In Nohria, N. and Eccles, R., editors, *Organ. networks Struct. form action*, pages 216–239. Harvard Business School Press, Boston.

Krackhardt, D. and Handcock, M. (2007). Heider vs Simmel: Emergent Features in Dynamic Structures. In Airoldi, E., Blei, D., Fienberg, S., Goldenberg, A., Xing, E., and Zheng, A., editors, *Stat. Netw. Anal. Model. Issues, New Dir.*, volume 4503 of *Lecture Notes in Computer Science*, pages 14–27. Springer Berlin / Heidelberg.

Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. (2009). Urban Gravity: a Model for Intercity Telecommunication Flows. *J. Stat. Mech. Theory Exp.*, 2009(07):L07003.

Krumme, C., Llorente, A., Cebrian, M., Pentland, A. S., and Moro, E. (2013). The predictability of consumer visitation patterns. *Sci. Rep.*, 3:1645.

Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '06*, page 611, New York, New York, USA. ACM.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proc. 19th Int. Conf. World wide web - WWW '10*, page 591, New York, New York, USA. ACM.

Lambiotte, R., Blondel, V., Dekerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Vandooren, P. (2008). Geographical dispersal of mobile communication networks. *Phys. A Stat. Mech. its Appl.*, 387(21):5317–5325.

Lancichinetti, A. and Fortunato, S. (2009a). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1):016118.

Lancichinetti, A. and Fortunato, S. (2009b). Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80(5):056117.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, 11(3):033015.

Lancichinetti, A., Radicchi, F., and Ramasco, J. J. (2010). Statistical significance of communities in networks. *Phys. Rev. E*, 81(4):046110.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS One*, 6(4):e18961.

Lazer, D. (2011). Networks in Political Science: Back to the Future. *PS Polit. Sci. Polit.*, 44(01):61–68.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–3.

Lehmann, J., Gonçalves, B., Ramasco, J. J., and Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proc. 21st Int. Conf. World Wide Web - WWW '12*, page 251, New York, New York, USA. ACM.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '09*, pages 497–506, Paris, France. ACM.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceeding 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '08*, pages 462–470, Las Vegas, Nevada, USA. ACM.

Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceeding 17th Int. Conf. World Wide Web - WWW '08*, page 915, New York, New York, USA. ACM.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proc. 19th Int. Conf. World wide web*, pages 641–650, New York, New York, USA. ACM.

Leung, I., Hui, P., Liò, P., and Crowcroft, J. (2009). Towards real-time community detection in large networks. *Phys. Rev. E*, 79(6):066107.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social networks. *Proc. Natl. Acad. Sci.*, 102(33):11623–8.

Lu, X., Bengtsson, L., and Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Natl. Acad. Sci.*, 109(29):11576.

Ludford, P. J., Cosley, D., Frankowski, D., and Terveen, L. (2004). Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 6, pages 631–638. ACM.

Macy, M. W. and Willer, R. (2002). From Factors to Actors : Computational Sociology and Agent-Based Modeling. *Annu. Rev. Sociol.*, 28(1):143–166.

Marsden, P. V. and Campbell, K. E. (1984). Measuring Tie Strength. *Soc. Forces*, 63(2):482.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569):910–3.

McDaid, A. and Hurley, N. (2010). Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion. In *2010 Int. Conf. Adv. Soc. Networks Anal. Min.*, pages 112–119. IEEE.

McMillan, D. W. and Chavis, D. M. (1986). Sense of community: A definition and theory. *J. Community Psychol.*, 14(1):6–23.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.*, 27(1):415–444.

Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In *Soc. Media Anal. KDD '10 Work.*, Washington, DC, USA. ACM.

Milgram, S. (1967). The Small World Problem. *Psychol. Today*, 1(60).

Miller, G. (2011). Social Scientists wade into the Tweet stream. *Science*, 333.

Miritello, G. (2013). *Temporal Patterns of Communication in Social Networks*. Springer Theses. Springer International Publishing, Heidelberg.

