School of Business and Management
Queen Mary College
University of London

Thesis submitted for the Degree of PhD

Structure and Evolution of Weighted Networks

Author:
Tore Opsahl

Supervisors:
Dr. Pietro Panzarasa
Prof. Christian Beck

May 15, 2009

Declaration

I certify that the thesis I have presented for examination for the Degree of PhD of University of London (Queen Mary College) is solely my own work other than where I have clearly indicated that it is joint work with others.

Tore Opsahl
May 15, 2009

Abstract

In my thesis, I will outline and critically discuss some of the research projects on which I have been working during my Ph.D. programme. All my projects draw on, and extend, recent theoretical and methodological advances in network science.

Many social network measures can only be applied to networks in which ties are either present or absent. One of these measures is the clustering coefficient. The first project represents an attempt to overcome this shortcoming by proposing a generalisation of the clustering coefficient that is explicitly based on the weight of ties, and is therefore suitable to the analysis of weighted networks.

A second project investigates the nature of the interactions among prominent nodes in a network. To this end, a new general measure is proposed aimed at evaluating whether, and the extent to which, the strongest ties in the network occur among these nodes.

A third project explores the network growth mechanisms that underpin the evolution of social interaction over time. We utilised a regression framework to assess these mechanisms in an online social network. In particular, we investigate the effects of triadic closure, preferential attachment, reciprocity, homophily, focus constraints, and reinforcement on tie generation.

The aim of these projects is to contribute to a better understanding of the principles that govern the global organisation and functioning of networks.

In addition, a fourth project is devoted to the development of an open-source software programme that can deal with weighted and longitudinal networks, and incorporates the methods proposed in the other chapters. This project has the potential to provide researchers with a common platform on which new methodological advances easily can be made.

Acknowledgements

The theme of this thesis is interdependence among elements. In fact, this thesis is not just a product of myself, but also of my interdependence with others. Without the support of a number of people, it would not have been possible to write. It is my pleasure to have the opportunity to express my gratitude to many of them here.

For my academic achievements, I would like to acknowledge the constant support from my supervisors. In particular, I thank Pietro Panzarasa for taking an active part of all the projects I have worked on. I have also had the pleasure to collaborate with people other than my supervisors. I worked with Vittoria Colizza and José J. Ramasco on the analysis and method presented in Chapter 3, Kathleen M. Carley on an empirical analysis of the online social network used throughout this thesis, and Martha J. Prevezer on a project related to knowledge transfer in emerging countries. In addition to these direct collaborations, I would also like to thank Filip Agneessens, Sinan Aral, Steve Borgatti, Ronald Burt, Mauro Faccioni Filho, Thomas Friemel, John Skvoretz, and Vanina Torlò for encouragement and helpful advice. In particular, I would like to thank Tom A. B. Snijders and Klaus Nielsen for insightful reading of this thesis and many productive remarks and suggestions.

I have also received feedback on my work at a number of conferences and workshops. I would like to express my gratitude to the participants at (in alphabetical order): Academy of Management's 2007 Annual Meeting (Philadelphia, PA, USA), American Sociological Association's 102nd Annual Meeting (New York City, NY, USA), Applications of Social Network Analysis 2006 (Institute of Mass Communication and Media Research, University of Zurich, Zurich, Switzerland), Cass Business School's Workshop on Scientific and Managerial Knowledge 2008 (City University, London, UK), Danish Research Unit for Industrial Dynamics (DRUID) Conference 2008 (Copenhagen Business School, Copenhagen, Denmark), European Conference

on Complex Systems (Saïd Business School, Oxford University, Oxford, UK), International Sunbelt Social Network Conference 26 (International Network for Social Network Analysis, Vancouver, Canada), International Sunbelt Social Network Conference 27 (International Network for Social Network Analysis, Corfu, Greece), International Sunbelt Social Network Conference 28 (International Network for Social Network Analysis, Tampa, FL, USA), International Sunbelt Social Network Conference 29 (International Network for Social Network Analysis, San Diego, CA, USA), International Workshop and Conference on Network Science 2008 (NetSci'08, Norwich BioScience Institutes, Norwich, UK), Nuffield/OII Networks Seminar Series (Oxford University, Oxford, UK), Social Network Analysis Forum 2005 (Centre for Criminology, Oxford University, Oxford, UK), Social Network Analysis Forum 2006 (University of Leeds, Leeds, UK), The ISI Foundation (Turin, Italy), UK Social Network Conference 2007 (Queen Mary College, University of London, London, UK), and UK Social Network Conference 2008 (University of Greenwich, London, UK)

On a social note, I would like to thank John, Claudius, and my family for their continuing support. Without them I would have lost focus. My peers and the administrative staff have also been a great source of support. In particular, I would like to extend my acknowledgements to Mariusz Jarmuzek, Geraldine Marks, Roland Miller, Jenny Murphy, Cathrine Seierstad, Lorna Soar, Steven Telford, and Eshref Trushin.

Contents

1 Introduction **9**
1.1 Weighted networks . . . 14
1.2 Longitudinal networks . . . 16
1.3 Projects and outline of thesis . . . 18
1.4 Network datasets . . . 21

2 Clustering in Weighted Networks **25**
2.1 Clustering coefficient . . . 28
2.2 Generalised clustering coefficient . . . 29
2.3 Empirical tests . . . 34
2.4 Directed networks . . . 39
2.5 Contribution to the literature . . . 43
2.6 Conclusion and discussion . . . 44

3 Prominence and Control: The Weighted Rich-club Effect **49**
3.1 The topological rich-club effect . . . 50
3.2 The weighted rich-club effect . . . 52
3.2.1 Null models . . . 53
3.2.2 Significance of effect . . . 58
3.3 Empirical tests . . . 59
3.3.1 Club of the most connected nodes . . . 60
3.3.2 Club of the most active nodes . . . 65
3.3.3 Club of the nodes with the highest average weight . . . 67
3.4 Contribution to the literature . . . 68
3.5 Conclusion and discussion . . . 70

4 Evolution of Networks **72**
4.1 Network growth mechanisms . . . 74
4.2 Cross-sectional binary networks . . . 76
4.3 Longitudinal binary networks . . . 82
4.4 Longitudinal weighted networks . . . 91
4.5 Sensitivity to the number of control cases . . . 96
4.6 Contribution to the literature . . . 98
4.7 Conclusion and discussion . . . 99

5 *tnet*: Software for Analysis of Weighted and Longitudinal networks **104**
5.1 Data structures . . . 106
5.2 Weighted network functions . . . 112
5.3 Longitudinal network functions . . . 116
5.4 Contribution to the literature . . . 120
5.5 Conclusion and discussion . . . 121

6 Concluding Remarks **123**

Bibliography **128**

Appendix **146**

A Presented and Published Papers **146**

B Appendix to Prominence and Control: The Weighted Rich-club Effect **149**
B.1 Directed Weight reshuffle when prominence is defined in terms of degree 149
B.2 Weighted rich-club effect in the *Network Science* collaboration network 151

C Source code of *tnet* 0.1.0 **152**
C.1 add_window_to_longitudinal_data . . . 153
C.2 as.longitudinal . . . 154
C.3 betweenness_w . . . 156
C.4 closeness_w . . . 158
C.5 clustering_w . . . 160
C.6 degree_w . . . 162
C.7 dichotomise . . . 164
C.8 distance_w . . . 165
C.9 longitudinal_data_to_edgelist . . . 167
C.10 rg_longitudinal . . . 168
C.11 rg_reshuffling_w . . . 171
C.12 rg_w . . . 174
C.13 shrink_to_weighted_network . . . 175
C.14 symmetrise . . . 176
C.15 tnet.growth.clogit . . . 178
C.16 weighted_richclub . . . 190

List of Figures

1 Example of a network with weighted ties . . . 16
2 Two weighted sample networks. . . . 32
3 Non-vacuous triplets centred around node i . . . 39
4 Weighted rich-club ordering . . . 54
5 Randomisation procedures . . . 57
6 Distribution of $\phi^w_{\text{null}}(r)$ . . . 59
7 Weighted rich-club ordering among the most connected nodes . . . 61
8 The prominent nodes in the network science collaboration network . . . 64
9 Weighted rich-club ordering among the most actively involved nodes . . . 66
10 Weighted rich-club ordering among the nodes with the highest average weight . . . 69
11 Example of dependence among ties . . . 77
12 Example of triplets . . . 85
13 In-degree distribution . . . 86

14 Example of weighted triplets 94
15 Significance with an increasing number of control nodes 97
16 Example of a directed and an undirected network with weighted ties . 107
17 Example of a longitudinal network 109
18 Impact of smoothing window on network measures 112
19 Example of distance in a weighted network 115
20 Example of network mechanisms 117
21 Weighted rich-club ordering among the most connected nodes 150
22 Weighted rich-club ordering among the most connected nodes in the *Network Science* collaboration network 151

List of Tables

1 Methods of calculating the triplet value, ω. 30
2 Simulations of the generalised clustering coefficient 31
3 Comparison between the generalised and the binary clustering coefficients. 37
4 Simulations of the generalised clustering coefficient (directed networks) 41
5 Triplets, τ, and triplet values, ω, in a directed network 42
6 Growth mechanisms in a binary network 88
7 Growth mechanisms in a weighted network 93
8 Format for directed weighted edgelists 107
9 Format for undirected weighted edgelists 107
10 Format for longitudinal data 110
11 The `as.longitudinal`-function 111

1 Introduction

As we move toward a highly competitive global economy where no person or firm acts in isolation (Drucker, 1993; Harris, 2001), it is vital to understand the systems in which people and firms interact. These systems can be represented as networks where the entities are called nodes and interactions among them are represented in terms of ties. More generally, a node can be a neuron, an individual, a group, an organisation, or even a country, whereas ties can take the form of friendship, communication, collaboration, alliance, or trade, to name only a few (Wasserman and Faust, 1994). For example, within an organisation, the network of social interactions among employees has been referred to as the organisation's informal structure (Hinds and Kiesler, 1995). This informal structure can reflect particular aspects of social interaction. For instance, a possible informal structure is the advice network within an organisation. The ties in this network are formed when an employee asks another for advice (Lazega, 2001). Another intra-organisational network is the one that maps the collaboration structure among employees (Cross and Parker, 2004). In this case, ties would be formed between employees collaborating on projects or other work-related tasks.

Many research problems within a wide range of disciplines can be framed within a network perspective (Wasserman and Faust, 1994; Watts, 2004). This has led to the development of social network analysis in sociology (Freeman, 2004), the analysis of complex networks in physics (Newman, 2003), and graph theory in mathematics and computer science (Bollobás, 1998; Leskovec et al., 2005). This research has been both theoretical (Erdős and Rényi, 1960; Granovetter, 1973; Heider, 1946; Luce and Perry, 1949) and empirical (Bernard et al., 1988; Foster et al., 1963; Milgram, 1967; Moreno, 1938; Watts and Strogatz, 1998). Stanley Milgram (1967) and his colleagues (Korte and Milgram, 1970; Travers and Milgram, 1969) conducted one of

the first empirical studies on the structure of social networks and the average shortest distance between nodes (i.e., the lowest number of ties that directly ($distance = 1$) or indirectly ($distance \geq 2$) separate nodes). He randomly chose people from a phone directory in Nebraska, a state located in the Midwestern United States. Each person was asked to send a letter to someone whom they knew on a first-name basis with the goal of ultimately reaching a stockbroker in Boston. The best strategy for the people was to send the letter to someone whom they perceived to be either socially or geographically closer to the stockbroker in some sense. A number of Milgram's letters did reach the stockbroker, and the average number of people that those letters had passed through was only about six. There were weaknesses and biases in Milgram's research. For example, letters which passed through more people were perhaps more likely to get lost or forgotten, or there might have been shorter paths for the letters to travel than the ones that were recorded. Nonetheless, Milgram's "six-degrees-of-separation" experiment is usually taken as evidence of the "small-world" effect (Watts and Strogatz, 1998). This effect refers to the fact that a short distance separates most people, even when the size of the population is very large and when people have a tendency to form small tightly knit groups. Milgram's research does not address the reasons why certain letters reached the stockbroker while others got lost, and the features of the network that impacted on this. A critical review of this work can be found in Wasserman and Faust (1994) and Watts (1999).

In addition to studying the shortest paths that connect nodes, a substantial body of work has concentrated on other features of networks (e.g., Bernard et al., 1988; Fararo and Sunshine, 1964; Foster et al., 1963). To this end, a number of complete social networks (i.e., networks for which information on a specific set of nodes and all ties among them is available) have been collected with a view to uncovering the global structural patterns that emerge from the ways in which individuals behave

at a local level. Among the early empirical studies are the studies by Foster et al. (1963) and Fararo and Sunshine (1964), who constructed maps of friendship networks among high-school students, and by Bernard et al. (1988) who did the same for communities of Utah Mormons, Native Americans, and Micronesian islanders. These efforts have been variously motivated. For example, some were aimed at bettering our understanding of human interaction patterns (Bernard et al., 1988; Fararo and Sunshine, 1964; Holland and Leinhardt, 1971; Lazarsfeld and Merton, 1954; Wasserman and Pattison, 1996); others focused on the spread of information and infectious diseases (Valente, 1995), and the effect of network on individuals' and organisations' performance (Burt, 1992; Coleman, 1988; Gulati and Gargiulo, 1999). More generally, scholars have been interested in the mechanisms governing tie generation and network evolution (Holland and Leinhardt, 1981; Powell et al., 2005; Snijders, 2001; Wasserman and Pattison, 1996). Early studies have investigated how people may benefit from participating in small dense groups (Coleman, 1988; Heider, 1946), for example by choosing new acquaintances that are already tied to current acquaintances, a process known as triadic closure (Davis, 1970; Heider, 1946; Holland and Leinhardt, 1970). In addition, a longstanding tradition of research has focussed on the effect that sharing demographic characteristics has on the existence of a tie (Lazarsfeld and Merton, 1954), and the term homophily was coined to indicate the tendency of individuals to interact with others socially similar to themselves (for a review, see McPherson et al., 2001). Scholars have also been concerned with the effects of focus constraints on tie generation, and empirically examined the tendency of social relationships to be established preferentially between individuals that share activities, roles, and social positions (Feld, 1981; Monge et al., 1985).

While directly probing the network structure and functioning, many of the early empirical studies of social networks suffered from two fundamental weaknesses.

These weaknesses stem from the data collections methods that were typically used to record social interactions, namely interviews and surveys (for a review, see Marsden, 1990). First, the methods were sensitive to subjective bias on the part of interviewees. In particular, what is considered an "acquaintance" or a "friend" can differ considerably from one person to another. This bias is usually referred to as the informant inaccuracy bias (Bernard et al., 1984). Second, survey instruments and direct observation methods are typically labour-intensive and onerous to administer. Thus, the number of nodes in the collected network is limited. In fact, the early social networks often comprised only a few tens (e.g., Bernard et al., 1988) or hundreds (e.g., Fararo and Sunshine, 1964) of people. An implication of the small network size is that it is difficult to perform robust statistical analysis, which in turn easily could result in biased conclusions.

To overcome these shortcomings, a number of researchers have studied networks for which there exist more precise definitions of connectedness (Watts and Strogatz, 1998), and collected much larger network datasets using archival records (Burt and Lin, 1977). Examples of such networks are the neural network of worms (Watts and Strogatz, 1998), electric power grid (Watts and Strogatz, 1998), the Internet (Albert et al., 1999; Broder et al., 2000), and the pattern of air traffic between airports (Amaral et al., 2000; Barrat et al., 2004). However, these networks suffer from a different problem: although they may loosely be regarded as social networks in the sense that their structure in some way reflects features of the society in which they are embedded, they do not measure actual ties among people. Many researchers are, of course, interested in these networks for their own sake, but to the extent that we want to find out more about individuals' interaction patterns, neural networks of worms, power grids, and computer networks are a poor substitute for the real thing.

In recent years, however, two considerable developments have prompted noteworthy advances in social network analysis. First, to explore patterns of interactions

among individuals in large-scale networks, Watts and Strogatz (1998) investigated the network of movie actors and Uzzi and Spiro (2005) studied the network of the artists working in Broadway musicals. In these networks, which has been meticulously constructed from archival data and contains thousands of people, two people are considered connected if they have been credited with appearance in the same movie or musical. In addition, a number of authors have studied scientists collaborating on patents and scientific publications (Barabási et al., 2002; Hall et al., 2001; Moody, 2004; Newman, 2001d). In these networks, some of which contain millions of scientists, a tie is considered to exist between two scientists if they have worked on a patent or paper together. However, while all these network are indeed made of people, they are not without limitations. A major weakness is the validity of ties. In other words, the appearance of two actors in the same movie or collaboration of two scientists on a publication does not necessarily imply that they are acquainted in any, but the most cursory fashion, or that their social relationship extends beyond the artistic or scientific endeavour. Furthermore, in co-authorship networks, the network structure can be biased by the fact that some scientists are listed as authors on an extremely large number of papers. This might be due to the common practice of some research institutes to list the directors' names on most of the papers published by scientists at these institutes (Newman, 2001d). These scientists have not necessarily interacted with all the other scientists they have co-authored with; however, they are likely to act as hubs by creating shortcuts among different groups of nodes, and thereby reduce the distance among nodes in the overall network.

Second, a remarkable deterioration of boundaries between disciplines has been responsible for new theoretical developments and a variety of new analytical tools for modelling the structure and behaviour of networks. Previously, on the one hand, social and behavioural scientists studied in-depth social relationships, whereas, on the other, physicists and applied mathematicians developed methods for analysing

large-scale networks. However, through interdisciplinary collaborations, these two sides have been brought together. In fact, Chapter 3 of this thesis is the outcome of a collaboration among social scientists and physicists. These two advances have contributed to the birth of what has been called the "new science of networks" (Watts, 2004).

1.1 Weighted networks

A major limitation of many methods used for studying large-scale networks stems from the fact that the strength of ties is not taken into account. Granovetter (1973) argued that the strength of a social tie is a function of its duration, emotional intensity, intimacy, and exchange of services. For non-social networks, the strength often reflects the function performed by the ties, e.g. carbon flow ($mg/m^2/day$) between species in food webs (Luczkowich et al., 2003; Nordlund, 2007) or the number of synapses and gap junctions in a neural networks (Watts and Strogatz, 1998). In infrastructure and information networks, variations in the strength of a tie depend on the flow of information, energy, people, and goods along that tie (Barrat et al., 2004; Guimerà et al., 2005; Pastor-Satorras and Vespignani, 2004). The strength of a tie is generally operationalised into a weight that is attached to the tie, thereby creating a weighted network (Wasserman and Faust, 1994).

Exploring the information that weights hold allows us to further our understanding of networks. In social networks, strong ties are often found among socially embedded individuals (Granovetter, 1973). In fact, Simmel (1950) argued that a strong tie cannot exist without other indirect ties (weak and strong). Strong ties facilitate change in the face of uncertainty (Krackhardt, 1992), reinforce obligations, expectations, and social norms (Coleman, 1988), and promote the transfer of complex and tacit knowledge by sustaining individuals' motivation to assist one another (Hansen, 1999; Reagans and McEvily, 2003). Strong ties also aid commu-

nication through, for example, the development of relationship-specific heuristics (Uzzi, 1997). In the behaviour of non-social networks, such as technological and transportation ones, strong ties also play crucial roles. For instance, in the network of routers on the Internet they form part of the large backbones that provide national or inter-continental connectivity (Pastor-Satorras and Vespignani, 2004), whereas in airport networks they represent major international or trans-oceanic routes (Barrat et al., 2004).

One of Granovetter's (1973) major findings is the idea that novel and explicit information is more likely to flow to individuals through weak ties than through strong ones. An individual's friends tend to move in the same circles, and therefore are likely to receive the same information that the individual already possesses. Conversely, acquaintances are likely to know people that the individual does not, and thus receive more novel information. If this information is explicit or codified, it can easily be transferred from the acquaintance to the individual (Levin and Cross, 2004). In light of these findings, the measures that scholars typically apply to study networks should be sensitive to tie strength and capture the difference between strong and weak ties. This will ensure that the full richness of the data is retained.

However, even though ties in many empirical networks have naturally a strength associated with them, most network measures can only be applied to binary networks, i.e. networks where ties are either present or absent (Scott, 2000; Wasserman and Faust, 1994). Therefore, researchers must convert weighted networks into binary ones. A common way of doing this is to dichotomise a weighted network into a series of binary ones by using a set of cutoffs. A tie is set to present if its weight is higher than the cut-off, and removed otherwise. As Figure 1 shows, a weighted network (a) can create a range of binary networks (b-d) depending on the cut-off. This figure illustrate how the subjectivity of the researcher's choice of the cut-off value will produce biases that may invalidate subsequent analysis.

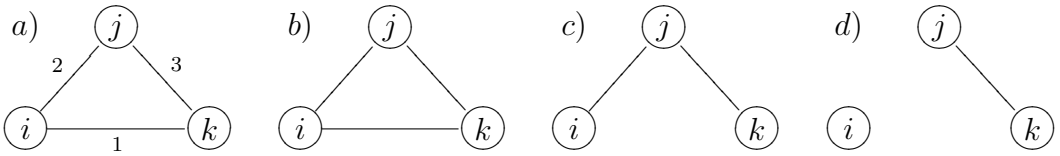


Figure 1: Example of a network with weighted ties ω (a) dichotomised into multiple binary networks based on different cut-offs (b: $\omega > 0$; c: $\omega > 1$; d: $\omega > 2$).

A better way to analyse weighted networks would be to redefine and generalise current methods to explicitly take weights into account. Currently, only a small set of measures have been generalised (Freeman et al., 1991; Newman, 2001c). For example, Newman (2001c) generalised Freeman's (1978) closeness measure by applying a method from computer science to define the distances among nodes (Dijkstra, 1959). The generalised methods would aid the analysis by removing the bias resulting from the subjective choice of the cut-off.

1.2 Longitudinal networks

Another limitation of traditional methods used to study networks is related to the way in which networks are collected. As described above, datasets are often collected through surveys or interviews at a single point in time. Consequently, most network measures have been developed for the purpose of analysing cross-sectional networks. However, networks evolve over time by the addition and removal of nodes, and the forming, strengthening, weakening, and ultimately, the severing of ties. For example, one network that is currently receiving a great deal of attention in the literature is the network of commercial airports (the nodes) that are tied together by scheduled flights (Amaral et al., 2000; Barrat et al., 2004; Guimerà et al., 2005; Opsahl et al., 2008). This network grows when new airports open and shrinks when old ones close down. Ties are created when new routes are started and reinforced if the capacity of an existing route is increased. Weakening of a tie occurs when airlines cut capacity, which ultimately results in the severing of the tie if all airlines terminate flights on

the route. Yet, the evolution of networks is typically not recorded. This implies that the dependency structure among ties is unknown.

Nevertheless, a number of measures aimed at examining the underpinning principles of tie generation have been developed. The first generation of methods aimed to detect a single structural feature in cross-sectional networks. For example, the clustering coefficient (for a review, see Chapter 2) identifies the extent to which triangles occur in a network. The coefficient obtained for an observed network can be compared to the expected value on a corresponding random network (Erdős and Rényi, 1959; Newman, 2003; Panzarasa et al., 2009; Solomonoff and Rapoport, 1951). If it is higher than the expected one, then scholars have often concluded that there is a mechanism that increases the likelihood of forming a tie between two nodes if they have ties to the same other node (triadic closure). However, this approach could be biased as other mechanisms could contribute the generation of triangles. For example, it could be the case that similar people were more likely to form ties with each other than randomly expected (homophily; Lazarsfeld and Merton, 1954; McPherson et al., 2001). If this was the case, a set of similar nodes is likely to be create a tightly knit group that would increase the level of clustering in the network, without any triadic closure effect. In fact, it would be difficult to assess whether, and the extent to which, the tightly knit group was formed due to, for example, triadic closure or homophily mechanisms.

This issue has motivated the development of a second generation of methods (p^* or Exponential Random Graph models). These methods allow for a multivariate analysis of mechanisms that might lead to tie generation (Holland and Leinhardt, 1981; Robins and Morris, 2007; Snijders, 2001; Wasserman and Pattison, 1996). They model the entire network and try to study how, and the extent to which, different mechanisms can be combined to produce the observed network. These mechanisms of tie generation include triadic closure, homophily, preferential attachment

(Barabási and Albert, 1999), and reciprocity (Gouldner, 1960; Plickert et al., 2007). However, these models suffer from criticism as they are difficult to interpret and extend, and have a range of unknown parameters (Hunter et al., 2008). Although there have been attempts to solve these issues (Snijders et al., 2006), a simpler, more flexible, methodologically sound framework is needed to improve the analysis of the underpinning principles of tie generation. This will ultimately improve our knowledge of the organisation and function of networks.

1.3 Projects and outline of thesis

In the last few years, I have focused on a number of research projects aimed at improving our understanding of networks. These include both theoretical and empirical studies as well as the development of a software package. In this thesis, I will highlight four of these projects, two of which are related to a structural analysis of weighted networks. The first project, presented in Chapter 2, is a generalisation of a much-studied network measure, namely the clustering coefficient, to weighted networks. The clustering coefficient is the fraction of triplets (i.e., three nodes that are tied together) that are part of triangles in a network (Luce and Perry, 1949; Wasserman and Faust, 1994). In social networks, the clustering coefficient tends to be higher than the one found if the ties were randomly formed among the people (Erdős and Rényi, 1959; Solomonoff and Rapoport, 1951). This implies that people have an increased likelihood of being tied together if they share a common contact. In sociology, it has been speculated that this is due to a person's cognitive need to create balance by introducing his or her contacts to each other (Heider, 1946; Holland and Leinhardt, 1971).

A limitation of the clustering coefficient is that it can only be applied to binary networks. This represents a major weakness as the richness of the information offered by the strength of ties is undoubtedly lost. In this thesis, I propose a new

measure that takes into account the strength of ties by incorporating the weights directly into its definition. The generalisation is flexible and allows the coefficient to be applied to different types of networks, including directed ones. In directed networks, ties do not simply connect two nodes together, but are formed by one node and terminates at another (Wasserman and Faust, 1994). For example, asking for advice is often seen as a directed tie (or arc) because it refers to social interactions in which knowledge flows from one person to another in a specific direction (Lazega, 2001). Conversely, collaboration is typically seen as two people forging an undirected tie (or edge) because it usually implies a two-way interaction between the nodes (Newman, 2001b).

The second project (Chapter 3) proposes a new measure for weighted networks called the weighted rich-club effect. Unlike the generalisation of the clustering coefficient, this measure detects a feature only found in weighted networks, namely whether the strongest ties in a network are shared among a subset of "prominent" nodes, e.g. nodes with the largest number of contacts. Numerous studies have shown that a number of properties in a wide range of networks are heterogeneously distributed across the nodes (Barabási and Albert, 1999; Barrat et al., 2004; Pareto, 1897; Pastor-Satorras and Vespignani, 2004; Simon, 1955; Zipf, 1935). Investigating the nature of the interactions among the nodes with the highest levels of a given property (i.e., the prominent ones) can provide useful insights into the network's organisation and functioning. Scholars have already started studying interactions among prominent nodes by investigating whether there is a tendency of the highly connected nodes to form more ties with each other than randomly expected (the topological rich-club phenomenon; Colizza et al., 2006; Zhou and Mondragon, 2004). Conversely, our proposed measure assesses whether the prominent nodes share the strongest ties in the network. For example, we ask: do prominent people attract and exchange among themselves the vast majority of resources available in a social

network, or do they tend to distribute resources homogeneously across the network? The measure helps answer this question by testing if the ties among prominent nodes are stronger or weaker than expected by chance.

The third project presented in Chapter 4 aims to provide a framework for studying the evolution of binary and weighted longitudinal networks (i.e., networks in which the exact sequence of the addition and removal of nodes, and the creation, strengthening, weakening, and severing of ties is known). Unlike cross-sectional networks with unknown dependency structure, the dependency among ties is known in this type of datasets. This feature allows us to identify the other nodes in the network and their properties at the time a node decides to form a tie. Thus, whether, and the extent to which, different properties affect the likelihood of receiving a tie can be probed directly. We empirically test this framework on an online social network created from a virtual community. This allows us to accurately study the interplay between a host of mechanisms that guide online behaviour and interpersonal dynamics.

During my PhD, I have used numerous software packages. For network analysis, these include *UCINET* (Borgatti et al., 2002), *Pajek* (Batagelj and Mrvar, 2007), *Siena* (Snijders et al., 2007), *Pnet* (Wang et al., 2005), and the *sna* (Butts, 2006) and *statnet* (Handcock et al., 2003) packages in *R* (R Development Team, 2008). To perform statistical and econometric analysis I have relied upon *Stata* (StataCorp, 2007) and general functions in *R*. In addition, I have programmed a number of functions in *R* and *Matlab* (MathWorks, Inc., 2007) to carry out a number of non-standard algorithms. In fact, my final project, presented in Chapter 5, is a software package named *tnet* which is a collection of most of the functions that I have written in *R*.

This software package was developed in response to the lack of open-source software programmes that can deal with weighted and longitudinal networks. For

example, the widely used *network*-package that many other open-source packages developed in *R*, notably the *sna* and *statnet*-packages, depend on, does not even have a data class for weighted networks (Butts, 2006; Butts et al., 2008; Handcock et al., 2003). Therefore, to allow for an assessment of weighted and/or longitudinal networks, a new platform is needed.

tnet is in itself an open-source package and allows others to add or modify functions. To this end, it contains two data structures, one for weighted networks and one for longitudinal ones, and a number of support functions. Based on these, researchers aiming to develop new measures or generalise existing ones to weighted or longitudinal networks have a platform on which they can easily do so. The package contains a set of functions for each of the two data structures. For the analysis of weighted networks, it includes a host of structural measures. Among these are functions to calculate the measures proposed in Chapters 2 and 3. Regarding the study of longitudinal networks, the package includes the framework proposed in Chapter 4. The goal of this project is to enable researchers to easily conduct a structural analysis of weighted and longitudinal networks by applying the measures proposed in this thesis and in the literature.

1.4 Network datasets

To test the methods proposed within this thesis, we have collected one weighted and longitudinal dataset. This dataset is an online social network created from a virtual community for students at University of California, Irvine, in the period between April to October 2004 (see Panzarasa et al., 2009, for a descriptive analysis). In this network, the nodes are $1,899$ students. When joining the community, each student was asked to create a profile. This profile contained a number of personal details. These details included the user's demographic characteristics, the user's list of friends, personal blogs, and forum postings. Based on these details, students

could search for others and base their decisions to communicate.

The ties are established when online messages $(59,835)$ are exchanged between the students. The weight of a directed tie is defined as the number of messages sent from one student to another. The maximum and average tie weight are 98 and 2.95, respectively. The students have on average 10.69 directed ties to others.

To ensure the protection of individuals and compliance with privacy laws, all individual identifiers were removed before we received the data. This included usernames, email and ip-addresses, and personal description. Each user was randomly assigned an identification number. In addition, the content of the online messages was not made available. For the messages, the information we received included only the identification numbers of sender and receiver, and the time at which the message was sent. Furthermore, two companies that gained access to the community through students with the purpose of mass-communication were excluded in order to filter out spamming activities. Moderators and other technical support staff with the only aim of facilitating the smooth functioning of the community were also excluded.

In addition, we have also relied on 6 dataset with 11 weighted networks used in the literature. The first three networks are from Freeman's EIES dataset (Freeman, 1978), also used in Wasserman and Faust (1994). This dataset was collected in 1978 and contains three networks of 32 researchers. The first is an acquaintance network of the group recorded at the beginning of the study (time 1). The second network is similar, but the data were recorded at the end of the study (time 2). The third is a frequency matrix of the number of messages sent between the researchers using an electronic communication tool. In the two acquaintance networks, all ties have a weight between 0 and 4. 4 represents a close personal friend of the researcher; 3 represents a friend; 2 represents a person the researcher had met; 1 represents a person the researcher had heard of, but not met; and 0 represents a person unknown to the researcher. In the frequency matrix, the average tie weight is 33.7 and the

maximum weight is 559.

The second dataset contains four networks are intra-organisational networks, two from a consulting company and two from a research team in a manufacturing company (Cross and Parker, 2004).[1] The consulting company had 46 employees who are the nodes in the first two networks. The ties in the first network are differentiated in terms of frequency of information or advice requests, whereas the ties in the second network reflect the value placed on the information or advice received. In both these networks, the directed ties are weighted on a scale from 0 to 5. The company had offices both in Europe and in the US. The US employees were divided in two tightly knit groups, while on the contrary, the European employees did not group together in the same way. The last two networks are based on a research team in a manufacturing company. The nodes in these networks are the 77 employees of the company. The ties in the first network are based on advice, whereas in the second network, they are based on the awareness of others' knowledge and skills. In both these networks, the directed ties are weighted on a scale from 0 to 6. The recording of the networks took place after an organisational restructuring, which meant that four separate units in different European countries had been combined. The research team was partitioned into strong communities based on employees' previous geographical location (Cross and Parker, 2004, pg. 15-17).

The third dataset includes a network representing political support in the US senate (101^{st} Congress, 1989/1990, also used in Skvoretz, 2002).[2] The nodes are 102 senators and ties are based on co-sponsorship on bills. The average tie value is 2.68 and the maximum value is 29.

The fourth dataset contains the neural network of the Caenorhabditis elegans worm. This was examined by Watts and Strogatz (1998) in their study of the small-

[1] We thank Andrew Parker at Stanford University for supplying this dataset.

[2] We thank John Skvoretz at University of South Florida for making this dataset available to us.

world phenomenon.[3] In this network, the nodes are 306 neurons and a tie joins two neurons if they are connected by either a synapse or a gap junction. The weight of a tie represents the number of these synapses and gap junctions between two neurons. The average weight is 3.74 and the maximum is 70.

The fifth dataset is the network of commercial airports in the United States (The US airport network). This network is publicly available on the website of the US Department of Transportation[4]. The nodes in this network are the 676 commercial airports in the U.S. Two airports are tied together if at least one fight was scheduled between them in 2002. The weight of each tie corresponds to the average number of seats per day available on the flights connecting the two airports (Barrat et al., 2004; Guimerà et al., 2005). There are a total of $3,523$ ties or scheduled routes consisting of one or more flights.

Finally, the sixth dataset is a scientific collaboration network collected by Newman (2001b,c). This network is created by the papers published on the arXiv repository in the area of condensed matter physics in the period from 1995 to 1999[5]. The nodes in this network are the authors of those papers, and a tie between two authors is established if they have co-authored at least one paper together. Following Newman (2001c), the weight attached to each tie is the sum over all the co-authored papers of the inverse of the size of the collaboration minus one. In other words, each paper increases the node strength of an author with 1, which is equally distributed across the weights attached to the ties directed towards the collaborators of that paper. For example, if an author writes a single paper with two others, the weight attached to the two ties would be 0.5. A consequence of this is that the node strength is a proxy of productivity, whereas the average strength reflects whether an author collaborates multiple times with a small group of others.

[3]This dataset was obtained from the Collective Dynamics Group's (Duncan Watts) website: `http://smallworld.sociology.columbia.edu/cdg/datasets/`

[4]`http://www.transtats.bts.gov/`

[5]`http://www.arxiv.org/`

2 Clustering in Weighted Networks[6]

This chapter has a methodological nature in that it builds on, and extends, a fundamental measure of network structure, namely the clustering coefficient, that has long received attention in both theoretical and empirical research. This measure assesses the degree to which nodes tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterised by a relatively high density of ties (Feld, 1981; Heider, 1946; Holland and Leinhardt, 1970; Freeman, 1992; Friedkin, 1984; Louch, 2000; Snijders, 2001; Snijders et al., 2006; Watts and Strogatz, 1998). More generally, one can ask: if there are three nodes in a network, i, j, and k, and i is tied to j and k, how likely is it that j and k are tied? In real-world networks, this likelihood tends to be greater than the average probability of a tie randomly established between two nodes (Holland and Leinhardt, 1971; Wasserman and Faust, 1994).

For social networks, scholars have investigated the mechanisms that are responsible for the increase in the probability that two people will be tied if they share an acquaintance (Holland and Leinhardt, 1971; Snijders, 2001; Snijders et al., 2006). The nature of these mechanisms can be social, as in the case of third-part referral (Davis, 1970; Heider, 1946), or non-social, as in the case of focus constraints (Feld, 1981). On the one hand, an individual may reduce cognitive stress by introducing his or her acquaintances to each other (Heider, 1946). Moreover, indirect ties foster trust, enhance a sense of belonging, facilitate the enforcement of social norms, and enable the creation of a common culture (Coleman, 1988). Burt (2005) found that reputation of a person is only maintained if his or her contacts can communicate or gossip. This also applies to inter-organisational networks where organisations in tightly knit groups create informal governance arrangements (Uzzi and Lancaster,

[6]An article based on this chapter has been published, see Opsahl and Panzarasa (2009).

2004). On the other, focus constraints refer to the increased likelihood of interaction and clustering among nodes that share the same physical, institutional, organisational or social environment. For example, people who share the same office are more likely to create independent dyadic ties leading to a heightened tendency towards clustering than people that reside in distant geographical locations (Feld, 1981).

Traditionally, the tendency of nodes to cluster together is measured using the global clustering coefficient (e.g. Feld, 1981; Karlberg, 1997, 1999; Louch, 2000; Newman, 2003) or the local clustering coefficient (Watts and Strogatz, 1998). This chapter deals with the former of these measures. Nevertheless, the local clustering coefficient is briefly introduced below to review and highlight differences among the two measures.

The local clustering coefficient is based on ego network density or local density (Scott, 2000; Uzzi and Spiro, 2005). For a node i, this is the fraction of the number of present ties over the total number of possible ties between node i's neighbours. For undirected networks, the local clustering coefficient is formally defined as:

$$C_i = \frac{\text{ties between node } i\text{'s neighbours}}{\text{node } i\text{'s neighbours } \times \text{ (node } i\text{'s neighbours } - 1)/2} \tag{1}$$

To obtain an overall coefficient for a network, the fractions for all the nodes in a network are averaged. The main advantage of this measure is that a score is assigned to each node. This enables researchers to study correlations with other nodal properties (e.g. Panzarasa et al., 2009) and perform regression analyses with the observations being the nodes of a network (e.g. Uzzi and Lancaster, 2004). However, this coefficient suffers from two major limitations. First, its outcome does not take into consideration the weight of the ties in the network. As a result, the same value of the coefficient might be attributed to networks that share the same topology, but differ in terms of how weights are distributed across ties and,

therefore, may be characterised by different likelihoods of friends being friends with each other. Second, the local clustering coefficient does not take into consideration the directionality of the ties connecting a node to its neighbours. A neighbour of node i might be: 1) a node that has directed a tie towards node i, 2) a node that node i has directed a tie towards, or 3) a node that has directed a tie towards node i and to whom node i has also directed a tie. Barrat et al. (2004) proposed a generalisation of the coefficient to take the weight to the ties into consideration. However, the issue of directionality still remains unsolved (Caldarelli, 2007).

Unlike the local clustering coefficient, the global coefficient is based on a clustering measure for directed networks: transitivity (Wasserman and Faust, 1994, 243). However, it is only defined for networks where ties are without weights. When the weights are attached to the ties, researchers have set an arbitrary cut-off level and then dichotomised the network by removing ties with weights that are below the cut-off, and then removing the weights from the remaining ties (this process is described in detail in Section 1.1). The result is a binary network consisting of ties that are either present (or equal to 1) or absent (or equal to 0; Scott, 2000). Doreian (1969) studied clustering in a weighted network by creating a series of binary networks from the original weighted network using different cut-offs. A sensitivity analysis can address some of the problems arising from the subjectivity inherent in the choice of the cut-off. However, it tells us little about the original weighted network, except that the value of the clustering coefficient changes at different cut-off levels. While we also conduct similar sensitivity analyses on various datasets, here we propose a generalisation that explicitly takes weights of ties into consideration and, for this reason, does not depend on a cut-off to dichotomise weighted networks.

In what follows, we start by discussing the existing literature on the global clustering coefficient in undirected and binary networks[7]. In Section 2.2 we propose

[7]Even though directionality of ties is a key advantage in choosing the global clustering coefficient over the local, for the sake of simplicity, we choose to start by focusing on undirected ties.

our generalised measure of clustering. We then test and compare the generalisation with the existing measure by using a number of empirical datasets based on weighted and undirected networks. In Section 2.4, we turn our attention to directed networks and discuss the existing literature on clustering in that type of network. We then extend our generalised measure to cover weighted and directed networks. Finally, Sections 2.5 and 2.6 highlight the contribution to the literature and offers a critical assessment of the main results.

2.1 Clustering coefficient

The global clustering coefficient is based on triplets of nodes. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties. The global clustering coefficient is the number of closed triplets over the total number of triplets (both open and closed). The number of closed triplets has also been referred to as $3 \times$ triangles in the literature. This is due to the fact that three closed triplets form a triangle, one triplet centred on each of the three nodes in the triangle. The first attempt to measure the coefficient was made by Luce and Perry (1949). For an undirected network, they showed that the total number of triplets could be found by summing the non-diagonal cells of a squared binary matrix. The number of closed triplets could be found by summing the diagonal of a cubed matrix. For clarity, we will refer to the global clustering coefficient as the *binary* clustering coefficient, C:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of triples}} = \frac{\sum \tau_{\Delta}}{\sum \tau}, \tag{2}$$

where $\sum \tau$ is number of triplets and $\sum \tau_{\Delta}$ is the subset of these triplets that are closed by the addition of a third tie. The coefficient takes values between 0 and 1. In a completely connected network $C = 1$ as all triplets are closed, whereas in classical random networks $C \to 0$ as the network grows. More specifically, in classical random

networks, the probabilities that dyads (i.e., pairs of nodes) are tied together are by definition independent (Erdős and Rényi, 1959; Solomonoff and Rapoport, 1951). Therefore, C is equal the probability of a tie in these networks (Newman, 2003).

A major limitation of the binary clustering coefficient is that it cannot be applied to a weighted network. As a result, the same outcome might be attributed to networks with different likelihoods of friends being friends with each other. This could bias the analysis of the network structure. In order to overcome this shortcoming, in the following section we will introduce a generalisation of the clustering coefficient that captures the richness of tie weights, while at the same time producing the same outcome as the binary clustering coefficient when all ties have the same weights (i.e., a binary network).

2.2 Generalised clustering coefficient

We can generalise the clustering coefficient, C, to take tie weights into consideration by rewriting Equation 2 and defining a triplet value, ω. It is vital to use an appropriate method for defining the value of a triplet as this impacts on the outcome of the coefficient. The method should be chosen based on the research question as well as the way in which the weights of the ties are operationalised. First, the triplet value, ω, can be defined as the arithmetic mean of the weights of the ties that make up the triplet. This is the simplest method of calculating the triplet value. However, this method is not sensitive to differences between the two tie weights as an extreme value can have a major impact on the triplet value. Second, ω can be defined as the geometric mean of the weights attached to the two ties. This method overcomes some of the sensitivity issues as the triplet made up by a tie with a low value and a tie with a high value will have a lower value than if the arithmetic mean were used. Additional methods defines the triplet value as the maximum or minimum value of the two weights. These two methods represent extreme cases. The "maximum"

method offsets a low tie weight and makes a triplet with a strong tie and a weak tie equal to a triplet with two strong ties. Conversely, the "minimum" method offsets a high tie weight by making triplets with a strong tie and a weak tie equal to triplets with two weak ties. Table 1 highlights the differences between the methods of defining the triplet value. We will explore them further at the end of Section 2.3.

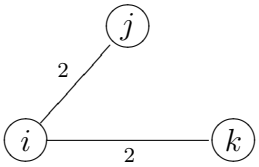
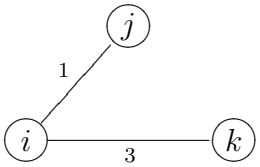
	Triplet value ω of	
		
Method		
Arithmetic mean	$(2+2)/2 = 2$	$(1+3)/2 = 2$
Geometric mean	$\sqrt{2\times 2} = 2$	$\sqrt{1\times 3} = 1.73$
Maximum	$\max(2,2) = 2$	$\max(1,3) = 3$
Minimum	$\min(2,2) = 2$	$\min(1,3) = 1$

Table 1: Methods of calculating the triplet value, ω.

Once an appropriate method for defining ω has been chosen, we can generalise the clustering coefficient to take ω into account as follows:

$$C_\omega = \frac{\textit{total value of} \text{ closed triplets}}{\textit{total value of} \text{ triplets}} = \frac{\sum_{\tau_\Delta} \omega}{\sum_{\tau} \omega} \tag{3}$$

The generalised clustering coefficient produces the same result as the binary clustering coefficient if applied to a binary network. This is because all triplets have the same value, $\omega = 1$, irrespective of the method used to calculated triplet value. In addition, the generalised coefficient shares the same properties of the binary coefficient. It still ranges between zero and one because neither numerator nor denominator of the fraction can be negative; moreover, every element that can be part of the numerator is part of the denominator. In a completely connected, network all triplets are closed as the third tie will always be present, e.g. between node j and node k in Table 1. Therefore, all triplets are part of both the numerator

and denominator: $C_\omega = \frac{1}{1} = 1$. To test whether $C_\omega \to 0$ as the size of a classical random network increases or, more specifically, C_ω equals the probability dyads to be tied together, we created a set of random networks with different sizes, but with a fixed degree. Since classical random networks are binary, we assigned a random weight between 1 and 10 to the ties in an effort to simulate a weighted network. We applied the generalised clustering coefficient to these networks, and found that $C_\omega \to 0$ as the network size increases. In particular, as shown by Table 2, we found that C_ω was very close to the probability of a tie in classical random networks. Furthermore, to assess the sensitivity to weights, we tested the generalised clustering coefficient on networks where the network structure was not randomised, but where the weights are randomly assigned to the ties. We found $C_\omega \approx C_{GT0}$, where C_{GT0} is the clustering coefficient calculated on binary networks where all ties with positive values are set to present.[8]

parameters		generalised clustering coefficient				maximum
N	p(tie)	$C_{\omega,am}$	$C_{\omega,gm}$	$C_{\omega,\max}$	$C_{\omega,\min}$	|p(tie)$-C_\omega$|
50	0.2040816	0.2026950	0.2026728	0.2027486	0.2025926	0.001489
100	0.1010101	0.1006199	0.1006022	0.1006502	0.1005631	0.000447
200	0.0502513	0.0502009	0.0501983	0.0502018	0.0501985	0.000053
400	0.0250627	0.0250521	0.0250474	0.0250553	0.0250460	0.000017
800	0.0125157	0.0124955	0.0124988	0.0124904	0.0125050	0.000025
1600	0.0062539	0.0062682	0.0062670	0.0062700	0.0062648	0.000016

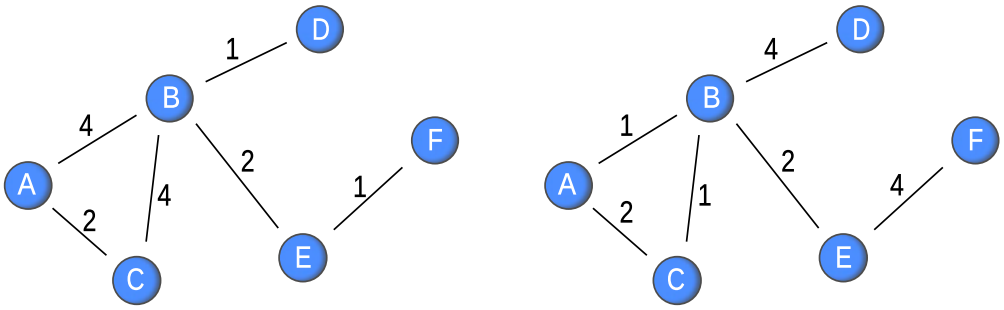
Table 2: Simulations of the generalised clustering coefficient on ensembles of classical random networks with 50, 100, 200, 400, 800, and 1,600 nodes and an average degree of 10 where ties are assigned a random weight between 1 and 10. Each ensemble contains 1,000 scenarios. The four methods for defining triplet value were not significantly different ($p > 0.7$).

In this chapter, we assume that weights are positive values and that a *higher* value is *better* than a lower one. In situations where the second part of this assump-

[8]This finding is based on the empirical networks presented in Section 2.3. For each network, we reshuffled the weights globally among the ties in the network, and calculated C_ω. This randomisation procedure maintains the topology of the observed network, and therefore C_{GT0} does not change. We compared C_ω (1,000 scenarios) to C and found that they were not statistically significantly different.

tion is not appropriate, e.g. when weights of ties refer to costs and lower values are better than higher ones, weights should be inverted (Newman, 2001c). Then, in the example of costs, a low cost will have a *higher* value than a high cost. Furthermore, we will use the absolute values of weights without normalising them (e.g., by dividing them by their maximum or average) as this would have no effect on the results of our analysis.

To illustrate and exemplify the applicability of the generalised clustering coefficient, Figure 2 shows two sample networks with six nodes and six weighted ties. In network *a*, the ties between the nodes that form the triangle have a higher weight than the average tie weight in the network, whereas the reverse is exemplified in network *b*.



(a) Stronger triplets being closed (b) Weaker triplets being closed

Figure 2: Two weighted sample networks.

Both sample networks have the same binary clustering coefficient if the networks are transformed by setting ties with a weight greater than 0 ($GT0$) to present:

$$C_{GT0} = \frac{3 \times 1}{9} = 0.33 \tag{4}$$

However, we believe it is not accurate to claim that both these networks have the same tendency of "friends to be friends themselves" (Freeman, 1992). Friendship can be assessed using the same criteria that Granovetter (1973) used for defining tie weights (duration, emotional intensity, intimacy, and exchange of services). If, for

example, the tie weights in the sample networks in Figure 2 correspond to duration, we can say that the nodes in network a are spending more time on other nodes that are themselves tied together than in network b. In turn, it could be argued that node B's friendships are transitive in network a, whereas the node's acquaintances are transitive in network b.

The generalised clustering coefficient shows a difference between the two sample networks. For network a and b in Figure 2, the coefficients obtained by using the "geometric mean" method for defining the triplet values ($C_{\omega,gm}$) are respectively:

$$C_{\omega,gm} = \frac{9.656854}{22.14214} \approx 0.44 \tag{5a}$$

$$C_{\omega,gm} = \frac{3.828427}{16.31371} \approx 0.23 \tag{5b}$$

The difference between the two coefficients results from the fact that the generalised clustering coefficient captures more information than the binary one. In fact, the difference between Equation 5a and 5b is a reflection of the differences in tie weights of the two sample networks. Burt (1992, 2005) defined a person as a less efficient information broker if he or she spends time with people that are themselves connected or part of the same group. Conversely, if the person spends time with disconnected people or people part of different groups, he or she could act as a gateway and control the information flowing between the people or groups. If the sample networks were knowledge networks, it can be argued that the people in network a are *less* efficient in being brokers and are in *less* advantageous positions to control information than the ones in network b.

The weight of the closing tie of a triplet is not considered in the proposed generalisation of the clustering coefficient. This is because we believe that the aim of the clustering coefficient is to assess the likelihood of the closing tie to be created,

and not the strength of this tie. Formally, we see the closing tie as a product of the triplet. In other words, as networks evolve over time by the creation and removal of ties, clustering occurs when a triplet exists and a third tie is created, so that the triplet is closed. However, in a cross-sectional network, we only observe a triangle and cannot determine which of the three triplets, that make up the triangle, occurred first. In effect, this means that the weight of the closed tie of a triplet is considered since it is part of the other two triplets in the triangle. Chapter 4 extends the proposed generalised coefficient to weighted longitudinal networks.

2.3 Empirical tests

We now test the proposed generalisation on ten of the networks outlined in Chapter 1.[9,10] We also compare the generalised clustering coefficient with the binary one measured with different cut-offs. In particular, we compare the generalised coefficient with the one found where all ties with a weight greater than 0 is set to present. This comparison allows us test a key assumption used by Granovetter's (1973) in his the strength of weak ties theory (Freeman, 1992). He assumed that a person's friends (i.e., strong ties) are more likely to be connected with each others than acquaintances (i.e., weak ties). Based on the assumption that acquaintances were

[9]We do not apply the proposed method to the scientific collaboration and the US airport network. The scientific collaboration network is a one-mode projection of a two-mode network that discounts the tie weights within large collaboration. This creates a number of biases for the weighted clustering coefficient as a single large collaboration would create many closed triplets composed of weak ties (see Section 1.4). Therefore, we choose not the use this dataset. The complete US airport network is not publicly available, and this Chapter was not created in collaboration with others who had access to the network. Thus, we could not use this network in this Chapter.