Miritello, G., Lara, R., Cebrian, M., and Moro, E. (2013). Limited communication capacity unveils strategies for human interaction. *Sci. Rep.*, 3:1950.

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2008). Growth of the flickr social network. In *Proc. first Work. Online Soc. networks - WOSP '08*, pages 25–30, Seattle, WA, USA. ACM.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proc. 7th ACM SIGCOMM Conf. Internet Meas. - IMC '07*, pages 29–42, San Diego, California, USA. ACM.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One*, 8(5):e64417.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms*, 6(2-3):161–180.

Moody, J. and White, D. R. (2003). Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. *Am. Sociol. Rev.*, 68(1):103.

Moreno, J. (1934). *Who shall survive?* Nervous and Mental Disease Publishing Company, Washington, DC, USA.

Mörters, P., Peres, Y., Schramm, O., and Werner, W. (2010). *Brownian Motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–8.

Negoescu, R.-A., Adams, B., Phung, D., Venkatesh, S., and Gatica-Perez, D. (2009). Flickr hypergroups. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 813–816, New York, NY, USA. ACM.

Negoescu, R. A. and Perez, D. G. (2008). Analyzing Flickr groups. In *Proc. 2008 Int. Conf. Content-based image video Retr. - CIVR '08*, pages 417–426, Niagara Falls, Canada. ACM.

Newman, M. (2002). Assortative Mixing in Networks. *Phys. Rev. Lett.*, 89(20):208701.

Newman, M. (2003a). Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126.

Newman, M. (2003b). The structure and function of complex networks. *SIAM Rev.*, 45:167–256.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, 46(5):323–351.

Newman, M. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103(23):8577–82.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113.

Newman, M. and Park, J. (2003). Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):9.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007a). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci.*, 104(18):7332–6.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007b). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci.*, 104(18):7332–6.

Padgett, J. and Ansell, C. (1993). Robust Action and the Rise of the Medici, 1400-1434. *Am. J. Sociol.*, 98(6):1259–1319.

Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

Pareto, V. (1896). *Cours d'Économie Politique*. F. Rouge, Lausanne.

Park, J. and Newman, M. (2004). Statistical mechanics of networks. *Phys. Rev. E*, 70(6):066117.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.*, 86(14):3200–3203.

Phithakkitnukoon, S., Smoreda, Z., and Olivier, P. (2012). Socio-geography of human mobility: a study using longitudinal mobile phone data. *PLoS One*, 7(6):e39253.

Prentice, D. A., Miller, D. T., and Lightdale, J. R. (1994). Asymmetries in Attachments to Groups and to their Members: Distinguishing between Common-Identity and Common-Bond Groups. *Personal. Soc. Psychol. Bull.*, 20(5):484–493.

Prieur, C., Pissard, N., Beuscart, J., and Cardon, D. (2008). Thematic and social indicators for Flickr groups. In *Proc. ICWSM*.

Pujol, J., Siganos, G., Erramilli, V., and Rodriguez, P. (2009). Scaling online social networks without pains. In *NETDB*.

Pujol, J. M., Erramilli, V., Siganos, G., Yang, X., Laoutaris, N., Chhabra, P., and Rodriguez, P. (2010). The little engine(s) that could. In *Proc. ACM SIGCOMM 2010 Conf. SIGCOMM - SIGCOMM '10*, page 375, New York, New York, USA. ACM.

Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. (2012a). Tracking "gross community happiness" from tweets. In *Proc. ACM 2012 Conf. Comput. Support. Coop. Work - CSCW '12*, page 965, New York, New York, USA. ACM.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *2011 IEEE Third Int'l Conf. Privacy, Secur. Risk Trust 2011 IEEE Third Int'l Conf. Soc. Comput.*, pages 180–185.