[10]Network measures and statistics are calculated on symmetrised matrices for directed networks. We choose to symmetrise using the sum method (i.e., tie weights are equal to the sum of the weights attached to the two possible directed ties between two nodes), because this method creates a wider range of values instead of the maximum method. For the binary coefficient, the networks are also dichotomised with different cut-offs (i.e., C_{GTX} refers to Eq. 2 where ties with a weight greater than X are set to present, whereas ties with a weight lower than or equal to X are removed). Unless otherwise specified, ties are set to present if their weight is greater than 0, and the generalised coefficient uses the geometric mean method for defining triplet value ($C_{\omega,gm}$). A function to calculate the binary and generalised clustering coefficient is available in the *tnet* package implemented in the open-source statistical programme *R* (see Chapter 5 for more details).

disconnected from a person's other contacts, he argued that they were more likely to bring novel information to the focal person than friends as they moved in different social circles, whereas friends moved in the same social circle.

The first three networks are Freeman's EIES networks (Freeman, 1978). The three networks are highly dense with densities (the randomly expected value of the clustering coefficient) of 0.78, 0.87, and 0.54, respectively. They also exhibit a fairly large tendency toward clustering; C_{GT0} for the three networks is 0.8417, 0.9010, and 0.6569, respectively. When the proposed generalisation is applied, all three networks experience an increase in clustering: $C_{\omega,gm} = 0.8593$, 0.9146, and 0.7654, respectively. For the acquaintance networks the increase is 2.1% and 1.5%, whereas for the frequency matrix the increase is relatively higher at 16.5%.

The fourth network is the online social network (Panzarasa et al., 2009). This network exhibits a density of 0.0077 and a clustering coefficient of $C_{GT0} = 0.0568$. We found a generalised clustering coefficient of $C_{\omega,gm} = 0.0694$. This represents an increase of 22.2%.

The next four networks are the intra-organisational networks (Cross and Parker, 2004). All four networks do exhibit a high clustering coefficient: C_{GT0} ranges between 0.6723 and 0.7242, and $C_{\omega,gm}$ between 0.7200 and 0.7750. This represents an average increase of 7.1% when the generalised coefficient is applied.

The ninth network represents political support in the US senate (101^{st} Congress, 1989/1990, also used in Skvoretz, 2002). This network has a density of 0.78. As the network is almost fully connected, it is difficult to draw conclusions from C_{GT0}. We found: $C_{GT0} = 0.8415$. The average undirected tie weight is 3.99 and the maximum weight is 42. The great difference between the mean and the maximum signals that most ties are relatively weak. This is an indication that a higher cut-off might be more appropriate. In Table 3, we list C calculated using higher cut-offs. Applying the generalised coefficient, we found $C_{\omega,gm} = 0.8726$. This represents an increase of

3.7%. A plausible explanation for this increase is the fact that party membership and ideologies represent a constraint on the strength of ties among senators (open triplets). In particular, senators belonging to different parties are likely to co-sponsor a limited number of bills. Thus, the value of closed triplets connecting senators from different parties tends to be small.

The tenth network is the neural network of the Caenorhabditis elegans worm. The density is 0.0460 in this network. We found: $C_{GT0} = 0.1807$ and $C_{\omega,gm} = 0.1748$. This is a decrease of 3.3%.

Table 3 sums up the empirical results. A number of observations are in order. First, the binary clustering coefficient, C_{GTX}, generally decreases as the cut-off, X, increases for all of the networks. The rate of decrease differs considerably among the networks. Moreover, the decrease is not linear. In fact, although not shown in Table 3 as only every second value of the cut-off is shown, the clustering coefficient even increases at some of the levels. In addition, the reliability of the results when high cut-offs are used should be questioned as there are only few triplets and triangles left in the network with high cut-offs. Thus, these findings from a sensitivity analysis of the binary clustering coefficient are difficult to interpret.

Second, there are variations in the generalised clustering coefficient, C_ω, when different methods for defining the triplet value are used. The highest C_ω is attained when the "minimum" method is used, whereas the lowest outcome is attained when the "maximum" method is used for most of the networks. This means that triplets consisting of two ties with approximately the same weight are likely to be closed. Different results are found for Freeman's frequency matrix, C.elegans' neural network, and the online social network, with the reverse found in the first two networks. A possible reason for this difference is that these three networks have the greatest range of tie weights (946, 71, and 183, respectively). If certain ties have extremely large weights attached to them, the fraction in Eq. 3 becomes sensitive to whether

network	C_ω				C					
	$C_{\omega,am}$	$C_{\omega,gm}$	$C_{\omega,\max}$	$C_{\omega,\min}$	C_{GT0}	C_{GT2}	C_{GT4}	C_{GT6}	C_{GT8}	C_{GT10}
Freeman EIES (time 1)	0.856	0.859	0.849	0.868	0.842	0.671	0.405	0.333		
Freeman EIES (time 2)	0.913	0.915	0.909	0.919	0.901	0.752	0.444	0.375		
Freeman EIES (messages)	0.782	0.765	0.791	0.745	0.657	0.648	0.593	0.573	0.559	0.541
Online community	0.069	0.069	0.069	0.069	0.057	0.043	0.035	0.029	0.030	0.029
Consulting (advice)	0.768	0.775	0.758	0.788	0.724	0.595	0.533	0.494	0.493	
Consulting (value)	0.728	0.732	0.718	0.744	0.706	0.693	0.631	0.577	0.479	
Research team (advice)	0.724	0.744	0.701	0.775	0.672	0.667	0.594	0.467	0.340	0.231
Research team (awareness)	0.710	0.720	0.690	0.742	0.672	0.660	0.668	0.671	0.651	0.569
101^{st} Congress	0.869	0.873	0.865	0.878	0.842	0.612	0.463	0.358	0.320	0.286
C.elegans' neural network	0.187	0.175	0.194	0.167	0.181	0.097	0.060	0.049	0.039	0.053

Table 3: Comparison between the generalised and the binary clustering coefficients.

these ties form part of closed or open triplets. Moreover, the similarity of the two tie weights in a triplet might affect the likelihood of closure. For example, there is greater variation in the weights of the second triplet in Table 1 than in the ones of the first triplet. If these two triplets are closed, the second triplet would add more to the fraction in Eq. 3 than the first triplet if the maximum method is used, whereas the reverse is true if the minimum method is used. Therefore, by obtaining the highest outcome using the "maximum" method for Freeman's frequency matrix and C.elegans' neural network, we argue that triplets consisting of ties with greater variation in weights are more likely to be closed than triplets consisting of ties with roughly the same weight.

Third, the generalised clustering coefficient is higher than the binary coefficient for all social networks. When networks are dichotomised by setting ties with weights greater than 0 to present, the binary clustering coefficient is a benchmark for the generalised one. As shown by simulations in Section 2.2, when the tie weights are randomly assigned to the ties, $C_\omega \approx C_{GT0}$. By comparing C_ω to C_{GT0}, we can assess whether strong triplets are more likely to be closed than weak triplets. If the generalised clustering coefficient is significantly higher than the binary clustering coefficient, strong triplets are more likely to be closed than weak ones, whereas when the reverse is the case, weak triplets are more likely to be closed than strong ones. The findings therefore support the claim by Granovetter (1973) that stronger ties are more likely to be part of transitive triplets in social networks than weak ones.

2.4 Directed networks

Directed data increase the difficulty of calculating the clustering coefficient. In directed networks, two directed ties might exist within a dyad – one in each direction. A network can be represented by an adjacency matrix, x. If a directed tie from node i to node j is present, the cell $x_{ij} = 1$. This matrix is a special case of the weighted adjacency matrix, w. In this matrix, the cell w_{ij} is the weighted of the tie x_{ij}. We define the triplet consisting of the two directed ties, x_{ji} and x_{ik}, as $\tau_{ji,ik}$, and the value of this triplet as $\omega_{ji,ik}$.

The binary clustering coefficient as stated in Equation 2 cannot be applied to directed data. A more refined measure to calculate clustering in directed networks is called transitivity, T (for a review, see Wasserman and Faust, 1994, pg. 243). Transitivity produces the same results as the binary clustering coefficient if applied to an undirected network (Feld, 1981; Newman, 2003) and shares the same properties. More specifically, $0 \leq T \leq 1$, in a completely connected network $T = 1$, and in a classical random network $T \to 0$ as the network size grows if the average degree is constant. Transitivity takes the direction of the ties into consideration by using a more sophisticated definition of a triplet. A triplet τ, centred on node i, must have one in-coming and one out-going tie, i.e. $x_{ki} = x_{ij} = 1$ or $x_{ji} = x_{ik} = 1$ as shown in Figures 3a and b, respectively.

Figure 3: Non-vacuous triplets centred around node i

Wasserman and Faust (1994) termed triplets that do not fulfill this condition as vacuous. These triplets are not part of the numerator nor of the denominator of the fraction. More specifically, when we are dealing with directed data there can be four

basic configurations of a triplet around an individual node i: $\tau_{ij,ik}$, $\tau_{ij,ki}$, $\tau_{ji,ik}$, and $\tau_{ji,ki}$. The configurations $\tau_{ij,ik}$ and $\tau_{ji,ki}$ form, respectively an out- and in-star, and therefore, are vacuous and not part of the fraction. Conversely, the configurations $\tau_{ij,ki}$ and $\tau_{ji,ik}$ are non-vacuous. These can be either transitive or intransitive.

Triplets defined according to Wasserman and Faust (1994) form chains of nodes. These triplets have been termed 2-path as they form chains of two directed ties between three nodes (Luce and Perry, 1949). A triplet is transitive if a tie is present from the first node to the last node of the chain, $x_{kj} = 1$ and $x_{jk} = 1$ for the triplets shown in Figure 3a and b, respectively. If some of the ties between the nodes in a triplet are reciprocated (i.e., there exist two directed ties within a single dyad), there might exist multiple triplets between the nodes.

Transitivity suffers from the same limitation as the binary clustering coefficient in that it cannot be applied to networks where weights are attached to ties. To overcome this shortcoming, we extend our proposed generalisation to directed and weighted networks by using the same definition of a triplet, τ, as transitivity. The triplet value, ω, is calculated using the same methods as stated in Section 2.2; however, we use the weights of the two directed ties that form the triplet instead of the two undirected ties. The generalised coefficient, T_ω, produces the same results as transitivity if applied to a binary and directed data, and the same result as the binary clustering coefficient if applied to binary and undirected data. Moreover, it still ranges between 0 and 1, and in a completely connected network we would still obtain $T_\omega = 1$, whereas in a classical random network $T_\omega \to 0$ as the network grows. In particular, we found that T_ω approximates the probability of a directed tie in classical random networks, as shown by Table 4.

To clarify which triplets are transitive and non-vacuous, Table 5 illustrates configurations of triplets centred on node i. The first four rows show the basic configurations mentioned above. The remaining rows show configurations of triplets

parameters		generalised clustering coefficient				maximum
N	p(tie)	$T_{\omega,am}$	$T_{\omega,gm}$	$T_{\omega,\max}$	$T_{\omega,\min}$	|p(tie)$-T_{\omega}$|
50	0.2040816	0.2032126	0.2032204	0.2032211	0.2031963	0.000885
100	0.1010101	0.1009727	0.1009777	0.1009813	0.1009575	0.000053
200	0.0502513	0.0502643	0.0502640	0.0502632	0.0502664	0.000015
400	0.0250627	0.0250830	0.0250873	0.0250760	0.0250960	0.000033
800	0.0125157	0.0125385	0.0125381	0.0125386	0.0125383	0.000023
1600	0.0062539	0.0062570	0.0062582	0.0062551	0.0062606	0.000007

Table 4: Simulations of the generalised clustering coefficient on ensembles of classical random directed networks with 50, 100, 200, 400, 800, and 1,600 nodes and an average out-degree (i.e., the number of directed ties originating from a node) of 10, where ties are assigned a random weight between 1 to 10. Each ensemble contains 1,000 scenarios. The four methods for defining triplet values were not statistically significantly different ($p > 0.7$).

where ties are reciprocated. More specifically, each additional directed tie doubles the number of triplets. In addition, the table shows which triplets are transitive under different conditions and which triplet values should be included in the fraction of Equation 3.

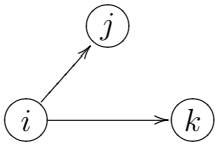
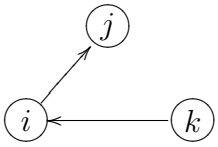
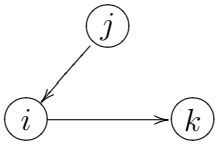
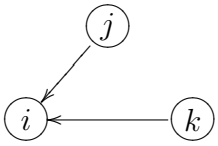
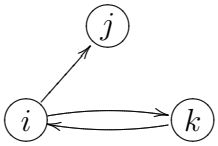
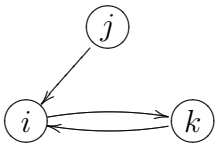
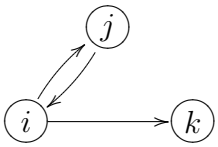
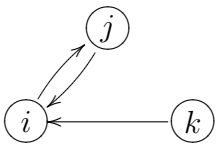
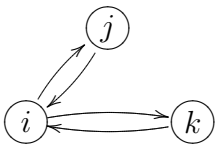
	Triplets	Denominator	Numerator if $w_{jk}=0$ $w_{kj}=0$	Numerator if $w_{jk}>0$ $w_{kj}=0$	Numerator if $w_{jk}=0$ $w_{kj}>0$	Numerator if $w_{jk}>0$ $w_{kj}>0$
	$\tau_{ij,ik}$	...	...	...	...	...
	$\tau_{ij,ki}$	$\omega_{ij,ki}$	0	0	$\omega_{ij,ki}$	$\omega_{ij,ki}$
	$\tau_{ji,ik}$	$\omega_{ji,ik}$	0	$\omega_{ji,ik}$	0	$\omega_{ji,ik}$
	$\tau_{ji,ki}$	...	...	...	...	...
	$\tau_{ij,ik}$ $\tau_{ij,ki}$	... $\omega_{ij,ki}$	... 0	... 0	... $\omega_{ij,ki}$	... $\omega_{ij,ki}$
	$\tau_{ji,ik}$ $\tau_{ji,ki}$	$\omega_{ji,ik}$ ...	0 ...	$\omega_{ji,ik}$ ...	0 ...	$\omega_{ji,ik}$ ...
	$\tau_{ij,ik}$ $\tau_{ji,ik}$	... $\omega_{ji,ik}$	... 0	... $\omega_{ji,ik}$	... 0	... $\omega_{ji,ik}$
	$\tau_{ij,ki}$ $\tau_{ji,ki}$	$\omega_{ij,ki}$ ...	0 ...	0 ...	$\omega_{ij,ki}$ ...	$\omega_{ij,ki}$ ...
	$\tau_{ij,ik}$ $\tau_{ij,ki}$ $\tau_{ji,ik}$ $\tau_{ji,ki}$	... $\omega_{ij,ki}$ $\omega_{ji,ik}$ ...	... 0 0 ...	... 0 $\omega_{ji,ik}$ ...	... $\omega_{ij,ki}$ 0 ...	... $\omega_{ij,ki}$ $\omega_{ji,ik}$ ...

Table 5: Triplets, τ, and triplet values, ω, in a directed network. $i \neq j \neq k$.

2.5 Contribution to the literature

Our generalisation of the clustering coefficient represents an improvement of current network methods in that it helps capture the richness and complexity contained within the structure of weighted networks. By using the generalised coefficient, we can thus obtain a better understanding of the topology of weighted networks. The explanatory power of the existing clustering coefficient is limited due to the fact that it can only be applied to binary networks. This shortcoming has been overcome by another clustering measure, the local clustering coefficient. However, this measure is sensitive to the number of contacts a node has (Soffer and Vázquez, 2005) and its applicability is restricted to undirected networks (for a review, see Caldarelli, 2007). The non-local, or global, clustering coefficient that we focus on does not suffer from this sensitivity, and it is also defined for both undirected and directed networks. Thus, the generalisation is a step forward towards a more fine-grained analysis of weighted networks.

To exemplify the applicability of the generalised coefficient, we applied it to a number of empirical datasets ranging from friendship networks to neural ones. In every social network that we tested, we found that the generalised coefficient obtained a higher value than the binary one calculated when all positive ties were set to present. This signals that strong ties are more likely to occur inside triangles rather than outside. This result validates Granovetter's (1973) assumption that weak ties are more likely to occur outside triangles than inside (Freeman, 1992). This assumption builds on Simmel's (1950) work and refers to the fact that a person is less likely to have common contacts with "acquaintances" than with "friends", and has implications for knowledge transfer. On the one hand, friends move in the same social circle and, therefore, their knowledge is likely to overlap (Coleman, 1988). On the other, acquaintances move in different social circles. This gives them access to novel pieces of information (Burt, 1992). By assessing whether strong ties occur

within triangles, we provide a method for quantitatively testing this assumption.

2.6 Conclusion and discussion

Relations to unique people are unique. We live in an increasingly connected world with an increasing number of contacts to whom we relate in different ways, with different frequencies, and for different reasons. Each social relationship bears a special meaning to us, and it would be overly simplistic and grossly unfair to treat every contact in the same manner. Therefore, it is important to capture this difference when studying social networks. We believe that social network measures should capture the richness of information that the weight of ties contains. There are a great number of networks where the weights of the ties are recorded (see Section 2.3, but also Ebel et al., 2002; Holme et al., 2004; Kossinets and Watts, 2006), nevertheless only a limited number of measures take the weight into account (e.g. Burt, 1992; Freeman et al., 1991; Nordlund, 2007; Yang and Knoke, 2001). Therefore, most measures can only be calculated on binary networks. This means that researchers must set a subjective cut-off, and ties whose weight falls below this cut-off are removed, and those whose weight is above are simply set to present (Doreian, 1969). However, this procedure is detrimental to the quality of the analysis, primarily because it compromises the information that the data holds.

To overcome this shortcoming of network measures applied to weighted networks, we offered a generalisation of the clustering coefficient that takes the weight of ties explicitly into account. It does so by attaching a value to each triplet. A triplet consists of two ties, either two undirected ties or two directed ties, depending on the nature of the network. The value of a triplet is based on the weight of these two ties. We proposed four methods for calculating this value. The first two methods utilise a mean algorithm, namely the arithmetic mean and the geometric mean. These methods discount a strong tie when it is coupled with a weak tie, with the "geomet-

ric mean" method discounting more than the "arithmetic mean" method. The last two methods represent extreme methods, namely the "maximum" and "minimum" methods. These methods are not sensitive to differences between the weights of the two ties that constitute the triplet. They simply set the value of the triplet equal to the maximum or the minimum weight of the two ties, respectively. The advantages and shortcomings of each of these four methods should be evaluated based on the research question and the type of network dataset at hand. For example, in a network where the weights correspond to the level of flow, and a weak tie would act as a bottleneck, the minimum method might be most appropriate to use. By contrast, when ties are weighted in terms of costs or time, it may be more suitable to apply the maximum method so as not to underestimate the value of triplets. The appropriateness of the two methods based on averages is linked to the question of whether extreme tie weights should be discounted. For example, the geometric mean might be more appropriate than the other methods when studying knowledge transfer in a social network. This is due to the fact that proportionally less knowledge is transferred over a strong tie (e.g., a tie with a weight of five is likely not transfer five times the knowledge as a tie with a weight of one; Granovetter, 1973). In an attempt to account for this feature, the geometric mean might be suitable as it discounts triplets with extreme weights (see Table 1). On the contrary, in networks where the weights are directly proportional to capacity, the arithmetic average might be more suitable. This is often the case in non-social networks, such as the US airport network.

The generalised coefficient produces the same results as the binary clustering coefficient when applied to a binary network. The binary coefficient divides the number of closed triplets by the total number of triplets, whereas the generalised coefficient divides the total value of the closed triplets by the total value of all triplets. When the data are binary, all the triplets have the same value, $\omega = 1$,

regardless of the method used to calculated the value. Therefore, the total value of closed triplets equals the number of closed triplets, and the total value of triplets equals the total number of triplets. Hence, the generalised coefficient equals the binary coefficient when applied to a binary network.

We measured and compared the binary and the generalised clustering coefficients on a number of empirical datasets where the weights of relations are recorded. First, we found that the binary coefficient generally decreased as the cut-off increased. However, as the rate of decrease varies, it is difficult to interpret this result. Second, we found that there were differences among the outcomes when different methods for defining the triplet value were used. The generalised coefficient using the "minimum" method yielded mostly the highest outcome, whereas when the "maximum" method was used, the lowest outcome was generally attained. There were three exceptions. These refer to the networks with the greatest range of tie weights. We speculated that this might have affected results because the difference between the weights of the ties that make up a triplet in these networks is likely to be relatively greater than in other networks. Third, we found that, in all social networks studied, the value of the generalized coefficient was greater than the value of the binary one. This signal that triplets composed out of stronger ties are more likely to be closed than triplets composed of weaker ties. This is not surprising as social interactions generally occur in groups larger than two, and if two people spend a great deal of time with the same third person, they are also likely to meet and develop a bond with each other. Moreover, this finding provide support in favour of an assumption used by Granovetter's (1973) when he formulated the strength of weak ties theory (Freeman, 1992). Our finding suggests that triplets compose of weak ties are more likely to be open than triplets composed of strong ties. Thus, acquaintances are less likely to be connected with a person's contacts than friends.

One of the advantages of the generalised clustering coefficient is also a limitation.

As opposed to the binary clustering coefficient, ties in weighted networks are not transformed. This becomes an issue when all the ties within a network are defined, even by a marginal weight, because the network is fully connected, and the coefficient is 1. The binary clustering coefficient does not have this issue as ties with a marginal weight are set to absent, and the network is therefore no longer fully connected. An example of a weighted, fully connected network, is a network consisting of cities and where ties among cities represent distances. Since there is a finite distance among all cities, all possible ties in this network are defined. The binary clustering coefficient overcomes this issue by setting weak ties (or long distances) to absent. A possible solution when applying the generalised coefficient (which does not normally transform the data) is exactly to apply this transformation and set weak ties to absent. However, the suitability and appropriateness of this solution depends on the data, the context in which the data were collected, and the research question.

We believe that researchers should carefully operationalise variables when dealing with research questions concerned with tie weights. Marsden and Campbell (1984) conducted a comparative analysis of Granovetter's (1973, pg. 1361) four criteria for defining tie weights. They found that emotional intensity was a better indicator of friendship than the other three criteria. Conversely, we believe that researchers should suitably assess which criterion represents the most appropriate measure for operationalising each variable of the study. This, in turn, will depend on the nature of the nodes, ties, and more generally on the context of the research setting.

In addition, the scale of the weights should be carefully defined. The scale should represent the chosen criteria to minimise subjective biases. A standard network question often used in the studies of advice networks is:

> *Please indicate how often you have turned to this person for information or advice on work-related topics in the past three months.*

with the scale: 0, Do not know this person; 1, Never; 2, Seldom; 3, Sometimes;

4, Often; 5, Very Often.[11] In this case, answers are prone to the inevitable bias that comes from the different ways in which different people assess duration and define the time-related scale. One way to overcome this problem is to design a scale that reflects actual time. For example, a better scale could be: 0, Never; 1, Once; 3, Monthly; 6, Bi-weekly; 12, Weekly. In turn, this scale, when compared to the former, is likely to yield a dataset that is richer in information, more robust against potential subjective biases, and more suitable for network studies that rely on generalised measures, such as our proposed clustering coefficient.

[11]Cross and Parker (2004) used this question to create the advice network in the consulting company used in Section 2.3.

3 Prominence and Control: The Weighted Rich-club Effect[12]

As described in the previous chapters, there is a growing sense of urgency within the scientific community to develop measures for weighted networks. Most efforts have been directed towards the generalisation of binary network measures (e.g. Barrat et al., 2004; Freeman et al., 1991; Newman, 2001c; Opsahl and Panzarasa, 2009); however, weighted networks can contain patterns and regularities that do not exist in binary networks. For example, the location of the strongest ties in a network is only possible to identify if the ties are weighted. Due to the increase in the number of collected weighted networks, the development of methods to detect new features of these networks is imperative.

This Chapter aims to study whether a subset of nodes exchange among themselves the strongest ties in the network. To select a subset of nodes, we rely on previous studies have shown that the elements of a wide range of systems, ranging from technological to economic and social ones, are organised into hierarchies (Pareto, 1897; Price, 1965; Simon, 1955; Zipf, 1935). For example, Pareto (1897) argued that 80% of the wealth within a society is controlled by 20% of the population. This hetrogeniety is also found for a variety of properties in a wide range of networks (Barabási and Albert, 1999; Barrat et al., 2004; Pastor-Satorras and Vespignani, 2004). For example, Barabási and Albert (1999) found that the vast majority of links on the Internet point towards a relatively small subset of webpages. Studying the nature of the interactions between the nodes at the top of the hierarchies (the prominent nodes) can provide useful insights into the network's organisation and functioning. In fact, scholars have already embarked on this avenue of investigation. In particular, Colizza et al. (2006) tested the tendency of highly connected nodes to

[12]This chapter is based on a collaboration with Vittoria Colizza, Pietro Panzarasa, and José J. Ramasco (see Opsahl et al., 2008).

form tighter interconnected groups than randomly expected. This property is known as the topological rich-club phenomenon (Colizza et al., 2006; Zhou and Mondragon, 2004). By allowing us to discover patterns of interactions (or their absence) at the top of the hierarchy, the rich-club phenomenon helps highlight organisational principles in the network. This approach, however, is limited by the binary nature of ties on which it draws, whereas a wealth of information is contained within the strength of ties (Barrat et al., 2004; Reagans and McEvily, 2003).

Given the relevance of the strong ties in many processes (see Section 1.1 for more details), this chapter proposes a new general measure aimed at evaluating whether, and the extent to which, strong ties occur among prominent nodes. Unlike the topological rich-club assessment which focused on the highly connected nodes, we do not limit the analysis to solely this group of nodes and define the prominent nodes as the ones that rank at the highest levels according to any ordering property in the network. This flexibility reflects the empirical findings that a wide range of ordering properties are heterogeneously distributed in networks (Pastor-Satorras and Vespignani, 2004; Serrano et al., 2007; Zlatic et al., 2008).

The analysis is undertaken within a two-fold framework. First, by focusing on an ordering property, a subset of prominent nodes is selected. Second, the preference of these nodes to direct their efforts towards one another, by forging ties that are stronger than randomly expected, is examined. By shifting attention from the binary structure to tie strength, this method thus extends previous research on the rich-club phenomenon, and provides a general framework for detecting non-trivial patterns of interaction among the prominent nodes of a weighted network.

3.1 The topological rich-club effect

This work draws on, and extends, the topological measure of the rich-club phenomenon (Colizza et al., 2006; Zhou and Mondragon, 2004), that quantifies the

extent to which the nodes with a high degree (i.e., the number of ties originating from a node, k) are connected with each other to a greater extent than randomly expected. In this measure, the prominent nodes are defined as the hubs that preside over many ties with other nodes. This choice was based on the discovery of skewed degree distributions (i.e., the probability that a given tie has degree k, $P(k)$) in a host of networks (Barabási and Albert, 1999; Dorogovtsev and Mendes, 2003, pg. 80-81).

Formally, the topological rich-club coefficient is the proportion of ties connecting prominent nodes, with respect to the maximum possible number of ties among them. For the set of prominent nodes with degree larger than k, $N_{>k}$, the coefficient is defined for undirected networks as (Zhou and Mondragon, 2004):

$$\phi(k) = \frac{2A_{>k}}{N_{>k}(N_{>k} - 1)} \tag{6}$$

where $A_{>k}$ represents the number of ties connecting the $N_{>k}$ prominent nodes. This equation is not enough to test whether highly connected nodes are connected with each other to a greater extent than randomly expected. Since the highly connected nodes have relatively many ties compared with the other nodes in the network, the likelihood that a tie is randomly located between them is higher than the likelihood of a tie in the overall network (Colizza et al., 2006). Therefore, in order to detect the non-random tendency towards the generation of rich-club structures, $\phi(k)$ measured on the observed network must be compared with the corresponding $\phi_{\text{null}}(k)$ obtained from an appropriate null model. The null model is typically used as a benchmark to assess whether a property measured in a real-world network deviates from what would be observed by chance (Amaral and Guimerà, 2006). For the topolgical rich-club measure, Colizza et al. (2006) proposed the following ratio:

$$\rho(k) = \frac{\phi(k)}{\phi_{\text{null}}(k)} \tag{7}$$

This ratio enables us to examine the extent to which the observed rich-club phenomenon diverges from what would be expected by chance.

A positive topological rich-club phenomenon has been found in networks of scientific collaborations among researchers (Colizza et al., 2006), in transportation networks (Colizza et al., 2006; Opsahl et al., 2008), in the Italian interbank network (De Masi et al., 2006), and in content-based networks (Balcan and Erzan, 2007). On the contrary, a negative tendency was found for the Internet, where highly connected routers are not typically connected with one another (Colizza et al., 2006). Biological networks, such as protein-protein interaction networks, do not show a consistent trend. Studies suggests that the trends are related to specific features of the organisms under study (Colizza et al., 2006; Guimerà et al., 2007; Wuchty, 2007).

3.2 The weighted rich-club effect

While Eq. 6 describes whether or not ties are established among prominent nodes, it does not measure the relative strength of these ties with respect to other ties in the network. Examining the intensity and capacity of interactions is however fundamental for understanding the organising principles underpinning the structure of weighted networks.

Moreover, the prominence of a node can be defined in terms not only of its degree, but also of other properties in weighted networks. This can include the strength of the nodes (i.e., the sum of the weights attached to the ties originating from the nodes; Zlatic et al., 2008) or the average weight (i.e., the ratio between the strength and degree of the node). To determine the relative strength of the ties connecting prominent nodes, we propose the following weighted rich-club coefficient, based on

a parameter r of node prominence:

$$\phi^w(r) = \frac{W_{>r}}{\sum_{l=1}^{A_{>r}} w_l^{\text{rank}}} \tag{8}$$

where the numerator is the total weight of the ties connecting the nodes that are prominent with respect to r. Given that the number of ties among the prominent nodes is $A_{>r}$, the denominator corresponds to the sum of the weights of the $A_{>r}$ strongest ties of the network. w_l^{rank} is an ordered vector of all the weights in the network. This vector is ordered accordingly to $w_l^{\text{rank}} \geq w_{l+1}^{\text{rank}}$ and $l = 1, 2, \ldots, A$, with A being the total number of ties in the network. Thus, Eq. 8 measures the fraction of weights shared by the prominent nodes compared with the total amount they could share if they were connected through the strongest ties of the network. $\phi^w(r)$ takes values ranging from 0 to 1. It is equal to 0 if none of the ties connecting the prominent nodes are among the strongest ones, whereas it is equal to 1 when the ties connecting the prominent nodes are the strongest available ones.

To illustrate the different elements of this coefficient, Figure 4 shows a sample network with 38 nodes out which 5 are designated as prominent ones. The prominent nodes are connected by 6 *internal* ties (highlighted ties in Figure 4a). Not all these 6 ties are among the 6 *strongest* ties in the network (highlighted ties in Figure 4b). The coefficient for this network would be the the sum of the weights attached to the *internal* ties (panel a) divided by the sum of the weights attached to the *strongest* ties (panel b).

3.2.1 Null models

In analogy with the topological rich-club coefficient, Eq. 8 might not enable us to test whether there is an actual tendency of the prominent nodes to be connected through the strongest ties in the network. This is due to the fact that some ordering

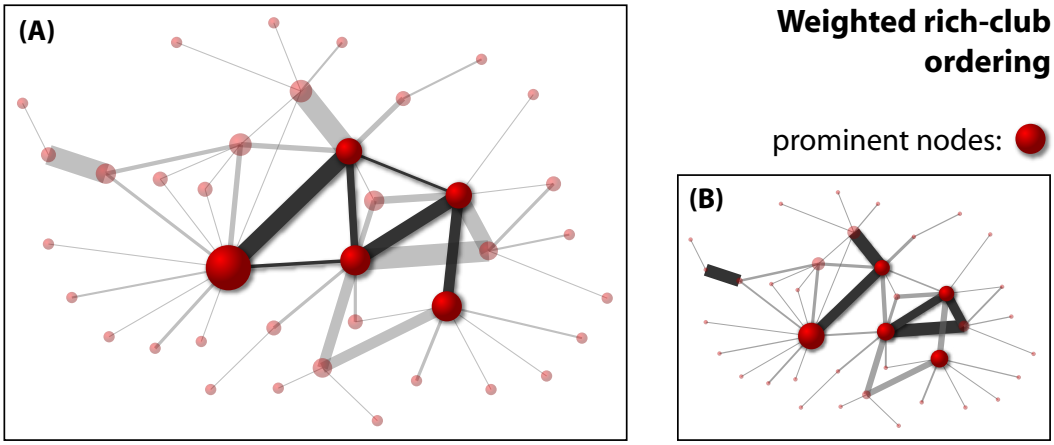


Figure 4: Weighted rich-club ordering. Schematic representation of a weighted network with 36 nodes. The size of nodes is proportional to their prominence, and width of ties to their weight. (A) The prominent nodes and the ties among them are highlighted. (B) The $A_{>r}$ strongest ties of the network are highlighted, where $A_{>r}$ is the number of ties among the prominent nodes.

properties are associated with the strength of ties. When this is the case, even random networks may display a signal. Therefore, to properly test the phenomenon, we need to assess the weighted rich-club effect observed in a real-world network against the effect found in an ensemble of random networks based on an appropriate null model. This model must generate networks that are random, but at the same time comparable to the observed network. In particular, our choice of an appropriate null model reflects the need to discount for associations between weights and ties. To this end, the ensemble of random networks produced by the null model must meet three main requirements. First, the random networks must have the same number of nodes and ties as the observed network. This ensures that the basic parameters of the networks are the same (Erdős and Rényi, 1960; Rapaport, 1953). Second, they must have the same weight distribution $P(w)$ (i.e., the probability that a given tie has weight w) as the observed network. This is a crucial constraint since we are looking for non-trivial intensity of interactions among nodes. Moreover, this guarantees that the vector of ordered weights remain the same. Third, the nodes in the selected

subset (the club) must be the same as in the observed network. This preserves the distribution $P(r)$ (i.e., the probability that a given node has prominence r). A null model that does not produce an ensemble of random networks that fulfill the above three requirements cannot produce networks comparable to the observed network, and thus does not allow for a proper weighted rich-club assessment (Colizza et al., 2006).

There are multiple null models that create random networks that meet the above conditions. Nonetheless, at the same time certain models are excluded. This is the case, for example, of models in which weighted ties are considered to be multiple binary ties (e.g., a tie with a weight of 3 is considered to be 3 binary ties; Newman, 2004a; Serrano, 2008). These models do not preserve the number of ties or the weight distribution $P(w)$ of the observed network.

In what follows, we introduce three null models for constructing random networks. The appropriateness of these models depends on the choice of the prominence parameter r. If the prominence of a node is given by its degree, we adopt the following two null models. A first procedure consists simply in reshuffling the weights globally in the network (Weight reshuffle null model). This null model maintains the topology of the observed network. Therefore, the number of ties originating from a node (degree) does not change.

A second procedure, which introduces a higher degree of randomisation, consists in reshuffling also the topology, reaching the maximally random network with the same degree distribution $P(k)$ as the observed network (Maslov and Sneppen, 2002; Newman, 2003). It does so by randomly selecting two ties, $i_1 \overset{}{\underset{w_1}{\rightarrow}} j_1$ and $i_2 \overset{}{\underset{w_2}{\rightarrow}} j_2$. The two ties are then rewired by setting $i_1 \overset{}{\underset{w_1}{\rightarrow}} j_2$ and $i_2 \overset{}{\underset{w_2}{\rightarrow}} j_1$. The weights are automatically redistributed by remaining attached to the reshuffled ties. However, if either of these ties is already formed, this step is reverted, and two new ties are selected. This condition guarantees that multiple ties are not formed between

two nodes, which ensures that the weight distribution $P(w)$ and degree distribution $P(k)$ remain unchanged. If this procedure is repeated enough times, the outcome is a corresponding random network (Weight & Tie reshuffle null model).[13]

While both randomisation procedures preserve $P(k)$ and $P(w)$ of an observed network that can be either directed or undirected, they differ in that the Weight & Tie reshuffle alters the location of the ties, and thereby destroys node-node topological correlations. Therefore, a rich-club coefficient based on the latter null model will mix the effect coming from the location of the strongest ties and that coming from the topology. We consider it here for the sake of comparison, since it is the method used to calculate the topological rich-club effect (Colizza et al., 2006), and also to check the effect of higher degrees of randomisation on the obtained results.

Inevitably, since weights are reshuffled globally, both procedures produce random networks in which the nodes do not maintain the same strength s as in the observed network. Therefore, when prominence is based on node strength, we need to introduce a third randomisation procedure that preserves this quantity. We construct a null model based on the randomisation of directed networks (Serrano et al., 2007) that preserves not only the topology and $P(w)$, but also the strength distribution $P(s)$ (i.e., the probability that a given node has strength s). To this end, we reshuffle weights locally for each node across its outgoing ties (Directed Weight reshuffle null model). In so doing, we also obtain null models where the average weight of outgoing ties is kept invariant. We extend this procedure to undirected networks by duplicating an undirected tie into two directed ties – one in each direction. It should be noted that this procedure breaks the weight symmetry in the two directions of an undirected tie (the topology remains invariant). The appropriateness of this method for undirected networks depends on the research setting and how

[13]Since this model is commonly used in the Physics literature, we apply it here. However, each of the random networks that can be produced with this model is not produced with an equal probability of the null model. For more details, see Snijders (2001) and Rao et al. (1996).

tie weights are defined. For example, its applicability to undirected transportation networks is justified by the typically directed nature of traffic flows (although the US airport network displays a high symmetry; Barrat et al., 2004). Conversely, in an undirected collaboration network this might not be appropriate. In particular, for projections of two-mode networks, it might be more appropriate to reshuffle the two-mode structure before projecting it onto a one-mode network (see Rao et al., 1996; Snijders, 1991). Nevertheless, we choose to apply this method due to the lack of better methods and a procedure which maintains $P(s)$ and the weight symmetry would constrain the produced random networks to an extent that they would differ from the observed network only slightly. Figure 5 shows a schematic representation of the three methods.

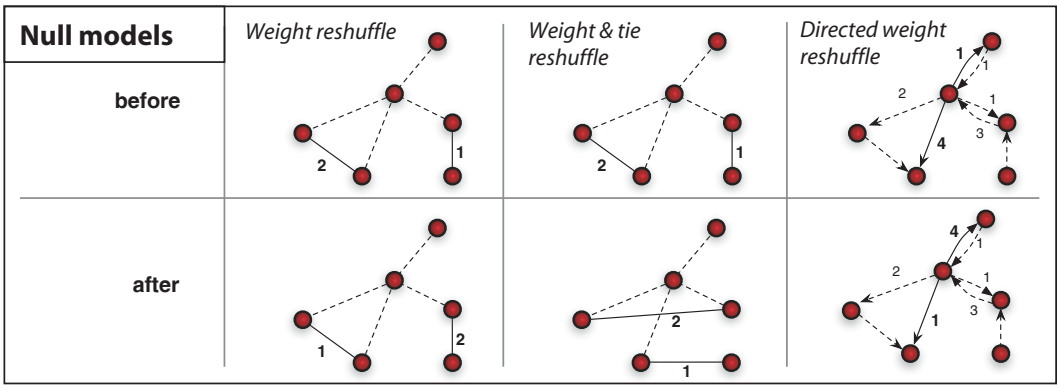


Figure 5: Randomisation procedures: Representation of a single reshuffle. Solid lines correspond to the ties that are randomised; the number attached to the ties refers to their weight. The Weight reshuffle procedure randomises the weights globally in the network, while keeping the topology intact. The Weight & Tie reshuffle procedure introduces a higher degree of randomisation by reshuffling also the topology, while preserving the degree of each node (Molloy and Reed, 1995; Newman and Park, 2003). Weights are automatically redistributed as they remain attached to the reshuffled ties. The Directed Weight reshuffle procedure reshuffles weights locally for each node across the ties originating from the node (Serrano et al., 2007). This figure is based on the third part of Figure 1 in Opsahl et al. (2008).

As with the topological rich-club coefficient, we define the normalised weighted rich-club coefficient as the ratio between the value obtained by Eq. 8 measured

on the observed network and the value obtained on an ensemble of corresponding random networks:

$$\rho^w(r) = \frac{\phi^w(r)}{\phi^w_{\text{null}}(r)}. \tag{9}$$

When ρ^w is larger than one, the observed network has a positive weighted rich-club ordering, with prominent nodes concentrating a disproportionately large part of their efforts towards other prominent nodes compared with what happens in the random null model. Conversely, if $\rho^w(r)$ is smaller than one, the ties among the prominent nodes are weaker than randomly expected.

3.2.2 Significance of effect

The randomly expected value, $\phi^w_{\text{null}}(r)$, is obtained by taking the average of $\phi^w(r)$ measured on many random networks created using an appropriate null model. Even though the random networks are based on the same null model, the values that constitute the average varies. This is because each sampled random network is different from each other. We have found that when few nodes and ties exist within the subset of prominent nodes, the values found for the random networks can differ considerably. In fact, a striking outcome of $\rho^w(r)$ might be reproduced in a large proportion of the random networks when the definition of prominence is very restrictive.

Here we analyse the variation in the values of $\phi^w_{\text{null}}(r)$. These values can be plotted as a distribution that shows the frequency of their occurrence. Figure 6 shows the distribution of $\phi^w_{\text{null}}(r)$ obtained from 1,000 random networks created from the online social network (see Section 1.4) using the Weight reshuffling.

The values of $p(\phi^w_{\text{null}}(r))$ can also been fitted by a probability density function. By analysing a number of the empirical networks outlined in Chapter 1, we have found that the distribution of $p(\phi^w_{\text{null}}(r))$ roughly follows a symmetric Gaussian function. If the value of $\phi^w(r)$ found in an observed network is rarely replicated in the networks generated by the null model, we argue that a statistically significant weighted rich-

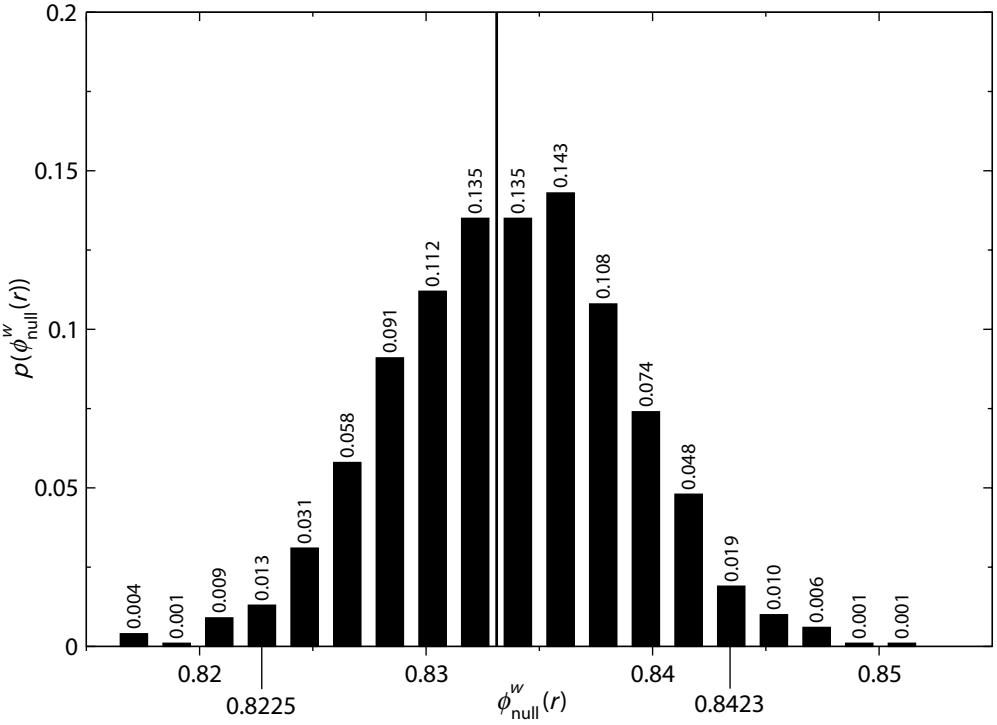


Figure 6: Distribution of $\phi^w_{\text{null}}(r)$. $p(\phi^w_{\text{null}}(r))$ is the fraction of random networks with a given $\phi^w_{\text{null}}(r)$. The thin vertical line represents the average value that would be used in Eq. 9.

club effect is present. A negative or a positive and significant effect is present if the value found for the the observed network is, respectively, lower or higher than the values found for the vast majority of random networks. More specifically, we argue that a significant weighted rich-club effect exists if the observed $\phi^w(r)$ is outside the 95% confidence interval of $\phi^w_{\text{null}}(r)$ measured on the random networks. For the distribution shown in Figure 6, a negative and significant weighted rich-club effect is found if $\phi^w(r)$ measured on the observed network is lower than 0.8225, and a positive and significant effect is found if $\phi^w(r)$ is higher than 0.8423.

3.3 Empirical tests

We apply the above framework to three of the real-world networks described in Chapter 1. The networks are: the US airport network, the scientific collaboration

network (Newman, 2001b,c), and the the online social network. We choose these networks due to their size. The remaining social networks had less than 100 nodes. We deem these networks too small.

3.3.1 Club of the most connected nodes

In analogy with the topological rich-club coefficient, we begin by defining the prominence parameter r as the degree of nodes. In so doing, we assess whether there is a tendency of highly connected nodes to forge stronger ties among one another than would be the case if weights were randomly attached to ties. We use the Weight reshuffle null model, since it is the simplest null model that preserves the prominence of nodes, i.e. their degree. We also consider the Weight & Tie reshuffle null model to show how an increased level of randomness affects the results.

Figure 7 reports the weighted and topological (inset) rich-club ratios for the three networks. The diagrams on the left show results based on the Weight reshuffle null model, whereas the ones on the right show results based on the Weight & Tie reshuffle null model[14]. The airport network shows a positive weighted rich-club ordering, as can be identified from the remarkable growth of both ρ^w when the subsets of high degree nodes become increasingly restrictive. Moreover, a mild topological effect is found. This is in agreement with previous studies that found correlations between weight of the ties and degrees of the nodes (Barrat et al., 2004; Guimerà et al., 2005; Wu et al., 2006). Routes among hub airports, with flights to many destinations, are the busiest ones in the U.S.

Conversely, while the scientific collaboration network has a strong positive topological rich-club effect, it does not exhibit any weighted rich-club ordering. This

[14]The Directed Weight reshuffle null model also maintains the degree distribution of the observed network. However, a random networks created using this null model is less different from the observed network than random networks created using the other two null models. In particular, if a node is only connected to one other node, the tie cannot be randomised. Nevertheless, similar results are found when this null model is used (see Appendix B.1).

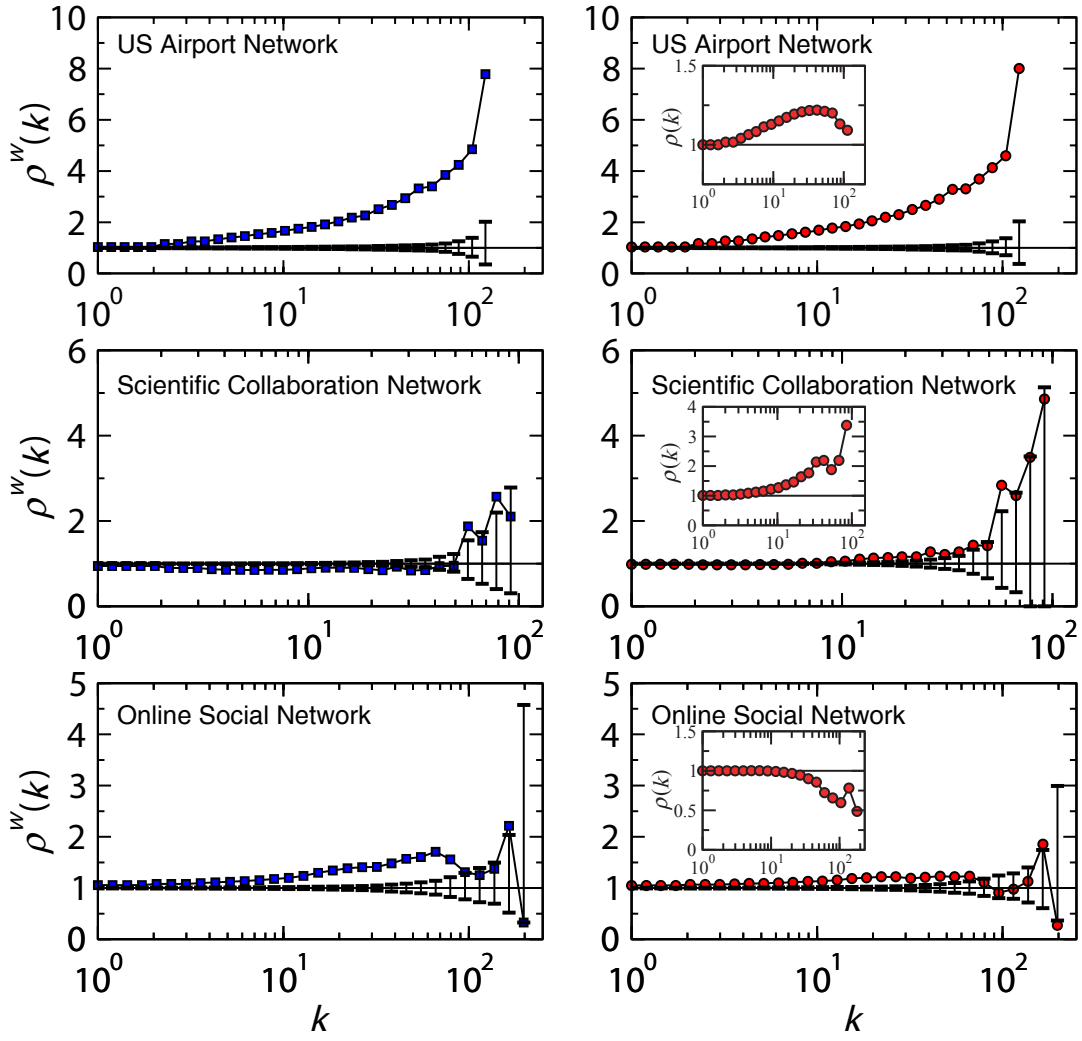


Figure 7: Weighted rich-club ordering among the most connected nodes in: the US airport network (top row); the scientific collaboration network (middle row); and the online social network (bottom row). The diagrams on in the left panel show results based on the Weight reshuffle null model and the diagrams in the right panel show results based on the Weight & Tie reshuffle null model. The error bars in the main diagrams refer to the 95% confidence intervals of $\phi^w_{\text{null}}(k)$. The boundaries of the intervals are divided by the average $\phi^w_{\text{null}}(k)$ in a similar fashion as $\phi^w(k)$ found in the observed networks (Eq. 9). Inset refers to the topological rich-club ordering.

suggests that the authors who collaborate with many others tend to collaborate among themselves. However, their collaborations are not stronger than randomly expected. As shown in the second row of Figure 7, the coefficient for the observed network is close to the value found in the random network: ρ^w remains flat around 1 for a large range of values of k. A substantial departure of the coefficient from the expected value is found only for very high values of k, where only 29 authors are classified as prominent. However, at this level, the observed coefficient does not depart from the random one in a statistical significant way. These results are in agreement with previous studies that showed that the strong ties in collaboration networks tend to be independent of the degrees of the nodes (Ramasco and Gonçalves, 2007; Ramasco, 2007).

Finally, the topological and weighted rich-club coefficients display strikingly different trends from each other for the online social network. While the topological coefficient decreases with k and remains below 1 throughout the whole range of degrees, the weighted coefficient shows a mild increasing trend. Very gregarious individuals, namely the ones that contact a large number of other users, poorly communicate with one another. However, when they do, they choose to forge ties that are stronger than randomly expected. When there are few nodes left in the network ($k \geq 180$), the coefficient fluctuates. Moreover, the observed coefficient does not significantly deviate from the random one in this range of k.

The limitations of defining prominence in terms of degree and the advantages of considering other ordering properties are illustrated by using the example of the scientific collaboration network. In this network, each paper is translated into a fully connected group of collaborators (or cliques). Therefore, the whole network can be represented as a set of cliques that overlap when authors write papers with different others. When a paper is co-written by a large number of authors, these authors take on a high degree and thus increase their chances to become classified

as prominent. However, due to the operationalisation of tie strength, the weight of these ties are weak (Newman, 2001c). For example, if 101 authors write only a single paper together, they would all receive a degree of 100, but the ties among them would only have a weight of 0.01. Thus, large collaborations tend to secure the prominent status for the authors, yet generate weaker ties among the prominent nodes than smaller collaborations.