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J. (2012b). The personality of popular facebook users. In *Proc. ACM 2012 Conf. Comput. Support. Coop. Work - CSCW '12*, page 955, New York, New York, USA. ACM.

Raghavan, U., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106.

Rapoport, A. (1957). Contribution to the theory of random and biased nets. *Bull. Math. Biophys.*, 19(4):257–277.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *Proc. 20th Int. Conf. companion World wide web - WWW '11*, page 249, New York, New York, USA. ACM.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. (2010). Characterizing and Modeling the Dynamics of Online Popularity. *Phys. Rev. Lett.*, 105(15):158701.

Ren, Y., Kraut, R., and Kiesler, S. (2007). Applying Common Identity and Bond Theory to Design of Online Communities. *Organ. Stud.*, 28(3):377–408.

Riger, S. and Lavrakas, P. J. (1981). Community ties: Patterns of attachment and social interaction in urban neighborhoods. *Am. J. Community Psychol.*, 9(1):55–66.

Rivera, M. T., Soderstrom, S. B., and Uzzi, B. (2010). Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annu. Rev. Sociol.*, 36(1):91–115.

Rivera-Alba, M., Vitaladevuni, S. N., Mishchenko, Y., Mischenko, Y., Lu, Z., Takemura, S.-Y., Scheffer, L., Meinertzhagen, I. a., Chklovskii, D. B., and de Polavieja, G. G. (2011). Wiring economy and volume exclusion determine neuronal placement in the Drosophila brain. *Curr. Biol.*, 21(23):2000–5.

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., and Almeida, V. (2011). On word-of-mouth based discovery of the web. In *Proc. 2011 ACM SIGCOMM Conf. Internet Meas. Conf. - IMC '11*, page 381, New York, New York, USA. ACM.

Romero, D. M. and Kleinberg, J. (2010). The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In *ICWSM*.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics. In *Proc. 20th Int. Conf. World wide web - WWW '11*, page 695, New York, New York, USA. ACM.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.*, 105(4):1118–23.

Rosvall, M. and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One*, 6(4):e18209.

Rozenfeld, H. D., Rybski, D., Andrade, J. S., Batty, M., Stanley, H. E., and Makse, H. a. (2008). Laws of population growth. *Proc. Natl. Acad. Sci.*, 105(48):18702–7.

Rybski, D., Buldyrev, S. V., Havlin, S., Liljeros, F., and Makse, H. a. (2009). Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci.*, 106(31):12640–5.

Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proc. fifth ACM Int. Conf. Web search data Min. - WSDM '12*, page 723, New York, New York, USA. ACM.

Saichev, A., Malevergne, Y., and Sornette, D. (2009). *Theory of Zipf's Law and Beyond*. Springer, New York.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–6.

Samuelson, P. (1954). The Pure Theory of Public Expenditure. *Rev. Econ. Stat.*, 36(4):387–389.

Sassenberg, K. (2002). Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics Theory Research And Practice*, 6(1):27–37.

Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. In *Proc. ICWSM*, pages 329–336.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. (2010). Folks in Folksonomies. In *Proc. third ACM Int. Conf. Web search data Min. - WSDM '10*, page 271, New York, New York, USA. ACM.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., and White, D. R. (2009). Economic networks: the new challenges. *Science*, 325(5939):422–5.

Serrano, M. A. and Boguñá, M. (2005). Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E*, 72(3):036133.

Simini, F., González, M. C., Maritan, A., and Barabási, A.-l. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

Simmel, G. (1950). *The Sociology of Georg Simmel*. Free Press of Glencoe.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42:425–440.

Söderberg, B. (2002). General formalism for inhomogeneous random graphs. *Phys. Rev. E*, 66(6):066121.

Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nat. Phys.*, 6(10):818–823.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science*, 327(5968):1018–21.

State, B., Weber, I., and Zagheni, E. (2013). Studying inter-national mobility through IP geolocation. In *Proc. sixth ACM Int. Conf. Web search data Min. - WSDM '13*, page 265, New York, New York, USA. ACM.