To illustrate this issue, we focus on a subset of the scientific collaboration network that includes only authors working on network theory and experiments (*network science* network; Newman, 2006). This network displays similar weighted and topological rich-club effects as the observed overall network (see Appendix B.2). Experimental papers on biological networks are written by a large number of authors, and therefore only one of these papers may suffice to substantially increase the topological rich-club ordering. Figure 8a and b show all nodes with $k \geq 10$ and $s \geq 5$, respectively, and the ties among them, in the *network science* network. The large clique in Figure 8a consists of 20 authors that are tied together by a single paper (Uetz et al., 2000). Only 3 of these authors are also tied together by an additional paper. The ties among the other 17 authors who only collaborated on the single publication have a weight of 0.05263, which is the minimum tie weight in the network. Therefore, the existence of large collaborations increases the number of weak ties in the numerator of the coefficient, which reduces the weighted rich-club effect. This can offset a possible weighted rich-club effect among the other prominent nodes. Therefore, another ordering property that does not classify the authors of a single large collaboration as prominent is needed. As can be seen in Figure 8b, if prominence is based on the strength of the nodes (which is the number of co-authored papers published by authors), the subset of nodes and, more importantly, the weight of the ties among them, change substantially.

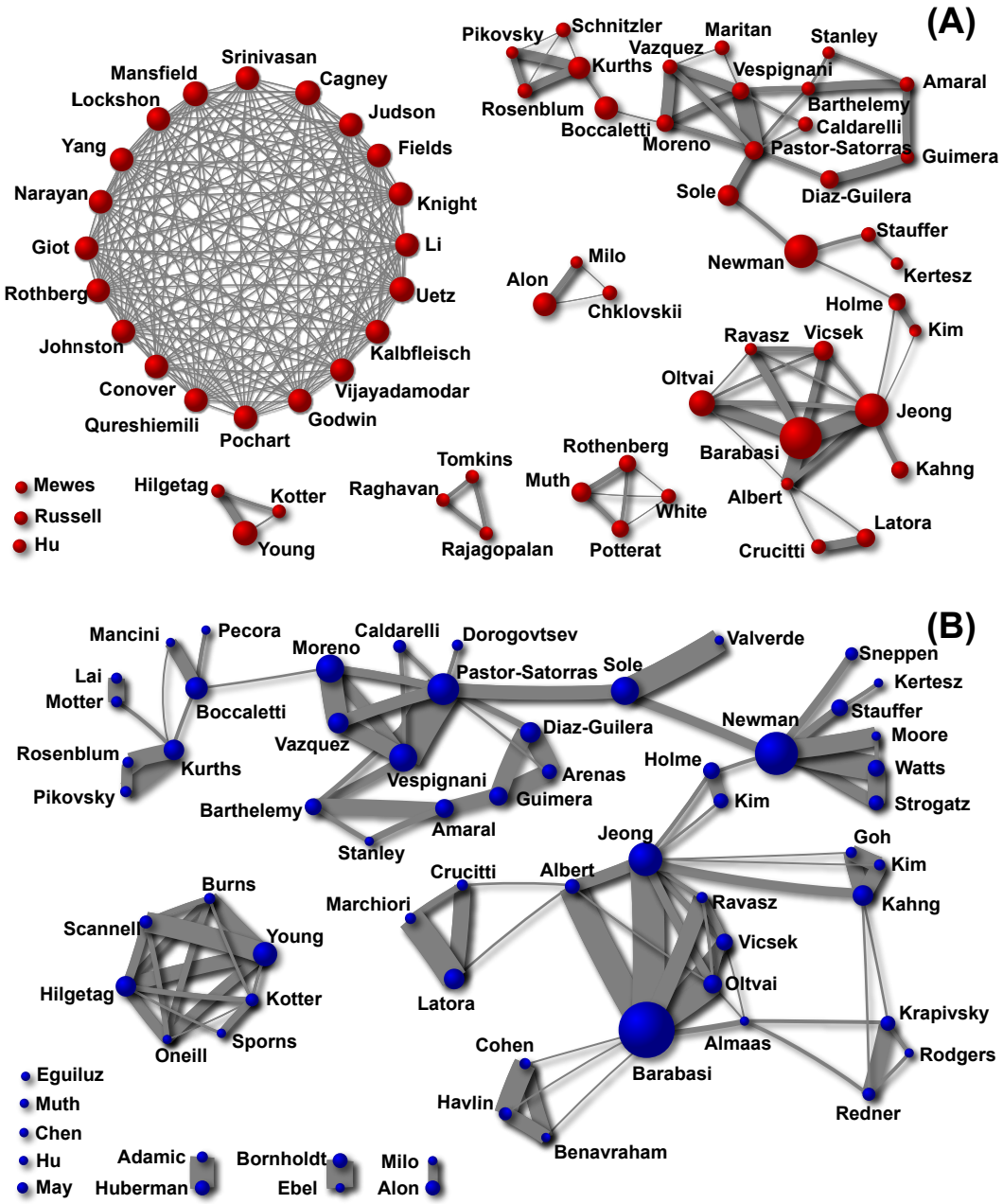


Figure 8: Subset of the prominent nodes in the network science collaboration network (Newman, 2006): A) based on degree ($k \geq 10$) and B) based on strength ($s \geq 5$). Only ties among the prominent nodes are shown. The size of the nodes corresponds to the prominence parameter – degree (A) or strength (B). The width of each tie is proportional to its weight. This figure is based on Figure 3 in Opsahl et al. (2008).

3.3.2 Club of the most active nodes

In light of this observation, the next step is thus to define prominence r in terms of node strength. In so doing, we shift our attention from the most connected to the most involved nodes in the network activity. The weighted rich-club assessment then measures whether these nodes direct their strongest ties preferentially towards each other. To ensure that the prominent nodes in the null model remain the same as in the observed network, we need to preserve $P(s)$ in addition to $P(w)$. Therefore, we adopt the Directed Weight reshuffle null model that also preserves this distribution. It is worth noting that the construction of this null model for the undirected scientific collaboration network is a methodological extension of the original procedure. As suggested in Section 3.2.1, a more appropriate null model might be reshuffling the two-mode structure instead the one-mode projection. However, only the one-mode structure of this network was available to us.

Figure 9 shows a positive weighted rich-club ordering for all the three networks analysed. Highly involved nodes preferentially direct their strongest ties towards one another, and this tendency becomes more pronounced as the number of prominent nodes decreases. The airport network exhibits a strong weighted rich-club effect. This result suggests that traffic is heavier among busy airports than randomly expected. Moreover, it corroborates the result obtained when prominence was defined in terms of k. This finding is not surprising given the previous result as node degree and strength are correlated in this network (Barrat et al., 2004). More generally, if node degree and strength are correlated, the subsets of prominent nodes obtained with the two definitions are likely to be composed of the same nodes. If this is the case, then the results will not differ.

Defining prominence in terms of strength is especially relevant for the scientific collaboration network, where tie strength is equal to the number of co-authored papers published by each author, and can therefore be seen as a measure of pro-

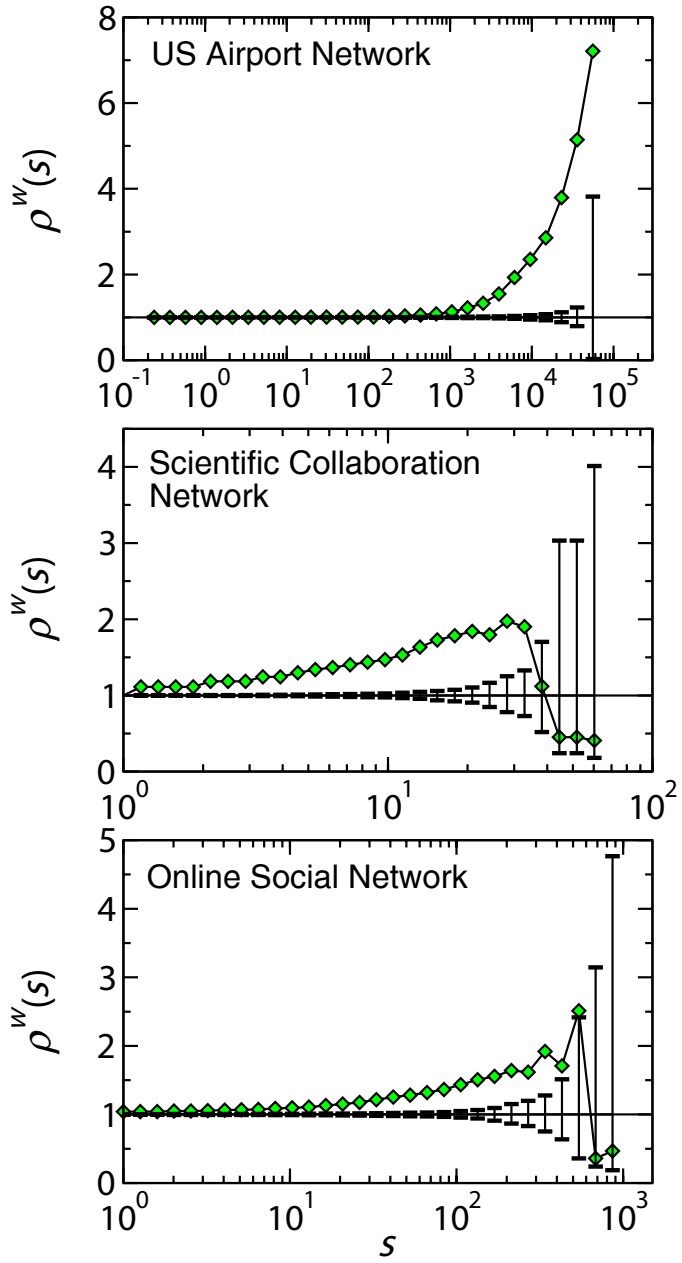


Figure 9: Weighted rich-club ordering among actively involved nodes in: the US Airport Network (top); the Scientific Collaboration Network (middle); and the Online Social Network (bottom). The error bars in the main diagrams refer to the 95% confidence intervals of $\phi^w_{\text{null}}(s)$. The boundaries of the intervals are divided by the average $\phi^w_{\text{null}}(s)$ in a similar fashion as $\phi^w(s)$ found in the observed networks (Eq. 9).

ductivity. In this case, the weighted rich-club ordering is positive among authors that published many papers, unlike what was found when prominence was defined in terms of number of collaborators. This signals that strength is a better parameter than degree for identifying a subset of nodes with stronger ties among themselves than randomly expected.

The online social network also reveals a pronounced positive weighted rich-club ordering, thus suggesting that active online users tend to communicate frequently with one another. This corroborates the findings when the definition of prominence was based on degree. A possible cause of similar results is correlation between node degree and strength. In fact, in this network, the pair-wise correlation between these two properties is 0.90.

For very high values of s, only few nodes are part of the subsets of prominent nodes in the three networks. This implies a high level of fluctuations in ρ^w for high values of s. It can be speculated that the drop observed in the social networks may be due to some form of competition among very prominent nodes. This might account for the reluctance of prominent authors to establish strong ties among themselves, as is suggested by the lack of interaction among the three most productive authors in Figure 8B: Barabási, Newman, and Vespignani.

3.3.3 Club of the nodes with the highest average weight

While node strength gives is a proxy of a node's involvement in the network activity, it does not distinguish between nodes with a large number of weak ties and nodes with a small number of strong ties, given the same value of node strength. In addition, due to high correlation between node degree and strength in the networks, the two ordering properties are likely produce similar subsets of prominent nodes. To address these issues, we define the prominence parameter r in terms of the average weight $\bar{w}$ (Ramasco and Morris, 2006). A positive coefficient suggests that

the nodes with strong ties choose to direct these with each other. We use the Directed Weight reshuffle null model to keep invariant $\bar{w}$ for each node, thus ensuring that the prominence of the nodes in the observed network and in its corresponding randomised version remain the same.

Figure 10 shows the weighted rich-club coefficient for all the three networks. The airport network displays a positive signal which substantially departs from the random baseline only at high values of the average weight. Airports characterised, on average, by very busy routes tend to direct these routes to one another.

Positive signals are also found for the scientific collaboration network and the online social network. In the collaboration network, authors that show the ability to commit themselves to their collaborators tend to forge strong ties among one another. In the online social network, strong bonds link online users that are capable of developing strong relationships.

3.4 Contribution to the literature

This project contributes to the literature by providing a method for assessing the extent to which prominent nodes direct their efforts towards each other. This is a feature of complex networks that was not detected previously. This project lies within the context of the rich-club perspective, which studies the properties of a subset of prominent or rich nodes. Much emphasis has been placed on the extent to which highly connected nodes of a network interact with one another, and to a lesser extent on the nature and strength of their interactions. The explanatory power of the rich-club perspective has so far been hindered by its assumption that only one class of nodes, namely the highly connected ones, are likely to play a crucial role in a network. In this chapter, we relaxed this assumption, and developed a novel framework for examining the tendency of prominent nodes to attract and secure control of resources. By exploring different definitions of prominence in a

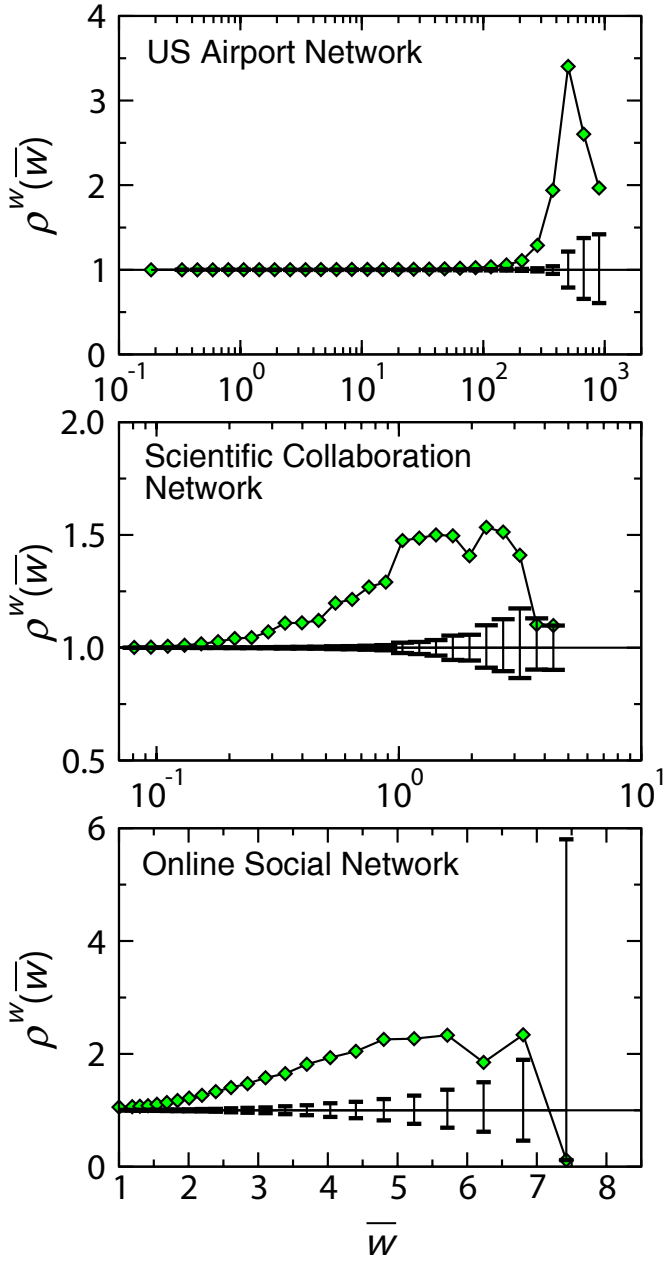


Figure 10: Weighted rich-club ordering among the nodes with the highest average weight in: the US Airport Network (top); the Scientific Collaboration Network (middle); and the Online Social Network (bottom). The error bars in the main diagrams refer to the 95% confidence intervals of $\phi^w_{\text{null}}(\bar{w})$. The boundaries of the intervals are divided by the average $\phi^w_{\text{null}}(\bar{w})$ in a similar fashion as $\phi^w(\bar{w})$ found in the observed networks (Eq. 9).

network, we are able to identify novel organising principles that help us deepen our understanding of the network's organisation and functioning.

The wide applicability of our method to a variety of empirical settings makes it a particularly suitable methodological tool for scientific investigations across multiple disciplinary fields. Results uncover mechanisms governing the distribution of air travel, the selection and intensity of scientific collaborations, and the way people commit themselves to one another in a virtual setting. The implications of our analysis for these three domains only represent a small fraction of the potential of the proposed method. In fact, this method can be applied to network data from a variety of fields, including biological sciences, social and political sciences, computer and information sciences, environmental sciences, spatial economy, and ecology, among others.

3.5 Conclusion and discussion

Previous research has shown the importance of the nature of ties in a variety of network-based processes (Barrat et al., 2004; Coleman, 1988; Granovetter, 1973; Hansen, 1999; Krackhardt, 1992; Pastor-Satorras and Vespignani, 2004; Simmel, 1950; Uzzi, 1997). Drawing on measures for detecting topological rich-club orderings, we proposed a new general framework for the study of patterns of interactions among selected nodes in weighted networks. We tested this framework on three networks from the domains of transportation, scientific collaboration, and online communication. We extracted increasingly restrictive subsets of prominent nodes based on their degree, strength and average weight. Then, for each of these subsets, we examined whether the prominent nodes were more prone to direct their efforts towards one another than would be expected if the targets of these efforts were chosen randomly. Our results show that two of the networks analysed display non-trivial weighted rich-club effects among highly connected nodes. Conversely, when

subsets were based on node strength and average weight, we found that prominent nodes tend to forge stronger ties with one another than randomly expected in all the networks.

The proposed method is widely applicable as it allows for an assessment of the control benefits of any subset of nodes. Prominence does not necessarily originate from network properties, such as node degree, strength, or average weight. To the extent that the nodes of a network can be ordered accordingly to a given property, our framework suggests several new ideas for future research. For example, how do performance, centrality, status, age, and size impact on the ability of nodes to control the strongest ties in a network? By providing insights into how prominent nodes, selected accordingly to different parameters, choose to direct their efforts towards one another, the proposed method represents a step towards furthering our understanding of the global organisation of complex networks.

4 Evolution of Networks[15]

Networks evolve as a result of the joining and leaving of nodes, and the creating, reinforcing, weakening, and severing of ties. For example, the online social networks created from the virtual community grew when people joined the community and shrunk when people left it. If ties are defined in terms of online messages, they are formed when a person sends a message to another for the first time, and reinforced if multiple messages are subsequently sent (Panzarasa et al., 2009). Since, online interaction datasets generally do not contain information about the severing or weakening of ties, it is often assumed that social relationships are severed if no is sent during a certain amount of time (Kossinets and Watts, 2006). Similarly, a weakening is assumed to occur if the rate of messages exchanged decreases.

Up until recently there have been only a few network datasets where the exact evolution of the network could be mapped (e.g. Hall et al., 2001; Holme et al., 2004; Kossinets and Watts, 2006; Onnela et al., 2007; Opsahl and Panzarasa, 2008; Panzarasa et al., 2009). Thus, a substantial part of the empirical work on social networks has been conducted on cross-sectional or static networks (e.g. Bernard et al., 1988; Fararo and Sunshine, 1964; Foster et al., 1963; Gouldner, 1960; Katz and Proctor, 1959; Lazarsfeld and Merton, 1954; McPherson et al., 2001). In these studies, a host of measures have been proposed to detect features of the network structure. Based on the existence of a particular feature, it has been speculated that certain mechanisms have underpinned tie creation. For example, the clustering coefficient (for a review, see Chapter 2) measures the extent to which triangles occur in a network. The coefficient for an observed network can be compared to the one found in a corresponding random network (Erdős and Rényi, 1959; Newman, 2003; Panzarasa et al., 2009; Solomonoff and Rapoport, 1951). If it is higher than the ran-

[15]We thank Tom Snijders for helpful comments that directed us towards developing the specific model used in this chapter.

domly expected one, then scholars have often concluded that there is a mechanism that increases the likelihood of forming a tie between two nodes if they have ties to the same other node (Heider, 1946; Holland and Leinhardt, 1971).

These measures fail to assess multiple effects that may influence the decision-making process of the nodes in a network. By using a conditional logistic regression framework, Powell et al. (2005) studied the effects of different mechanisms on tie generation among organisations. However, the framework that they developed was not general as their network was not a prototypical social network. In particular, they combined a one-mode network (i.e., one set of nodes and ties among those nodes) with a two-mode network (i.e., two sets of nodes with ties only between nodes in the different sets). Moreover, the network was only recorded at yearly intervals and, therefore, they could not account for the dependence among ties occurring in the same year. In addition, the network was undirected. This implies that the decision to form a tie does not rest with one node, but with both the nodes which are connected by the tie. In this Chapter, we develop a general and flexible framework for one-mode networks. We focus on the individual choices that a single node makes when creating a tie, and thereby we concentrate on directed networks. This enables us to assess mechanisms that cannot be tested in undirected networks, such as reciprocity.

The rest of this Chapter is organised as follows. First, we highlight a number of mechanisms thought to guide tie generation in social networks. In Section 4.2 we describe methods used to study combinations of these mechanisms in binary cross-sectional networks. We then turn our attention towards longitudinal network data, where the exact sequence of ties is known, and propose to use an established regression framework to examine the decisions made by individual nodes at the local level to initiate a social tie (binary analysis). We empirically test the proposed method on the online social network outlined in Chapter 1. We are able to use this

network as each message (or tie) was recorded with the exact time at which it was created. This allows us to extract the exact sequence in which nodes and ties were added to the network. In Section 4.4 we extend our analysis to weighted networks and conduct a second assessment of the online social network by also taking into consideration messages used to reinforce existing social ties. The following section tests the sensitivity to a methodological and computational constraint. Finally, we highlight the contribution to the literature and offer a critical assessment of the main results.

4.1 Network growth mechanisms

A number of mechanisms have been thought to affect nodes' decision-making. First, as mentioned in Chapter 2, there tend to be many more triangles of nodes in networks than one would expect by chance (Davis, 1970; Erdős and Rényi, 1960; Heider, 1946; Rapaport, 1953). This might be the result of individuals' desire to maintain balance among ties with others ("my friends' friends are my friends"; Hallinan, 1974; Heider, 1946). Another explanation refers to third-part referral (Davis, 1970; Holland and Leinhardt, 1971). A person's social circles are likely to overlap through social occasions, during which, a person might introduce people from the different circles to each other. In the literature, the mechanism underpinning the creation of triangles is known as triadic closure (Burt, 2005; Heider, 1946; Holland and Leinhardt, 1971).

Second, a number of empirical studies have shown that ties are not equally distributed across nodes (Barabási and Albert, 1999; Barabási et al., 2002; Dorogovtsev and Mendes, 2003; Jeong et al., 2003). The majority of nodes attract only few ties, whereas there are a select minority of nodes that receive a disproportionately large number of ties (Barabási and Albert, 1999). A possible explanation for this type of skewed distribution is that *popularity is attractive* (Dorogovtsev and Mendes, 2003). In other words, a virtuous effect occurs for the "popular" nodes whereby they re-

ceive relatively more ties than others as the network evolves. This mechanism has been independently rediscovered several times in different areas of research. Simon (1955) called this mechanism the "Gibrat principle" after French economist Robert Gibrat (1904-1980). Gibrat argued that the proportional change in the firm size is the same for all firms in an industry. This implies that if one firm doubles its size in 10 years, all other firms will also double in size. Thus, the biggest firms will become bigger and bigger relative to all the others. Merton (1968) referred to this concept as the "Matthew effect", after the first part of the biblical edict, "For everyone who has will be given more, and he will have an abundance. Whoever does not have, even what he has will be taken from him." (Matthew, 25:29). More recently, Barabási and Albert (1999) coined the term "preferential attachment", which states that nodes preferential "attach" or create ties with high degree nodes. In particular, they showed that a high number of hyper-links on the Internet point towards a small number of pages.

Third, in a directed network, two directed ties can exist between two nodes in a dyad – one in each direction. It has been found that in most networks, there are more dyads with two directed ties than randomly expected (Gouldner, 1960; Holland and Leinhardt, 1981; Plickert et al., 2007; Wasserman and Faust, 1994). In fact, the vast majority of dyads in the airport network studied in Chapter 3 have either two or no directed ties. Extremely few dyads consist of a single directed tie (Barrat et al., 2004; Guimerà et al., 2005). The mechanism that increases the likelihood of creating a second directed tie in a dyad is referred to as reciprocity (Gouldner, 1960; Holland and Leinhardt, 1981; Plickert et al., 2007).

Fourth, socially similar people have been found to create ties with each other to a greater extent than randomly expected (Lazarsfeld and Merton, 1954; McPherson et al., 2001). It has been argued that this is due to the fact that similarity generates a baseline level of interpersonal attraction (McPherson et al., 2001). In addition, it

has been shown that similarity bring about more stable and stronger ties between people than randomly expected (Hallinan and Kubitschek, 1988; Hinds et al., 2000; Reagans and McEvily, 2003; Lazarsfeld and Merton, 1954). Moreover, individuals are likely to participate in joint activities with others who have similar interests as they receive validation of their attitudes and beliefs (Aboud and Mendelson, 1996). The mechanism that is responsible for the generation of ties among similar nodes has been named homophily.

Fifth, another mechanism that is thought to affect tie generation is focus constraint (Feld, 1981; Kalmijn and Flap, 2001). This mechanism is defined as the increased likelihood of a tie being present among people that share activities, roles, social positions, and geographical location. For example, a group of people working in the same office are more likely to interact than simply a group of people in geographically distance places.

The outcomes of measures designed to test these mechanisms individually can be biased as only a single feature of the network structure is described. It could be the case that triadic closure and homophily affected the likelihood of a tie in a network. If so, an observed triangle formed among three similar people can be attributed to both triadic closure and homophily, and it would be difficult to assess whether, and the extent to which, the triangle was formed due to either of these mechanisms.

4.2 Cross-sectional binary networks

Despite the identification of these mechanisms, the literature does not provide many statistical models for appropriately analysing the combined effects of these mechanisms and the evolution of networks in general (Frank and Strauss, 1986; Robins and Pattison, 2001; Snijders, 2001; Snijders et al., 2008; Wasserman and Pattison, 1996). The lack of appropriate models stems from the fact that most networks are collected in a cross-sectional way (e.g., a snapshot of the network is collected at a

single point in time). Thus, the dependency structure among the ties is unknown. This increases the complexity of modelling efforts. For example, if a triangle is observed in a static network, it is not possible to determine which of the three ties closed the triplet formed by the other two ties.

A basic assumption often applied in statistical models is that observations are independent of each other. One model with this assumption is the logistic regression model. This is a statistical model that allows for a discrete choice or binary dependent variable (Hosmer and Lemeshow, 2000; Long and Freese, 2003). Thus, it could be applied to a binary network to predict the presence or absence of ties in a binary network, where the observations would be all possible ties. The dependent variable for the tie x_{ij} would be equal to 1 if a tie is present between node i and node j, and 0 otherwise. A number of independent variables could be identified relating to the mechanisms outlined above, such as the number of common friends. However, in this case, ties are not independent of each other as they share common nodes. For example, the three ties, x_{ij}, x_{ji}, and x_{ik}, in Figure 11 share the node i. Therefore, the estimated coefficients from a logistic regression would not be efficient, resulting in unreliable standard errors (Mood et al., 1974).

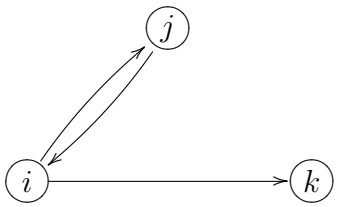


Figure 11: Example of dependence among ties. All the ties share the common node i.

Holland and Leinhardt (1981) were among the first to propose a model where ties were seen as a stochastic function of node or network properties. Their p_1-model appropriately included the dependence among the two possible directed ties in a dyad (e.g., x_{ij} and x_{ji} in Figure 11), thus allowing for more reliable estimates of reciprocity. A major shortcoming of the p_1-model was that dyads were still

considered to be independent of each other, even if they shared a common node. This meant that other effects, such as triadic closure, could not be properly assessed (see chapter 15 of Wasserman and Faust, 1994, for a criticism).

To overcome the problems of the p_1-model, Frank and Strauss (1986) applied methods from spatial statistics and statistical mechanics to networks, and developed a model that included higher-order dependency structures (Robins and Morris, 2007). This was a Markov model that assumed conditional dependencies between any two dyads that shared a common node. Thus, this model allowed for inclusion of both reciprocity and triadic closure. This model was further developed into the p^*-model or Exponential Random Graph model (ERG model; Pattison and Wasserman, 1999; Robins and Pattison, 2001; Wasserman and Pattison, 1996).

The ERG model has extensively been used to predict the combined effects of mechanisms that guide tie formation in networks (see the special issue in the journal *Social Networks* edited by Robins and Morris, 2007, for a review of developments and applications). This model takes the observed cross-sectional network structure as the dependent variable, and tries to model the factors that led to its creation[16]. To this end, it estimates coefficients for different terms that in turn create an ensemble of random networks which most closely resembles the observed one (Wasserman and Pattison, 1996).

These terms are network sub-configurations that operationalise different mechanisms thought to guide tie formation. The most common terms for directed networks are density, star-configurations, reciprocity, triadic closure, homophily, and focus constraint effects (Robins and Morris, 2007). First, the density term is a "constant" in the model that controls for the number of ties in the network. This term should be included to limit degeneracy. A degenerate model is a model that predicts a fully connected (i.e., all possible ties are formed) or empty (i.e., no ties are formed)

[16]We will focus on the ERG model for the remaining of this chapter as the p_1-model is a special case of this model, and not suited for effects such as triadic closure.

network (Handcock, 2003; Snijders, 2002). When this occurs, the coefficients of the model cannot be relied upon.

Second, a set of different star-configurations are often included to account for expansiveness and popularity mechanisms in a network (skewed degree distributions). For example, a 2-in-star configuration is the number of occurrences of two ties terminate at the same node. This term would account for popularity in a network. The star-configurations have caused a number of interpretation and modelling difficulties. For example, for undirected networks, Robins et al. (2005) found models with positive 2-star parameters and negative 3-star parameters. From this result, their "substantive interpretation" was that nodes tend to expansive (positive two-star), but experience a cost in forming too many ties (negative three-star) (p. 913). In addition, when these terms are included, the models often do not converge (i.e., stable parameters that described the network could not be found). To overcome these modelling issues, alternating k-stars were introduced by Robins et al. (2007). However, these did not improve the ease of interpretation as they are a function of positive even-numbered star configurations and negative odd-number ones.

Third, the number of dyads consisting of two directed ties is often included in the model. If it is not and reciprocity is an effect in the network, then there would be more dyads with at least one tie. Thus, to control for the total number of connected dyads in the network and to account for the effect of reciprocity, this term is often included.

Fourth, the likelihood of observing triangles in networks can be accounted for by using a number of terms. The most common term is the number of transitive triplets in a network. This term would control for the tendency of two nodes to be tied if they tied to the same other node (see Chapter 2 and Holland and Leinhardt, 1971).

Finally, various terms have been included to describe homophily and focus con-

straints based on covariates or attributes of the nodes contained in the datasets. For example, if information about nodes' gender are included in a social network dataset, it is possible to test whether two people of the same gender are more likely to form a tie than two people of different gender.

This model is formalised as follow:

$$P\{Y = y\} = \frac{\exp(\theta Z(y))}{\kappa(\theta)} \tag{10}$$

where Y is an ensemble of random networks, y is the observed network, θ is a vector of coefficients or statistical parameters, $Z(y)$ are the counts of the sub-configurations in the network, and $\kappa(\theta)$ is a normalisation factor. The normalising factor is the sum of the expression in the numerator calculated for all possible realisations of random networks based on the chosen θ. This guarantees that the probabilities sum to 1. The values for θ are estimated by maximising the log-likelihood.

This model has been variously implemented in the three publicly available software packages that estimate ERG models: *Pnet* (Wang et al., 2005), *Siena* (Snijders et al., 2007), and *Statnet* (Handcock et al., 2003). In *Siena*, in addition to the above list of terms and a range of less common terms, interaction terms can be introduced in a similar fashion as in a standard multivariate regression. This allows for a better understanding of how various terms combine to affect the likelihood of a tie.

Furthermore, a related model to ERG model has been proposed by Snijders (2001) and collaborators (Snijders et al., 2008). They proposed a panel model that takes multiple observations of a network over time to study network evolution. This model is referred to as the SIENA model (implemented in the *Siena* software; Snijders et al., 2007)[17]. The panel model has a number of advantages. It allows for a deeper understanding of different processes, such as selection and influence

[17]This model might have been more appropriate for Powell et al. (2005) since they used yearly snapshots.

(Steglich et al., 2007). Selection terms are seen as factors that "attracts" ties, whereas influence terms are factors responsible for the changing of behaviour as a result of a social tie. The classical example is that of 129 pupils at a high school in Scotland where the social structure was observed at each year (1995-1997) from when they were roughly 13 to 15 (Pearson and Michell, 2000; Pearson and West, 2003). Using this model, Steglich et al. (2007) were able to differentiate whether smoking behaviour and alcohol consumption made pupils friends with each other (selection) or whether social ties affected the smoking behaviour and consumption of alcohol (influence).

A main shortcoming of the ERG and SIENA models is that the denominator or normalising factor cannot feasibly be calculated (Snijders, 2002; Wasserman and Pattison, 1996). This shortcoming stems from the fact that only a single snapshot of the network (or a limited number of snapshots for the SIENA model) is known. A consequence of this fact is that the exact sequence in which ties were formed and severed is unknown, which in turn, requires the entire network to be modelled at the same time (the dependent variable in the model is the entire observed network). The maximum likelihood procedure cannot be exact due to the extremely large number of possible realisations that random networks can take if there are more than a few nodes. Therefore, approximation procedures have been applied to detect coefficients that are close to the exact ones. Traditionally, the models relied upon a pseudo-likelihood estimation procedure (Wasserman and Pattison, 1996). However, this method has been found to estimate coefficients that are not close to the exact ones, and to produce unreliable standard errors (Snijders, 2002). Currently, a Markov chain Monte Carlo (MCMC) procedure is the recommended method for attaining coefficients that are close to the exact ones as well as determining their significance.

4.3 Longitudinal binary networks

We propose to use a model that is based on the conditional logistic regression framework (Breslow, 1996; Cox and Hinkley, 1974; Hosmer and Lemeshow, 2000) or discrete choice modelling (McFadden, 1973) to investigate growth mechanisms in the network evolution. This is a particular type of statistical model that tests whether chosen options or cases have certain properties that a set of other options or cases do not have. For example, suppose that a number of people can choose their mode of transportation to work. For each option that a person can choose, there are a set of known parameters specific to that option. This could include duration, cost, inconvenience, energy consumption, and comfort among others. In addition, for each person, the chosen option is known. Then, a conditional logistic regression could be used to probe which parameters, and the extent to which they guide the choice that a person makes.

This model can also be used to study other decision processes. In fact, we can apply this model to investigate why nodes form *new* ties with certain other nodes (i.e., binary ties). The components of this model are as follows. At a given time t, a node i decides to form a tie. This tie can be directed towards the set of available nodes in the network at that time A_t. Since we are studying a binary network, we assume that this set includes all the nodes in the network at time t that node i is currently not tied to. The node that receives the tie, node j, can have a number of properties, $Z_{j,t-1}$. These properties might include terms related the mechanisms outlined above, such as the in-degree of node j (preferential attachment) and whether node j has already formed a tie with node i (reciprocity). The purpose of the conditional logistic regression model is to see whether the properties of node j, $Z_{j,t-1}$, stand out from the properties of all the available nodes, $Z_{A_t,t-1}$. We choose to use the properties observed before the tie is created (Z_{t-1}) as we seek to understand why node j was the one that was selected by node i. We formalise the model as

follows:

$$P\{j_t = j|Z_{t-1}\} = \frac{\exp(\beta' Z_{j,t-1})}{\sum_{h \in A_t} \exp(\beta' Z_{h,t-1})} \tag{11}$$

where β' is a vector of coefficients. The coefficients that best fit the data are found by maximising the log of the equation (Hosmer and Lemeshow, 2000).

This model is analogous with other models that have been suggested in the literature. First, Snijders (2001) defined a model similar to the above model when proposing the SIENA model. Although the data used in the SIENA model is multiple snapshots of the network structure, the model is defined for network data where the exact sequence of ties is known, and estimation is used for the panel data. Second, Butts (2008) suggested a hazard or survival model for studying the network evolution. This model was applied to the radio communication data collection on the World Trade Center disaster. However, unlike Butts' model, the model above is concerned with directed one-mode networks, in which the aim is to detect preferences of nodes in how they direct ties toward other.

The conditional logistic regression model suffers from a number of limitations. A major one is that each tie can be directed towards many possible nodes (i.e., the set A_t is large). This implies that the ratio been the realised option or the observed tie (dependent variable equal to 1) and all the others (dependent variable equal to 0) is very small. In fact, most conditional logistic regressions in epidemiological studies have a ratio between 1 : 1 and 1 : 5 (Hosmer and Lemeshow, 2000). A lower ratio could create a number of issues for estimation of the coefficients (King and Zeng, 2001). In particular, the logistic regression models can greatly underestimate the probability of events when the ratio is very small.

For the online social network the ratio is extremely small. A user can choose to direct a tie to one of the 1,898 other users when all users are included in the network. Thus, the ratio can be as small as 1 : 1897. Moreover, the number of observations in the regression would be extremely large. Again, in the case of

the online social network with 20,296 binary directed ties, the size of the sample would be approximately 24 million observation (or 38.5 million observations if all nodes are considered available from $t = 1$). A common method to overcome these two limitations is to use a matched sample (Hosmer and Lemeshow, 2000; Powell et al., 2005). To this end, for each realised case, a number of other observations are picked as control cases. This method can also be applied to our proposed use of the conditional logistic regression framework. For each tie that is formed, a number of control nodes are selected. These should be randomly selected from the available nodes in the network[18]. In addition to making the sample balanced for each tie, this also ensures that the method can scale to large datasets.

We now consider a number of network growth mechanisms in an effort to understand how *new* ties were formed in the online social network (Panzarasa et al., 2009).[19] For example, in Chapter 2 the clustering coefficient revealed an above random likelihood of a generating a tie between two nodes that share a common contact in the network. In what follows, we first test a number of mechanisms independently, and then together, within the conditional logistic regression framework.

First, as mentioned above, the network was found to exhibit a high number of triangles as a clustering coefficient of 0.0568 was obtained for the undirected network (Eq. 2 in Chapter 2). This is over 7 times larger than what we would expect in a corresponding classical random network (Erdős and Rényi, 1960). Based on this finding, we hypothesise that the likelihood of a tie from one person to another increases as a function of the number of common friends the two people share. In particular, we define $Z_{h,t-1}$ as the number of nodes that 1) node i is already tied to,

[18]A sensitivity analysis of the number of control cases is conducted in Section 4.5. Unless otherwise specified, in the remaining of this chapter, for each observed tie we include 19 control cases. Thus, the sample includes 20 observations for each observed tie. The analysis was repeated multiple times with different sets of control nodes and results were consistent.

[19]Although, it can be reasonably argued that most networks develop over time, this is the only network available to us where the exact sequence of ties is known. Therefore, we can only apply the method within this Chapter to this network and not the other networks presented in Section 1.4.

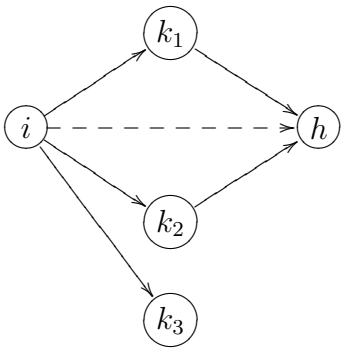


Figure 12: Example of triplets that originated at the creator and terminated at the target of a possible tie. Node i is tied to three nodes (k_1, k_2, and k_3), two of which are tied to node h. Therefore, $Z_{h,t-1}$ is equal to 2 for the $(i \rightarrow h)$-dyad when triadic closure is tested independently.

and 2) are already tied to node h (see Figure 12 for a schematic representation). We found support for this hypothesis. The coefficient that yielded the highest maximum log-likelihood was 0.233 with a standard error of 0.0389 ($p < 0.001$). This means that people in the online community are more likely to create a tie to others with whom their existing contacts are already tied to. More specifically, given two users, for each additional common contact they share, the likelihood of a tie between them increases by 26% ($e^{0.233}$ equals an odds ratio of 1.26).

Second, the ties are not homogeneously distributed across the nodes in the online community (Panzarasa et al., 2009). While the majority of nodes in the network is connected to few others, there is a subset of extremely well-connected nodes. More specifically, the in-degree distribution follows a power-law function, $p(k^I) \propto (k^I)^{-\tau}$, with an exponent τ of less that 2 (see the solid line in Figure 13). This suggests that the network is "scale-free" (Barabási et al., 2002). This distribution has been replicated in growing random networks where nodes that already have received relatively many ties have a higher probability of attracting new ties than nodes with fewer ties (Barabási et al., 2002; Dorogovtsev and Mendes, 2003; Jeong et al., 2003). Thus, it has been speculated that a "popularity is attractive"-mechanism is the cause of the observed distribution. Since we found a scale-free distribution for the online

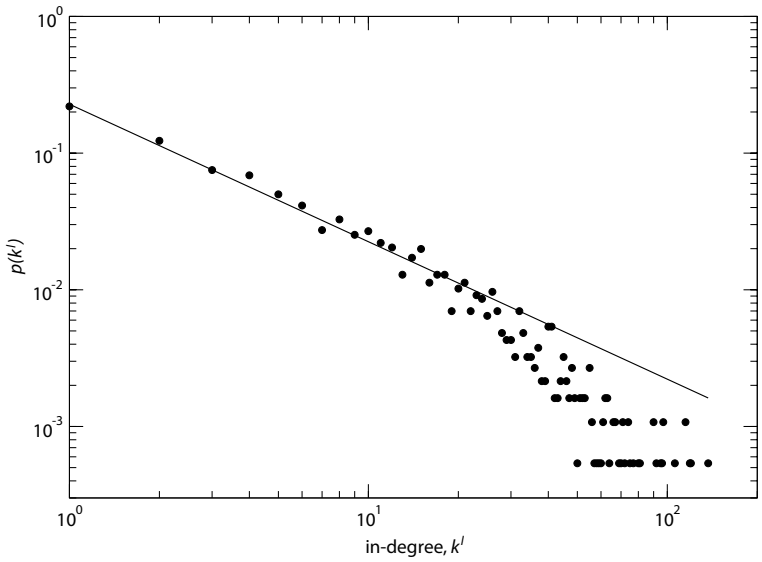


Figure 13: In-degree distribution: $p(k^I)$ is the fraction of nodes with a certain in-degree, k^I. The distribution is fitted by $p(k^I) \propto (k^I)^{1.005}$. The scales are log-log to emphasise the power-law shape of the distribution.

social network, we hypothesise that users in the online community were attracted to other users that were "popular". More specifically, we define $Z_{h,t-1}$ as the number of nodes that are already tied to node h before a possible tie is created at t. We found support for this hypothesis as a coefficient of 0.037 was obtained with a standard error of 0.0006 ($p < 0.001$). For each additional tie that terminates at a node, the likelihood that a node will receive a new tie increases with 3.7%.

Third, 12,916 of the 20,296 directed ties in the online community occurred within a dyad with another tie (reciprocated). In a classical random network the number of reciprocated ties is likely to be much smaller. More specifically, since each directed tie is independent of all the other ties, the probability of a tie being reciprocated is equal to the probability of a tie (Erdős and Rényi, 1960; Rapaport, 1953). For the online social network, the probability of a directed tie is 0.0056. Thus, the expected number of reciprocated ties is 114. Based on the large difference between the observed and expected number of reciprocated ties, we hypothesise that reciprocity exerted a strong effect on the evolution of the network. $Z_{h,t-1}$ was

defined as 1 if node h had already directed a tie towards node i, and 0 otherwise. We found support for this hypothesis as a coefficient of 4.85 was obtained with a standard error of 0.105 ($p < 0.001$). This implies that reciprocated ties are 128 times more likely to be formed than non-reciprocated ties.

Fourth, we focus on homophily, namely the tendency of similar nodes to create ties among themselves (Lazarsfeld and Merton, 1954; Louch, 2000; McPherson et al., 2001). This mechanism has been found to be responsible for the creation of tightly knit groups of similar individuals in a social network (Kossinets and Watts, 2006). This might be due to similar background (Hallinan and Kubitschek, 1988; Lazarsfeld and Merton, 1954; McPherson et al., 2001). Drawing on these empirical studies, we hypothesise that the likelihood of a tie between two people increases as a function of their social similarity. When students registered for the online community, they supplied a number of demographic characteristics or attributes, v. These included the individuals' gender, age, year of study, region of origin, and marital status. For each ordinal attribute (i.e., age and year of study), we constructed a similarity index. This index was defined the $(i \rightarrow h)$-dyad as 1 minus the standardised absolute difference between $v(i)$ and $v(h)$[20]: $1 - \frac{|v(i)-v(h)|}{\max(v)-\min(v)}$. For the other nominal attributes, we constructed a dummy term indicating whether the two nodes had the same value, e.g., set to 1 if two nodes have the same gender and 0 otherwise. By testing these terms independently, we found that the coefficients for all attributes, excluding gender, were positive and significant (see Models 4-7 in Table 6). Conversely, having the same gender significantly decreased the likelihood of creating a tie (see Model 8). This implies that a male (female) user was more likely to communicate with a female (male) one than with another male (female) user.

Finally, we tested the effects of focus constraints on network evolution. This mechanism is responsible for the increase of the likelihood that institutionally or

[20]This formula is also used by *Siena* (Snijders, 2001).

	Model				
Variables	1	2	3	4	5
Triadic closure	0.233*** (0.039)				
In-degree		0.036*** (0.001)			
Reciprocity			4.85*** (0.105)		
Similar age				1.07*** (0.202)	
Similar year of study					1.30*** (0.070)
Wald χ^2	36***	2,433***	2,145***	28***	346***

	Model				
Variables	6	7	8	9	10
Triadic closure					0.038 (0.042)
In-degree					0.029*** (0.001)
Reciprocity					4.605*** (0.112)
Similar age					0.142 (0.127)
Similar year of study					1.205*** (0.073)
Same area of origin	0.556*** (0.025)				0.492*** (0.037)
Same marital status		0.289*** (0.024)			0.343*** (0.036)
Same gender			-1.29*** (0.052)		-1.272*** (0.062)
Same school				0.278*** (0.028)	0.308*** (0.035)
Wald χ^2	482***	143***	626***	102	3,105***

Table 6: Growth mechanisms in a binary network tested in a conditional logistic regression framework with 20 observations for each observed tie. To ensure comparability across the models, the same set of control nodes was used. Robust standard errors adjusted for clusters based on the sender of a tie are in parentheses.
† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. N=405,340.

geographically bounded nodes form ties among themselves (Feld, 1981; Monge et al., 1985). As triadic closure and homophily, this mechanism is therefore responsible for the generation of groups of well-connected nodes. When students registered on the web site of the community, in addition to the demographic attributes, they also supplied information about the course they attended. Based on this information, we hypothesise that two students belonging to the same school have a larger likelihood of creating a tie than two students from different schools. We found support for this hypothesis as a coefficient of 0.278 was obtained with a standard error of 0.028 ($p < 0.001$). This implies that belonging to the same school increases the likelihood of forming a tie with 32%.

A limitation of the analysis we conducted so far lies in the fact that it focuses on a single mechanism in turn. However, networks are likely to evolve as a result of a combination of mechanisms (Snijders, 2001; Wasserman and Pattison, 1996). This limitation can be easily overcome since we are using a regression framework which allows for multiple independent variables (Cox and Hinkley, 1974; Hosmer and Lemeshow, 2000). Thus, we can to assess the likelihood of a new tie as a function of two or more terms.

Model 10 in Table 6 shows the coefficients and significance of all the terms previously tested independently when tested together. A number of observations are in order. First, the two terms with the highest absolute z-scores (i.e., the coefficient divided by the standard error) are network effects, namely reciprocity ($z = 41$) and in-degree ($z = 35$). This signals the importance of path dependency in the evolution of the network.

Second, the two terms *similar age* and *triadic closure* lost their significance. These were also the terms with the lowest absolute z-score in the independent tests. A possible reason for the loss of significance is multicollinearity. If two terms are representing the same underlying factor in the data, the maximum likelihood esti-

mation would not be able to determine which of the two terms is responsible for the increase in likelihood of a tie. This might be the root of the non-significance of both these terms. Conceptually, there is a link between the age of a person and the year of study as most people in the US join university when they leave school at the age of 18. In fact, between *similar age* and *similar year of study* the pair-wise correlation coefficient is 0.2649, whereas between *similar age* and all other terms the coefficients are less than 0.1. The results in Table 6 might indicate that it is not age of people that determines whether ties are formed, but the advancement in their university education.

The *triadic closure* term is associated with the term *indegree*. Given an equal number of triplets originating from the creator of a tie, the higher the in-degree of a target node, the more triplets are likely to terminate at that node. Therefore, the term *triadic closure* should be correlated with the term *indegree*. In fact, the pair-wise correlation coefficient between *triadic closure* and *indegree* terms is 0.2998, whereas the coefficients between *triadic closure* and all other terms are less than 0.1. Thus, the positive effect of *triadic closure* in the independent test might be a reflection of the correlation with *indegree*. It is rare to find a social network without a triadic closure effect. However, this network is a special kind of social network where a person communicates individually with his or hers contacts and never in a group. Therefore, the contacts of a person cannot be observed by each other, and they might not be aware of each other.

Third, the size of the in-degree effect is reduced when other mechanisms are included. Unlike an increase of 3.7%, when combined with other measures, each additional in-degree only increases the likelihood of receiving a tie by 2.9%. This result should be assessed in the light of the controversal debate between two camps within the social networks community. On the one hand, there are scholars, especially statistical physicists, that tend to argue in favour of the ubiquity of preferential at-

tachment as a growth mechanism that explains how a variety of real-world networks evolve (Barabási et al., 2002; Dorogovtsev and Mendes, 2003). On the other, it is argued, especially by social scientists, that preferential attachment is inaccurately measured (Borgatti et al., 2006) or that other forces are at work in driving network evolution, such as homophily and triadic closure (Powell et al., 2005; Kossinets and Watts, 2006). Our results indicate that these arguments are not mutually exclusive and together they may well suggest that the evolution of the network is governed by the contribution of multiple mechanisms.

4.4 Longitudinal weighted networks

An aspect of social networks that is often overlooked is the weight of ties as most empirical studies are in fact based on binary network datasets, and therefore give no indication about the role of ties of different strength in shaping the structure and function of the network (see Section 1.1). However, strong and weak ties are associated with different properties, and it is therefore important to distinguish between them. For example, strong ties have been associated with trust and transfer of tacit knowledge (Reagans and McEvily, 2003; Levin and Cross, 2004; Uzzi and Spiro, 2005), whereas weak ties have been linked with access to novel explicit knowledge (Burt, 1992; Granovetter, 1973).

The ERG and SIENA models currently cannot be applied to weighted networks (an extension has been proposed by Snijders and Steglich, 2008). Conversely, the model proposed in this Chapter can easily be extended to weighted networks by relaxing some of its assumptions. In weighted networks, multiple ties can exists from one node to another. This means that a tie is not only formed when a node interact with another for the first time, but every time an interaction occurs. Multiple ties reinforce the tie already existing from the creating node to the target node. Therefore, in a weighted network, we assume that A_t includes all other nodes in

the network when a tie is created, even those that the creator of the tie is already connected to.

Moreover, in weighted networks additional mechanisms might be responsible for tie generation. In the online community, an average user has sent 31.5 messages to 10.7 people. This means that two-thirds of messages were used to reinforce existing ties. Since the network is sparse and on average ties have been reinforced roughly two times, we hypothesise that reinforcement is likely to affect the likelihood of a future tie. In other words, we believe that a new tie is more likely to occur between two nodes that are already connected than between nodes that are disconnected. We found support for this hypothesis as we obtained a positive and significant coefficient of 0.7598 with a standard error of 0.0759 ($p < 0.001$). This coefficient translates into an odds ratio of 2.14 ($e^{0.0759}$), which suggests that each previous message sent between two nodes roughly doubles the likelihood of another message being sent.

However, this result is obtained by testing reinforcement independently without taking into account the other mechanisms that proved to be significant for the evolution of the binary network. Thus, the next step is to include multiple effects in our assessment of weighted networks. Model 2 in Table 7 shows reinforcement modelled together with the terms used in the binary analysis. Even though in this case we found a smaller coefficient for reinforcement, the coefficient remained positive and extremely significant ($z = 8.7$).