Szell, M., Lambiotte, R., and Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci.*, 107(31).

Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. (2012). Understandiang mobility in a social petri dish. *Sci. Rep.*, 2:457.

Tajfel, H. (1982). *Social identity and intergroup relations*, volume 7. Cambridge University Press.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter networks. *Soc. Networks*, 34(1):73–81.

Tang, L., Wang, X., and Liu, H. (2011). Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol.*, 3(1):15:1–15:25.

Taraborelli, D. (2011). Viable web communities: two case studies. In Deffuant, G. and Gilbert, N., editors, *Viability Resil. Complex Syst.*, pages 75–105. Kluwer.

Tessone, C. J., Geipel, M. M., and Schweitzer, F. (2011). Sustainable growth in complex networks. *Europhys. Lett.*, 96(5):58005.

Toivonen, R., Onnela, J., Saramaki, J., Hyvonen, J., and Kaski, K. (2006). A model for social networks. *Phys. A Stat. Theor. Phys.*, 371(2):851–860.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 4(4):425–443.

Ugander, J., Backstrom, L., Marlow, C., and Kleinberg, J. (2012). Structural diversity in social contagion. *Proc. Natl. Acad. Sci.*, 109(16):5962–6.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The Anatomy of the Facebook Social Graph. *Arxiv Prepr. arXiv1111.4503*.

Utz, S. and Sassenberg, K. (2002). Distributive justice in common-bond and common-identity groups. *Group Processes and Intergroup Relations*, 5(2):151–162.

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *Am. Sociol. Rev.*, 61(4):674–698.

Van Dijk, J. (2006). *The network society: social aspects of new media*. London: Sage Publications Ltd.

Vedres, B. and Stark, D. (2010). Structural Folds: Generative Disruption in Overlapping Groups1. *Am. J. Sociol.*, 115(4):1150–1190.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425–8.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the Evolution of User Interaction in Facebook. In *Proc. 2nd ACM Work. Online Soc. networks - WOSN '09*, page 37, Barcelona, Spain. ACM.

Volkovich, Y., Scellato, S., Laniado, D., Mascolo, C., and Kaltenbrunner, A. (2012). The length of bridge ties: structural and geographic properties of online social interactions. In *Proc. ICWSM*, volume 12, pages 346–353.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. In *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11*, page 1100, New York, New York, USA. ACM.

Wang, D., Song, C., and a. L. Barabasi (2013). Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132.

Wang, P., González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–6.

Watts, D. J. (2007). A twenty-first century science. *Nature*, 445(7127):489.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2.

Wegner, D. M. (1987). Transactive Memory: A Contemporary Analysis of the Group Mind. In Mullen, B. and Goethals, G. R., editors, *Theor. Gr. Behav.*, pages 185–208. Springer New York, New York, NY.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., and Haythornthwaite, C. (1996). Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. *Annu. Rev. Sociol.*, 22(1):213–238.

Welser, H. T., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8(2).

Weng, L., Menczer, F., and Ahn, Y.-y. (2013). Virality prediction and community structure in social networks. *Sci. Rep.*, 3:2522.

White, D. R. and Harary, F. (2001). The Cohesiveness of Blocks In Social Networks: Node Connectivity and Conditional Density. *Sociol. Methodol.*, 31(1):305–359.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proc. fourth ACM Eur. Conf. Comput. Syst. - EuroSys '09*, pages 205–218, Nuremberg, Germany. ACM.

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proc. 20th Int. Conf. World wide web - WWW '11*, page 705, New York, New York, USA. ACM.

Yang, J. and Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *Proc. ACM SIGKDD Work. Min. Data Semant. - MDS '12*, pages 1–8, New York, New York, USA. ACM.

Zipf, G. (1946). The P 1 P 2/D hypothesis: on the intercity movement of persons. *Am. Sociol. Rev.*, 11(6):677–686.

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Reading MA.