All of the terms that were previously used in the analysis of network evolution were designed for binary networks. However, some of them can be generalised to weighted networks. As shown in Chapter 2, a triplet value can be used to differentiate triplets. We proposed four methods for determining the triplet value. It can be the arithmetic mean, geometric mean, the maximum, or the minimum of the two weights that compose the triplet. Building on Opsahl and Panzarasa (2009), Snijders and Steglich (2008) used the concept of a triplet value to model triadic

	Model				
Variable	1	2	3	4	5
Reinforcement	0.760***	0.2168***	0.2162***	0.2177***	0.2107***
	(0.076)	(0.0249)	(0.0249)	(0.0251)	(0.0253)
Triadic closure		0.0470†			0.0871**
		(0.0274)			(0.0292)
Weighted triadic closure (Min)			0.0271*		
			(0.0124)		
Weighted triadic closure (GM)				0.0008	
				(0.0036)	
Node in-degree		0.0213***	0.0212***	0.0220***	
		(0.0009)	(0.0009)	(0.0009)	
Node in-strength					0.0047***
					(0.0002)
Reciprocity		4.0676***	4.0684***	4.0705***	4.1055***
		(0.0868)	(0.0870)	(0.0871)	(0.0869)
Similar age		-0.0662	-0.0693	-0.0662	-0.0719
		(0.1078)	(0.1072)	(0.1077)	(0.1072)
Similar year of study		1.0732***	1.0732***	1.0770***	1.0840***
		(0.0725)	(0.0727)	(0.0727)	(0.0731)
Same area of origin		0.4258***	0.4267***	0.4274***	0.4410***
		(0.0309)	(0.0309)	(0.0310)	(0.0307)
Same marital status		0.2916***	0.2924***	0.2904***	0.3003***
		(0.0287)	(0.0288)	(0.0287)	(0.0286)
Same gender		-1.1189***	-1.1220***	-1.1070***	-1.1340***
		(0.0545)	(0.0548)	(0.0555)	(0.0549)
Same school		0.2494***	0.2487***	0.2493***	0.2545***
		(0.0335)	(0.0336)	(0.0336)	(0.0334)
Wald χ^2	100	3,541***	3,367***	3,494***	3,533***

Table 7: Growth mechanisms in a weighted network tested in a conditional logistic regression framework with 20 observations for each observed tie. To ensure comparability across models, the same set of control nodes was used in all of them. Robust standard errors adjusted for clusters based on the sender of a tie are in parentheses. † $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. N=1,196,400

closure in a proposed extension of the SIENA model to weighted networks. They used the minimum and geometric mean methods for defining the triplet value and for the $(i \rightarrow h)$-dyad summed the values of the triplets that originate at node i and terminate at node h. These terms can be formalised as follows:

$$\text{Weighted triadic closure (minimum)}(i,h) = \sum_k \min(w_{ik}, w_{kh}) \tag{12a}$$

$$\text{Weighted triadic closure (geometric mean)}(i,h) = \sum_k \sqrt{w_{ik} \times w_{kh}} \tag{12b}$$

where i is the creating node and h is the target node of a possible tie, and k represents any other node that i is tied to.

To illustrate these two effects, Figure 14 exemplifies a possible directed tie between node i and h (dashed line). Node i is tied to three other nodes k_1, k_2 and k_3. The first two of these nodes are tied to the target node h. The two terms would be equal to 5 and 6.29, respectively, for this tie[21].

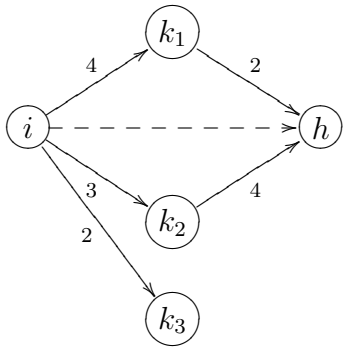


Figure 14: Example of weighted triplets that originate at node i and terminate at target node h of a possible tie.

Following our results for the binary network, we first hypothesise that the two terms, formalised in Equations 12a and b, increase the likelihood of a tie. In addition, we hypothesise that if the two generalised terms replace the simpler binary term (the number of common friends; the term *triadic closure* as defined in the previous

[21]The calculations are: $\min(4,2)+\min(3,4)+\min(2,0) = 5$ and $\sqrt{4 \times 2}+\sqrt{3 \times 4}+\sqrt{2 \times 0} \approx 6.29$ for the two methods, respectively.

section) the models produce a higher Wald χ^2 due to the finding in Chapter 2 that strong ties are more likely than weak ties to be part of closed triplets in the online social network (see Section 2.3). As shown by Models 3 and 4 in Table 7, we found only partial support for the first hypothesis[22]. Moreover, Models 3 and 4 had a lower Wald χ^2 than Model 2. Therefore, we did not find support in favour of our second hypothesis that the generalised triadic closure terms were better than the binary term.

Furthermore, for weighted networks in-degree is often redefined as node in-strength (see Chapter 3, and Barrat et al., 2004; Newman, 2004a; Opsahl et al., 2008). A node's in-strength is the sum of the weights attached to the ties that terminate at the node. This is equal to the in-degree of a node if all ties carries a weight of 1, i.e. if the network is binary. Therefore, this term can only be included in an assessment of weighted networks. In the light of the improvement of results in Chapter 3 when out-strength replaced out-degree, here we hypothesise that a similar improvement of results is likely to occur when in-strength is used instead of in-degree. However, as shown by Model 5 in Table 7, we did not find support for this hypothesis. In fact, a lower overall Wald χ^2 was found when in-degree was replaced with in-strength. A possible explanation for this is that it is not the total number of messages received, but the number of unique contacts that most accurately gives an indication of a user's popularity.

Finally, the results from the analysis of the binary network are consistent with what was found for the weighted network. First, the two terms with the highest absolute z-scores are still network effects: *reinforcement* and *reciprocity*. Second, the term *similar age* remains not significant, while the term *triadic closure* is just within the $p < 0.10$ level. This suggests that people's age is indeed not a relevant predictor of ties, and triadic closure does not have a strong statistically significant effect on

[22]We chose not to test the three triadic closure terms together due to high correlation. This correlation is not surprising since the three terms are operationalisations of the same mechanism.

network growth. Third, the magnitude of the in-degree effect is further mitigated in the analysis of the weighted network. In Model 2 of Table 7, the increase in likelihood of receiving a tie only increases of 2.15% for each additional in-degree of a node. This suggests that the explanatory power of preferential attachment as a mechanism for network dynamics is mitigated when a broader prespective is adopted that takes into account not only the reinforcement mechanism, but also the weight of ties and ties used to reinforce.

4.5 Sensitivity to the number of control cases

The findings presented in the previous sections are based on a sample of 20 observations for each observed tie: the tie that was formed and 19 randomly selected non-formed ties. However, in the literature several rules of thumb exist with regard to the appropriate number of control observations (Cosslett, 1981; Hosmer and Lemeshow, 2000; King and Zeng, 2001). Cosslett (1981) argued that the optimal number of control cases is same as the number of realised cases (i.e., observed ties). This implies that the sample of observations in the regression is strictly balanced (i.e., a ratio of 1 : 1). However, King and Zeng (2001) argued for a sensitivity analysis of the number of control cases. The "optimal" number was found when an additional control case did not decrease the standard errors (or increase the significance).

We have undertaken a similar sensitivity analysis for the assessment of both the binary and weighted networks. Figure 15 shows the significance (z-score) of in-degree tested independently in a) the binary network, and b) the weighted network, when the number of control cases increases. As shown, the marginal increase in z-scores is very small when approximately 20 control cases are used. Thus, additional control cases would not add value to the analysis. In fact, it might introduce measuring errors (King and Zeng, 2001) and would increase the computational requirements.

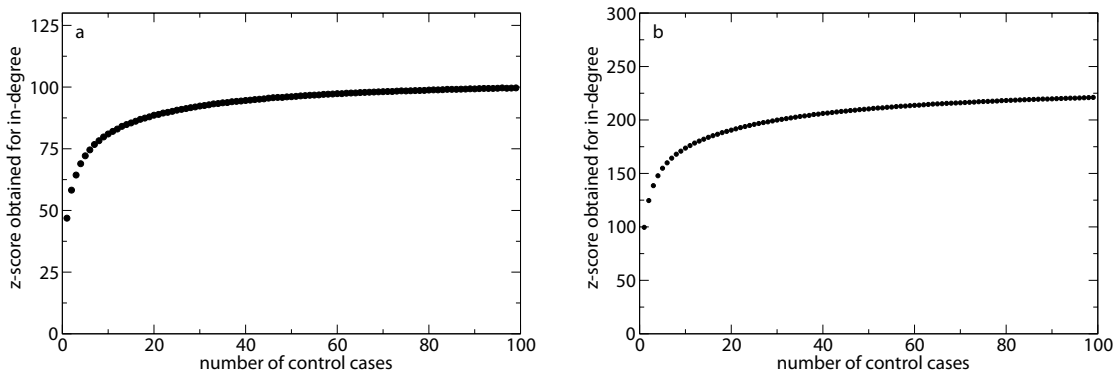


Figure 15: Significance (z-score) of in-degree independently tested in the binary (a) and the weighted (b) network with an increasing number of control cases. The z-scores are the average values obtained with 30 regressions with different sets of control nodes.

First, King and Zeng (2001) found that the statistical properties of logistic regression models is not invariant to the unconditional mean of the dependent variable. They showed through a number of simulations that the estimates obtained by traditional logistic regression models is biased. In fact, the estimates are biased in a specific direction: they are underestimated. This bias can be corrected through a number of methods; however, the problem can be overcome altogether by using a limited number of control cases.

Second, since the number of observations in the regression is the number of observed ties multiplied by the number of control cases plus one, the computational requirements of running a regression model is dependent upon the number of control cases. If there is no constraint on the number of control nodes taken into consideration (i.e., it is the number of other nodes in the network), the number of observations can become extremely large and demand a great deal of memory. In fact, a regression of the binary network with 20,296 observed ties when A_t was not bounded required 36 gigabyte of memory in the statistical programme R (R Development Team, 2008). This is beyond the capabilities of Windows XP, and generally outside the scope of computers running Linux or Unix. Thus, dedicated

servers must be used to estimate models of this size. Yet, the online social network is a fairly small network. For larger networks, such as the network of 2.9 million utility patents connected through 16.5 million citations used by Hall et al. (2001), it would not be feasible to estimate a model. If all patents were available from $t = 1$ and A_t was not bounded, there would be 48,310 billion observations in the sample. A regression model of this size would certainly be outside the scope of the current computational capabilities.

4.6 Contribution to the literature

The contribution of this project is two-fold. First, we proposed a general method for analysing the evolution of longitudinal network data. The method is based on a regression framework that enables us to test how different mechanisms jointly drive network evolution. It is not the first time this regression framework has been used to study network evolution. Powell et al. (2005) studied tie generation in an inter-organisational network of pharmaceutical firms. However, due to the nature of their data, the model they applied was not generalisable to other datasets. More specifically, they combined one- and two-mode data. Moreover, they included four different terms to account for preferential attachment, and calculated homophily based on similarity with previous partners. In addition, the data were only recorded at yearly intervals and, therefore, the dependence among ties occurring in the same year was not taken into consideration. The method that we proposed in this Chapter is more general and a special case of the actor-perspective applied by the SIENA model (Snijders, 2001).

Second, we have empirically tested six growth mechanisms independently and jointly in binary and weighted networks constructed from a self-organising virtual community. These networks are over four times larger than the one used by Powell et al. (2005) and the exact sequence of ties is known. Two often overlooked aspects

of network analysis are reciprocity and reinforcement. Since communication in the community was directed, and repeated interactions were recorded, we could test for these two aspects. Our findings allow for a better understanding of the underpinning principles that guide interpersonal dynamics in an online environment. In particular, we found little or no effect of triadic closure depending on model specification. This is surprising as the number of common friends has generally a strong effect on the evolution of social networks (Steglich et al., 2007). Nevertheless, this result comes from an online social network. Online behaviour might differ from offline behaviour, and different mechanisms might underpin tie formation in the two contexts. By shedding light on these principles and results, managers and moderators of online communities are able to device better strategies for improving communication and knowledge transfer.

4.7 Conclusion and discussion

In this Chapter, we applied a well-tested regression framework, often used in epidemiology (conditional logistic regression; Cox and Hinkley, 1974) and economics (choice modelling; McFadden, 1973), to the study of the principles underpinning network evolution. In economics, this framework is used to model the decision-making process of people in various settings, whereas in epidemiology it is used to model the factors that affect a specific medical condition, such as a particular cancer (e.g., Pastides et al., 1983). We proposed to apply this framework to the decisional processes of nodes when creating ties.

Nodes do not form ties with others randomly (Erdős and Rényi, 1960; Holland and Leinhardt, 1981; Solomonoff and Rapoport, 1951; Snijders, 2001; Wasserman and Pattison, 1996). Conversely, they are likely to direct their ties towards other nodes with whom they have common contacts (Heider, 1946), popular nodes (Barabási et al., 2002), similar nodes (McPherson et al., 2001), and nodes with whom

they share geographical location or institutional context (Feld, 1981). However, due to the fact that most networks are collected at a single point in time, it is difficult to analyse the choices that nodes make when forming a tie. In fact, a range of methods have been developed to simulate the sequence in which ties are formed, and then analyse the decision process of nodes (Snijders, 2001; Snijders et al., 2008).

We did not pursue this line of investigation. Conversely, we studied a new type of network data that have become available in recent years: longitudinal network data (e.g. Hall et al., 2001; Holme et al., 2004; Kossinets and Watts, 2006; Onnela et al., 2007; Panzarasa et al., 2009). A special feature of this type of data is that the exact sequence of ties is known and, therefore, we can measure the properties, such as in-degree, of all nodes available in the network to whom a node could have directed a tie at the time the tie was formed. Based on the properties of potential targets of a tie, we were able to model the decision processes of nodes.

This method was applied to a social network constructed from a virtual community. The nodes of the network were college students who could create ties with each other by send online messages. This network is a prototypical evolving social network where the nodes are people who are in control of their outgoing ties. Therefore, it allowed us to explore the general regularities governing the initiation and progression of interpersonal dynamics.

The findings clarify and extend past research by focusing on critical issues that tend to be overlooked in studies of the evolution of networks, such as directionality and reinforcement. We first dichotomised the network and tested independently a number of mechanisms leading to the generation of *new* ties. We found support for triadic closure, popularity, reciprocity, homophily, and focus constraint when tested independently. The only exception was gender homophily which proved to have a negative effect on the evolution of the network.

A key benefit of using a regression framework is the possibility of testing multiple

effects at the same time. This allowed us to study how effects jointly drive network evolution. Results were consistent with the independent tests, except for the *triadic closure* and *similar age* terms that became insignificant. We speculated that the significance results obtained in the independent tests were due to multicollinearity. Moreover, the effect of in-degree was mitigated in the multivariate analysis.

Furthermore, we extended the method to cover weighted networks by relaxing some of the assumptions made by the model and adding terms specifically designed for weighted networks. We found strong support in favour of reinforcement, thereby suggesting that students are much more likely to communicate with someone they have already communicated with. Moreover, the results were roughly consistent with the analysis of the binary network. In particular, we found a further mitigation of in-degree. In addition, we generalised triadic closure using two of the methods for calculating triplet values proposed in Chapter 2. The generalised term based on the minimum method for calculating triplet values was significant ($p < 0.05$). Nevertheless, the models with the generalised terms produced lower Wald χ^2 than the ones with the binary term.

The lack of a strong positive and significant effect of having common friends on tie generation in the online social network is surprising as this generally a strong effect in networks (Snijders, 2001; Wasserman and Pattison, 1996). In most offline social settings, communication occurs in groups larger than two. In these settings, the contacts of an individual can observe each other. Conversely, in the virtual community, individuals could only communicate one-to-one. Thus, an individual's contacts could not observe each other. Furthermore, in May 2008, the virtual community Facebook launched a service called People You May Know[23], which proposed new possible contacts to an user. This feature was heavily based on common contacts. However, in September when a new interface was launched, this service did

[23] `http://blog.facebook.com/blog.php?post=15610312130`

no longer form part of the first page users see when logging in. The reduce of focus might suggest a lack of use due to inaccuracy by relying on common friends.

The empirical analysis we conducted is not without limitations. We could not verify that the messages did indeed reflect genuine interpersonal communication. A possible method for verifying whether this is the case is to study message content. However, due to privacy reasons, we could not study the content of messages. Moreover, the information supplied when the students registered for the community was not validated. Only students' email addresses were validated to guarantee that they were in fact students at the university. In addition, the dataset does not contain any information about the weakening or severing of ties. Thus, a tie created at the very start of the community was assumed to remain in the network till the end.

The proposed method is not without limitations. The main one is that required data are difficult to obtain. However, due to the increase in use of electronic medium for social interaction, such as social network sites (Leskovec et al., 2005; Wellman, 1999), and the rise of machine-readable databases with interaction data, such as online repositories of scientific papers (e.g. Newman, 2001a), we believe that this type of data is likely to become more common in the future. In addition, as we are exploring the decision processes of nodes, this method relies on networks where the nodes are in charge of their ties. This is not always the case. For example, in the movie network (Watts and Strogatz, 1998) or the Broadway musical network (Uzzi and Spiro, 2005), ties among people might not entirely reflect people's decisions as it is the casting directors that design the teams of people working together on projects.

The method developed in this chapter is general and flexible. From an actor-based perspective, researchers can test additional growth mechanisms. These could include the geodesic distance among nodes (Wasserman and Pattison, 1996) or dyadic covariates (Snijders, 2001). Furthermore, the method is not limited to social networks. For example, if the sequence in which neurons create synapses and gap

junctions can be recorded, this method might yield new and interesting findings. Moreover, the method is not limited to an actor-based perspective. A dyad-based perspective might be adopted to study undirected networks. This would require new terms that take into consideration the decisional process of both nodes when forming a tie. An example of an undirected network where the exact sequence of ties is possible to map is the airport network used in Chapter 3 as routes start and terminate at specific points in time. Moreover, in this network, the weakening of ties (or decrease of capacity) also occurs at specific times.

5 *tnet*: Software for Analysis of Weighted and Longitudinal networks

As seen in the previous chapters of this thesis, information of tie weights as well as on the exact time ties are created or severed enables us to uncover and study interesting network properties through novel methods. However, not only are few network measures applicable to weighted and/or longitudinal networks, but there is also a lack of integrated software programmes that can deal with these types of networks. To the best of our knowledge, there are no network analysis programmes that can deal with weighted networks and allow for users to create their own functions. On the one hand, programmes like UCINET and Pajek have a small set of functions for weighted networks, but they do not allow for users to programme additional functions (Batagelj and Mrvar, 2007; Borgatti et al., 2002). Therefore, researchers proposing new measures must create stand-alone programmes to deal with a single aspect of weighted networks (e.g., Newman, 2001c).

On the other, a number of packages dealing with network analysis have been created within the open-source statistical programme *R*, notably the *sna* and *statnet*-packages (Butts, 2006; Handcock et al., 2003). These packages allow researchers to create additional functions on top of existing ones. This ability reduces the time spent on programming greatly, and let researchers focus on the contribution to the literature instead. For example, if someone has already written a function for identifying the shortest paths in a network, a researcher that would like to extend this measures can simply work on this code without programming the function from scratch. However, the *sna* and *statnet*-packages rely on the basic *network*-package for data structures to represent networks (Butts et al., 2008). This basic package does not have data classes for weighted or longitudinal networks. Therefore, to allow researchers to easily create new functions for weighted and longitudinal networks, a

new platform is needed.

This new platform should be able to handle both types of datasets. To this end and to disseminate the methods proposed in this thesis, *tnet* was created. Although this is a user-written package in *R* similar to the *sna* and *statnet*-packages, it does not rely on the *network*-package. It utilises its own data structures: one for weighted static networks and one for longitudinal networks. The longitudinal network structure can represent both binary and weighted networks.

For each of these two structures, *tnet* contains a set of functions. First, for the analysis of weighted networks, in addition to the functions to calculate the measures proposed in Chapter 2 and 3, a set of centrality measures (Barrat et al., 2004; Newman, 2001c), random network generators, and support functions are included. Second, for the analysis of longitudinal networks, functions include the framework proposed in Chapter 4 to study the tie generating mechanisms in longitudinal networks as well as random network generators and support functions.

The rest of this chapter is organised as follows. First, the two data structures and their supporting functions are introduced. Then, the functions for studying structural properties of weighted networks are presented. Only functions which have not been thoroughly described in previous chapters will be presented in detail. In Section 5.3, functions to deal with longitudinal networks are presented. Finally, we will highlight the contributions of this package to the scientific community and offer some concluding remarks. Appendix C includes the source code and the specific details for running the various functions within *tnet*. Moreover, the supporting website (`http://opsahl.co.uk/tnet/`) includes download instructions, examples, and guides to transfer data from other software programmes.

5.1 Data structures

tnet uses two basic data structures depending on the nature of the data. The first one is used to represent weighted static networks. Since most networks are sparse (i.e., the number of ties is much lower than the squared number of nodes: $A \ll N^2$), we opted for an edgelist format instead of a matrix one. This is a format that records the sender and receiver of established ties. The main advantage of this format is that it can scale to networks with many nodes as it is the number of ties, not nodes, that determine the size of the data object. Although many programmes can read edgelists, most network analysis programmes rely on an internal matrix representation, e.g. UCINET and the *network*-package (Borgatti et al., 2002; Butts et al., 2008). Conversely, Pajek, which was designed to analyse large-scale sparse networks, specifically uses an internal edgelist representation (Batagelj and Mrvar, 2007).

A binary edgelist consists of two columns that represent pairs of nodes that are tied together (e.g., the edgelist1-format in UCINET's dl files; Borgatti et al., 2002). When a directed network is represented, the first column represents the nodes that create the ties, whereas the second column represents the target nodes. Thus, an edgelist is an $A \times 2$ matrix.

This type of list has been extended to cover weighted networks by adding a third column representing the weight of the ties (Borgatti et al., 2002). In an effort to stay consistent with existing data structures, this is also the structure used by *tnet*. The class of the object in *R* should be `data.frame`. This class of object allows the different columns of a table to be of different classes themselves, such as integer and numeric. Thus, the class is more efficient at storing data than a matrix which requires all columns to be numeric. The first two columns of object are assumed to be integers (i.e., the identification number of the node creating the tie and the identification number of the node receiving the tie, respectively). The third column

can be real numbers (numeric) that represent the weights attached to the ties.

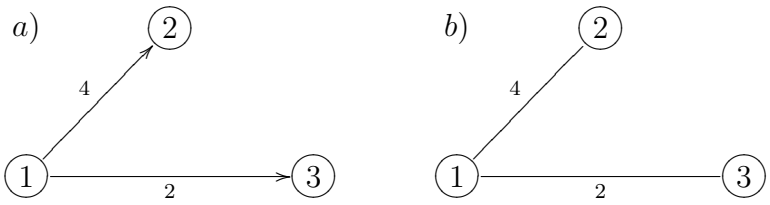


Figure 16: Example of a directed (a) and an undirected (b) network with weighted ties.

To illustrate the edgelist structure, we represent the directed network in Figure 16a by using the following matrix:

```
1    2    4
1    3    2
```

Table 8: Format for directed weighted edgelists: a $A \times 3$ `data.frame` object in *R*.

To represent an undirected network, each undirected tie must be included twice – one in each direction. Therefore, the undirected network in Figure 16b should be represented by the following matrix:

```
1    2    4
2    1    4
1    3    2
3    1    2
```

Table 9: Format for undirected weighted edgelists: a $2A \times 3$ `data.frame` object in *R*.

There are a number of functions that help the users to convert other formats into the weighted edgelist format. For example, if the dataset is undirected, but there is only one entry for each tie in the edgelist, the `symmetrise` function adds a second entry of the edge with the identification numbers of the creator and target nodes reversed. Moreover, if the dataset is similar to an edgelist, but with only two

columns (representing the identification numbers of the creator and target nodes) and multiple entries of the same tie refer to the weight of that tie (e.g., if a tie has a weight of 3, it is included three times), then the `shrink_to_weighted_network`-function allows the users to convert the edgelist into the correct format.

To allow for a comparison between weighted and binary network measures, the `dichotomise`-function creates a binary network from a weighted one. It does so by removing the ties in a weighted edgelist that fall below a certain cut-off and sets the weight to 1 for the remaining ones.

The second data structure is the longitudinal one. This structure represents network data where the sequence of ties is known, i.e. each tie is associated with a timestamp. For example, these network datasets include those obtained by studying phone and online interaction log files. In recent years, there has been a rise in the availability of this type of datasets (e.g. Ebel et al., 2002; Hall et al., 2001; Holme et al., 2004; Kossinets and Watts, 2006; Onnela et al., 2007; Panzarasa et al., 2009). This is mainly due to the fact that people communicate more through electronic mediums (Wellman, 1999), and the providers of these mediums are often required by law to keep a log of the activity where the time is explicitly stated. Yet, few methods have been developed for this type of datasets, and no general programmes have been developed to study them.

Figure 17 exemplifies the first six time steps of such a network. At $t = 1$ a tie is created from node 1 to node 2 among the only two nodes in the network. An isolate (node 3) joins the network at $t = 2$, and the tie from node 1 to node 2 is reinforced, whereas in the following time step, the tie is weakened. At $t = 4$, node 1 creates a tie with the isolate node, node 3. Node 3 reciprocates this tie at $t = 5$, and direct another tie towards node 2 ($t = 6$). The last tie closes the non-vacuous triad starting at node 3 to node 2 through node 1 (Wasserman and Faust, 1994, 243).

The longitudinal data structure is a $T \times 5$ `data.frame` object, where T is the

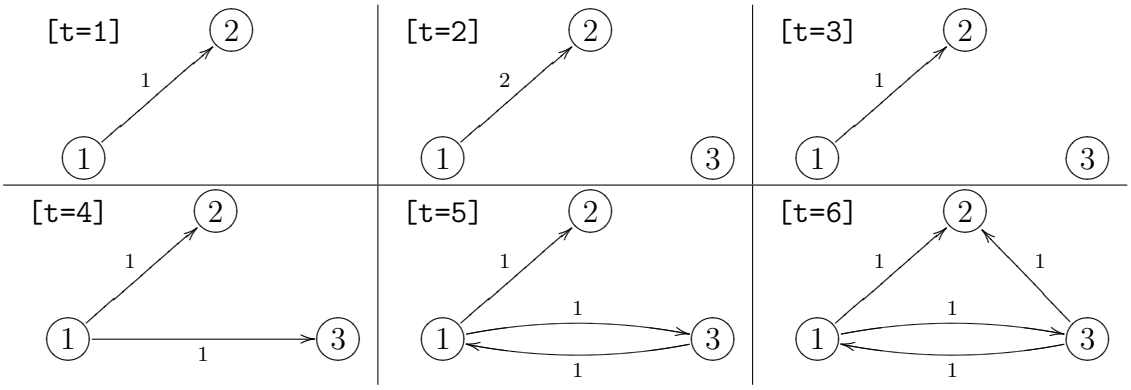


Figure 17: Example of the first 6 time steps in a longitudinal network.

total number of time steps. The first column represent the time at which the tie was created and should be a string with a standard Unix/SQL timestamp, i.e. YYYY-MM-DD HH:MM:SS where YYYY is the four digit year, MM is the two digit representation of the month, DD is the two digit day of the month, HH is the two digit hour of the day in 24-hour format, MM is the minutes, and SS is the seconds. The second and third column are integers that represent the identification numbers of the creator and target nodes, respectively. These numbers must follow a strict sequence, e.g. the two nodes connecting at $t = 1$ must be node 1 and 2. An increase in this sequence reflects the inclusion of a new node. The fourth column is numeric and measures the *increment* to the weight of the tie. Currently, all the functions in *tnet* only interpret this when the *incremental* value is either +1 or -1. In other words, +1 refers to creation or reinforcement of the tie, whereas -1 reflects a weakening or severing of the tie. The fifth column represents the total number of nodes that have been active in the network at a given time. This column is important as it allows for inclusion of isolates in the network (i.e., nodes not represented in columns 2 and 3). Finally, nodes can join and withdraw from the network at specific times. This is signalled by a self-loop (i.e., when the second and third column have the same identification number). The fourth column controls whether it the node is joining (+1) or withdrawing (-1).

The sample network shown in Figure 17 has the following structure:

```
2008-09-12 13:45:00   1   2   1   2
2008-09-12 13:46:31   1   2   1   3
2008-09-12 13:49:27   1   2  -1   3
2008-09-12 13:58:14   1   3   1   3
2008-09-12 13:52:17   3   1   1   3
2008-09-12 13:54:26   3   2   1   3
```

Table 10: Format for longitudinal data: a $T \times 5$ `data.frame` object in *R*.

In analogy with the functions that help the user comply with the criteria for creating the appropriate weighted edgelist, there is also a function that transforms longitudinal data into the format outlined above. The `as.longitudinal`-function does two operations. First, it adds the missing columns if either the fourth and fifth or just the fifth column are missing in the dataset, e.g. if only the time, the creator, and target are known. If this is the case, it assumes that all ties carry a weight of 1 (fourth column), and that there are no isolates in the network. Second, it reorganises the identification numbers so that they follow a strict sequence. If the identification numbers of nodes were not entered in a order strict, the function will reassign the identification numbers and give a warning message. A table containing the association between the original and the new identification numbers will be attached to the data object as an attribute called `order`. This table is used by the other functions in *tnet* to rearrange node attribute data when this type of data is included in the analysis. Table 11 shows how the `as.longitudinal`-function works on an object where the fourth and fifth column are missing and the identification numbers are not in sequence.

A main issue with longitudinal network datasets is that the severing of ties is often not recorded. For example, we know that an email was sent, but we do not know when the tie between the sender and receiver is broken. There are several ways

```
>data
 2008-09-12 13:45:00   1   5
 2008-09-12 13:46:31   4   2
 2008-09-12 13:49:27   1   2

>transformed.data <- as.longitudinal(data)

>transformed.data
 2008-09-12 13:45:00   1   2   1   2
 2008-09-12 13:46:31   3   4   1   4
 2008-09-12 13:49:27   1   4   1   4

>attributes(transformed.data)$order
   1   1
   5   2
   4   3
   2   4
```

Table 11: How the `as.longitudinal`-function works on an object called `data`. Lines starting with `>` refer to commands in *R*. The first command displays the object. This object is then the input of the `as.longitudinal`-function. The third command displays this output of the function, `transformed.data`. The last command displays the attribute `order`, which is the key between the old and new node identification numbers.

of estimating the severing of ties, such as the introduction of a smoothing window (e.g. Kossinets and Watts, 2006; Panzarasa et al., 2009). This method assumes that social ties are severed if there has been a prolonged period of time with no interaction (i.e. the length of the window). A function that allows for the severing of ties after a set amount of time and adds negative ties (i.e., where the fourth column is -1) in longitudinal datasets is `add_window_to_longitudinal_data`.

Since longitudinal datasets are richer in detail than static weighted ones, the function `longitudinal_data_to_edgelist` was created to transform a longitudinal network into a static network. This function is particularly useful when studying the evolution of static network measures over time (e.g. Kossinets and Watts, 2006; Panzarasa et al., 2009). A possible application of `add_window_to_longitudinal_data` and `longitudinal_data_to_edgelist`-functions to the Online Social Network (see

Chapter 1) is illustrated in Figure 18. The Figure shows both cumulative networks (i.e., ties are never severed) and networks constructed with smoothing windows of 2, 3, and 6 weeks. Both panels in Figure 18 highlight the vulnerability of network measures to the use of a smoothing window. Panel a suggests that there is only a small core of users that actively use the virtual community at the end of the observation period. An analysis of the cumulative network at that point would be heavily influenced by the majority of users that only used the network in the first 6 weeks, and would not reflect the current activities that are occurring in the community. This could bias network measures and, ultimately, the analysis. Panel b shows the evolution of one possible measure, the clustering coefficient (Chapter 2). In particular, the clustering coefficient measured on the active core fluctuates greatly and is mostly below the value found in the cumulative network.

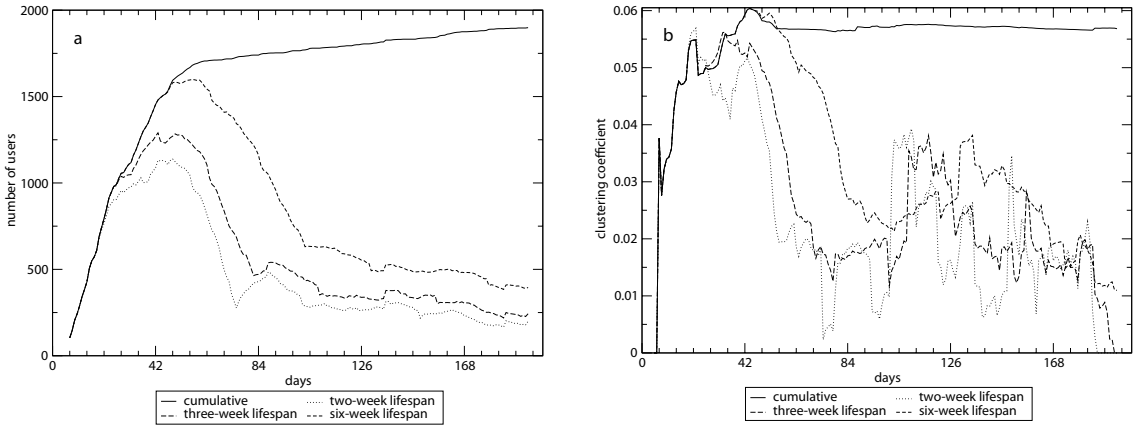


Figure 18: Impact of smoothing windows on network measures in the Online Social Network: a) number of nodes with degree greater than 0; b) the clustering coefficient (Eq. 2). These figures are based on Panzarasa et al. (2009).

5.2 Weighted network functions

For weighted networks, the main functions are transitivity (`clustering_w`), the weighted rich-club effect (`weighted_richclub`), and centrality measures (`degree_w`,

`closeness_w`, and `betweenness_w`). The transitivity and the weighted rich-club effect functions calculate the coefficients proposed in Chapters 2 and 3, respectively. However, since all the weighted centrality measures have not been discussed in detail in this thesis, they will be briefly introduced here. They are generalisations of the three measures that Freeman (1978) originally proposed, namely degree, closeness and betweenness. The original measures were formalised as follows:

$$C_{\mathrm{D}}(i) = \sum_{j} x_{ij} \tag{13a}$$

$$C_{\mathrm{C}}(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1} \tag{13b}$$

$$C_{\mathrm{B}}(i) = \frac{g_{jk}(i)}{g_{jk}} \tag{13c}$$

where i is the focal node, j and k are all the other nodes in the network, x is the adjacency matrix with $x_{ij} = 1$ if node i is tied to node j and 0 otherwise, $d(ij)$ is the shortest distance between the focal node and node j, g_{jk} is the number of shortest paths between node j and node k, and $g_{jk}(i)$ is the number of those paths that go through the focal node (Freeman, 1978).

These measures have been extended to weighted network as follows. First, "degree" in weighted networks is often taken as the sum of weights (see Chapter 3 and Barrat et al., 2004; Caldarelli, 2007; Newman, 2004a; Opsahl et al., 2008), and labelled node strength. This property has been formalised as follows:

$$C_{\mathrm{D}}^{\omega}(i) = \sum_{j} \omega_{ij} \tag{14}$$

where ω is the weighted adjacency matrix, in which w_{ij} is greater than 0 if node i is tied to node j, and the value is the weight of the tie, which represents the strength of the relation between the two nodes. The function `degree_w` calculates both this

and Freeman's (1978) version of degree centrality.

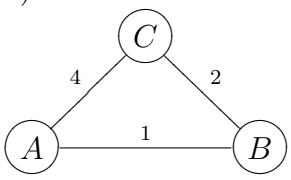
Second, both the closeness and betweenness measures rely on the calculation of shortest distances in the network (Wasserman and Faust, 1994). Closeness centrality is the inverse of the sum of the shortest distances from a focal node to all the other reachable nodes in the network, whereas betweenness centrality measures the number of shortest paths among all other nodes that the focal node is part of. Therefore, a first step towards extending these measures to weighted networks is to generalise how shortest distances are defined.

There has been great interest in distances among nodes in networks, and in particular the shortest distance between two nodes (Dijkstra, 1959; Katz, 1953; Peay, 1980; Yang and Knoke, 2001). A key assumption used when analysing the shortest distances is that intermediary nodes increase the cost of the interaction between two nodes. First, the more intermediary nodes, the more time the interaction between two nodes takes. Second, the intermediaries can distort information or delay the interaction between other nodes (Simmel, 1950; Burt, 1992). Since all ties have the same weight in binary networks, the shortest path for interaction between two nodes is through the smallest number of intermediary nodes. However, a complication arises when the ties in a network do not have the same weight attached to them. For instance, diseases are more likely to be transferred from one person to another if they have frequent interaction (Valente, 1995). This has implications for the diffusion in networks if a backbone of strong ties exists. In fact, it has been shown that nodes heavily involved in the network activity tend to be strongly connected with one another (see Chapter 3 and Opsahl et al., 2008). Therefore, diffusion is likely to occur quicker among these nodes than would be the case if all ties carried the same weight.

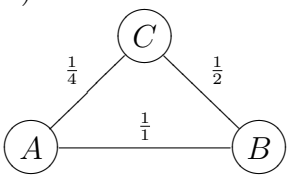
There have been several attempts to calculate shortest distances in weighted networks (Katz, 1953; Dijkstra, 1959; Peay, 1980; Fisek et al., 1992; Yang and Knoke,

2001). Dijkstra (1959) proposed an algorithm that finds the path of least resistance. This algorithm was defined for networks where the weights represented costs of transmitting, e.g. number of miles for distance calculations in GPS devices or time to transfer Internet traffic between routers (Ash, 1997). The distance between two nodes is the sum of the costs attached to each tie. Newman (2001c) applied Dijkstra's algorithm to a scientific collaboration network by inverting the positive weights in the network (see Figure 19a and b). Thus, high values represented weak or costly ties, whereas low values represented strong or cheap ties. For example, if two nodes are connected through a tie that has a weight twice as large as the weight of the tie connecting another pair of nodes, the distance between the former nodes is considered to be half the distance between the latter (e.g., the AC-dyad and the BC-dyad). The `distance_w`-function calculates the weighted distance matrix based on this method (see Figure 19c)[24].

a)



b)



c)

	A	B	C
A		0.75	0.25
B	0.75		0.50
C	0.25	0.50	

Figure 19: Example of distance in a weighted network. a) A sample network with positive tie weights. b) The sample network with inverted tie weights. c) The weighted distance matrix: the sum of inverted tie weights attached to the path of least resistance. Although node A and node B are directly connected, the path through node C carries a lower cost ($A \rightarrow C \rightarrow B = 0.75$ whereas $A \rightarrow B = 1$).

By using a distance matrix computed with Dijkstra's algorithm, Newman (2001c) extended the closeness measure by taking the inverse row sums of this matrix. He applied both Freeman's closeness measure and the generalised one to a coauthorship network. He found that different authors had the highest closeness score in the binary and weighted assessment. This highlights the importance of considering the

[24]This function relies on the *RBGL* package that is freely available through *R*'s repositories, *CRAN*.

weights, and suggests that the binary measure is not a good proxy of the weighted one. This algorithm is implemented in *tnet* as the `closeness_w` function.

In the case of betweenness, if the researcher assumes that the flow in the network occurs over the paths that Dijkstra's algorithm identifies, then it is possible to use this algorithm to find the nodes that funnels the flow in the network. I have extended Freeman's (1978) betweenness measure by counting the number of paths found by Dijkstra's algorithm on a weighted network instead of the number found on a binary network. This extension is implemented as the `betweenness_w` function[25].

In addition, there are two functions that create random weighted networks. First, `rg_w` takes a set of properties, and according to these properties it produces a random network. These properties included the number of nodes and ties, the range of weights, and whether the resulting network should be directed. Second, `rg_reshuffling_w` takes an observed network and randomises certain properties, such as the creator node, target node, or the location of the weights. In fact, the latter function is used by `weighted_richclub` when creating corresponding random networks (see Chapter 3).

5.3 Longitudinal network functions

For longitudinal networks, the main function implemented in *tnet* is the method presented in Chapter 4. This function is called `tnet.growth.clogit` and probes the underlying mechanisms that guide nodes' choices in where to direct their ties. This function is general and not limited to the analysis conducted in Chapter 4 or online communication data. The only requirement is that the data represents a directed network where the exact sequence in which ties were formed is known. The data can be both binary and weighted.

[25]This function relies upon the `sp.between`-function in the *RBGL* package. This function currently does not find more than one shortest path connecting two nodes even if an equally short path exists. This limitation affects the `betweenness_w`-function.

As outlined in Chapter 4, multiple mechanisms are thought to guide the formation of ties. These mechanisms are based on the network as well as information about the nodes. Each of these mechanisms can in turn be operationalised as a set terms or independent variables. These terms are given through a vector to the `tnet.growth.clogit` function – at least one term must be included. The network terms that are implemented are illustrated in Figure 20. Node i creates a tie to a possible target node h (the focal dyad, dashed line). The k_1, k_2, and k_3 nodes are the ones that node i is already tied to, and the l_1, l_2, l_3, and l_4 nodes are the ones that node i is not tied to, but are already tied to h.

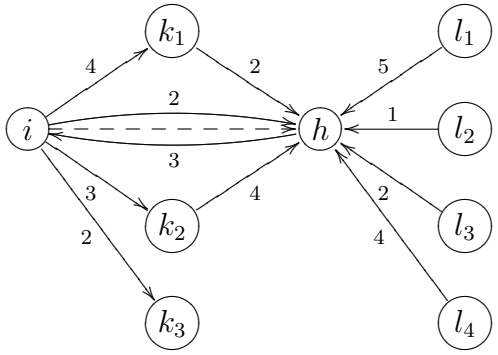


Figure 20: Example of network mechanisms: a focal dyad ($i \rightarrow h$) and the surrounding other nodes before a possible tie is created.

First, three terms are included to account for the increased likelihood of forming a tie if two nodes share common contacts (Heider, 1946). The basic term is the number of node i's contacts that are tied to node h. This term is named `"triadic.closure"`. In the sample network in Figure 20, this term is equal to 2 for the focal dyad. However, as shown in Chapter 2, clustering can be generalised to weighted networks. This generalisation requires the use of a triplet value, which is based on the weights of the two ties that indirectly connect a node to another. Currently, the `tnet.growth.clogit`-function includes two triplet terms where the triplet values are based on the "minimum" (`"triadic.closure.w.min"`; equal to 5 in the example) and "geometric mean" (`"triadic.closure.w.gm"`; equal to 6.29 in

the example) methods. These two methods are also illustrated by Figure 14a and b, respectively.

Second, two terms are included to measure the effects of popularity in networks (Dorogovtsev and Mendes, 2003). The first term is the in-degree of nodes. This term has simply been named `"indegree"`. In this example, this would be equal to 6. In weighted networks, this term can be extended to the in-strength of a node, which is the sum of weights attached to the ties terminating a node (`"instrength"`; equal to 18 in this example).

Third, terms are included to test reciprocity (Gouldner, 1960; Plickert et al., 2007). The simplest is a dummy term that is equal to 1 if a tie exists from node h to node i, and 0 otherwise (`"reciprocity"`). This term can also be extended to weighted networks to the weight of the tie from node h to node i (`"reciprocityw"`). Since a tie is present from node h to node i with a weight of 3 in the example, these terms would be equal to 1 and 3, respectively.

Fourth, ties can be reinforced in weighted networks (Krackhardt, 1992). Therefore, the weight of the tie from node i to node h can be included (`"reinforcement"`) to account for the increase in likelihood of directing a tie towards another node that already have been contacted. In the example, this would be equal to 2.

Moreover, the effect of node similarity by sharing or having a similar node attribute on the likelihood of forming a tie can be included through the inclusion of demographic (homophily) or positional (focus constraints) attributes. The node attribute must be transformed into dyadic terms. Two methods have been programmed in `tnet.growth.clogit`-function to do this process. The first method is to create a dummy term that is equal to 1 if the two nodes have the same value of the attribute, and 0 otherwise. This method can be used for both nominal and ordinal attributes. The second method is specifically designed for ordinal attributes. It takes 1 minus the standardised difference between the values of the attributes for

the two nodes. For example, if people's age is known in a social network of 30 to 40 year-old individuals, then the first term would be equal to 1 if two people (i and h) were of the same age ($age(i) = age(h)$), whereas the second term would be equal to: $1 - \frac{|age(i)-age(h)|}{40-30}$. Terms based on these two methods are included by loading the node attribute as separate vectors in the R-session, and adding the name of these vectors prefixed with `same.` or `simi.` to the list of terms, respectively. For example, if the age of the people is loaded as a vector named `ageofperson`, the term to study the effect of similar age on the likelihood of a tie being formed would be `"simi.ageofperson"`.

However, if these two methods are not sufficient for creating a dyadic term, then the `tnet.growth.clogit`-function also allows for the inclusion of a user-created matrix where a dyad term is explicitly defined. This should be an $N \times N$ matrix with, for example, the log of the number of miles separating two nodes (e.g. Sorenson and Stuart, 2001). If this matrix was named `logmiles`, then `"dyad.logmiles"` should be added to the list of terms to study the impact of geographical closeness on tie formation.

The `tnet.growth.clogit`-function can easily be extended by the inclusion of additional terms (or independent variables). The simplest way to do this is to output a table with all the observations and variables instead of the results from a regression[26]. Then the researcher can add variables to this table, and then run a regression in either *R* or other statistical programmes[27]. For example, if interaction terms are to be included, this can easily be done using this method. Moreover, a researcher with basic knowledge of *R* programming can also alter the source code. Each term is coded as a module and additional modules can easily be inserted in the

[26]This is done by setting the switch `regression` to `FALSE`.

[27]The `clogit` command in *Stata* is able to run a conditional logistic regression (StataCorp, 2007). The easiest way to export the regression table from *R* to *Stata* is to use the *foreign*-package's `write.dta` function, e.g. `write.dta(output, file="c:/statafile.dta")`, where `output` is the regression table and `c:/statafile.dta` is the location of the file.

code. The supporting website contains the specific details for inserting user-created modules[28].

In addition to the `tnet.growth.clogit`-function, *tnet* also includes a function to randomise a longitudinal network, `rg.longitudinal`. This function is flexible and allows the random network to be constrained in a number of ways. First, either creator or target nodes can be maintained. This guarantees that the out-degree or in-degree distributions are maintained at every t. Second, it can keep the size of the network invariant by maintaining either the available nodes in the network (regardless of whether they are connected) or the number of connected nodes at every t. Third, it can maintain the weight distribution at every t. It does so by finding the duplication of ties (reinforced ties) in the network, and replicating a randomised tie when the observed tie was reinforced. For example, if a tie is formed between two nodes at $t = 5$ and reinforced again at $t = 9$ in the observed network, then the tie at $t = 5$ is randomised and the tie at $t = 9$ is equal to the randomised one at $t = 5$.

5.4 Contribution to the literature

The software package *tnet* aims to help researchers and practitioners in two ways. First, it provides a simple tool for conducting a structural analysis of weighted networks, and investigating the mechanisms underpinning tie generation in longitudinal networks. This reduces the difficulty of taking the weights of ties and evolution into consideration, and has the potential of greatly lowering the cost of performing these types of analysis. As a result, more researchers and practitioners can apply the methods presented in this thesis to conduct network analysis.

Second, *tnet* makes available a platform that can handle both weighted and longitudinal network datasets. In so doing, researchers aiming to generalise network

[28] `http://opsahl.co.uk/tnet/content/view/45/25/`

measures to these types of networks do not have to programme basic functions. Instead, they can focus on the new measures. For example, to study a new network growth mechanism in the evolution of a longitudinal network, researchers only need to write a small module to the `tnet.growth.clogit`-function[29]. This gives researchers the opportunity to save time and reduce their effort. By lowering the cost of generalising measures for weighted and longitudinal networks, *tnet* might help drive this process.

tnet can be downloaded at `http://opsahl.co.uk/tnet/` and hopefully soon on the *CRAN*-servers (Comprehensive R Archive Network). This will increase its visibility and enable wider dissemination.

5.5 Conclusion and discussion

tnet represents a step forward towards incorporating two aspects into mainstream network analysis, namely weight of ties and network evolution. First, the software package enables researchers to conduct a structural analysis of weighted networks. This analysis is not limited to a description of weighted networks (e.g., Panzarasa et al., 2009), but can also be used for investigating the impact of network structure on performance (Ahuja, 2000; Panzarasa and Opsahl, 2007). The importance of using the richness encoded in the weight of ties was highlighted by Newman (2004b) who showed that different authors had the highest closeness score when Freeman's (1978) binary measure and a generalised version (Newman, 2001c) were applied to a coauthorship network.

Second, *tnet* provides a simple and methodologically sound framework for studying different growth mechanisms in both binary and weighted longitudinal networks. This type of analysis can produce new findings. For example, it is surprising to find that the number of common friends does not significantly increase the probability

[29] `http://opsahl.co.uk/tnet/content/view/45/25/`

of forming a tie in the online social network used in Chapter 4. By making the function available, this has the potential of enabling practitioners to make better choices and formulate appropriate strategies and policies. In particular, it might help to calibrate algorithms like the ones used by Facebook's service People You May Know[30].

The software package is not without limitations. First, there are many more network measures that have already been generalised to weighted networks, but not yet included in the package, e.g. the betweenness measure by Freeman et al. (1991). By incorporating these measures, the relevancy of *tnet* would undoubtedly increase. Second, even though functions have been checked, we cannot rule out bugs. Third, this package is meant to become a truly open-source project with many contributors. However, we have not yet had the opportunity to invite others to join us as it has just become public.

The already existing functions are not without limitations either. For example, networks analysed by the `tnet.growth.clogit`-function cannot contain negative ties, i.e. the weakening or severing of ties. Currently, we are working on including this feature. The incorporation of the `add_window_to_longitudinal_data`-function forms an integral part of this effort, which allows for the inclusion of a smoothing window to a longitudinal dataset. Moreover, we are also adding the option of including specific node-level variable. For example, if the gender of people in a social network is known, a dummy variable signalling whether the creator node is female or male could be included in the regression.

[30] `http://blog.facebook.com/blog.php?post=15610312130`

6 Concluding Remarks

We are living in an interconnected world where people can make use of technologies to expand their personal networks beyond the boundaries that existed just a decade ago. This has prompted interest in networks from a wide range of disciplines, such as sociology and psychology as well as statistical physics and mathematics. There has also been a surge in the development of a diverse range of methods that can be applied to the study of a variety of networks, from neural networks, to social networks. Most of these methods are only applicable to single snapshots of the binary network structure. This is a major limitation as, in most empirical network datasets, ties can be differentiated by attaching a weight to them and are formed, reinforced, weakened, and severed over time. By discarding the weights, the analysis is limited to the presence or absence of ties (Freeman, 1978). Moreover, by not knowing or recording the evolution of the network, the difficulty of modelling growth mechanisms increases (Snijders, 2002; Wasserman and Pattison, 1996).

The chapters within this thesis represent a step forward for the analysis of weighted and longitudinal networks. In Chapter 2, we proposed a generalisation of the clustering coefficient to weighted networks. The clustering coefficient examines the tendency of nodes to form triangles. Often when the coefficient is applied to a weighted network, the network is first made binary by using a subjective cut-off: ties with weights above the cut-off are set to present, whereas ties with weights below are removed. This reduces the richness of the data as some ties are removed and the remaining ones cannot be differentiated. Instead of changing the data, the clustering coefficient should be generalised to take into account weights. Chapter 2 was devoted precisely to develop a generalised coefficient.

The second project explored associations between prominence and control over the strongest ties in three real-world networks, namely the US airport network,

a scientific collaboration network, and an online social network (Chapter 3). We build on the topological rich-club perspective that assesses whether the highly connected nodes (the prominent ones) form a club with more ties than expected by chance (Colizza et al., 2006). This framework was extending in three ways. First, we explored multiple definitions of prominence. This enabled us to detect novel and different results. Second, instead of limiting the analysis to the network topology, we examined whether the prominent nodes shared stronger ties than we would expect by chance. By exchanging the strongest ties, the prominent nodes secure control over the majority of resources flowing in the network. Third, in the rich-club framework, the coefficient obtained for the observed networks was compared to the average of coefficients found on a large number of random networks. When there are few prominent nodes, the coefficients obtained for the random networks might vary considerably. In fact, sometimes a striking result might be replicated in a non negligible number of the random networks. Therefore, we measured the 95% confidence interval of the coefficients found on the random networks. This enabled us to test whether or not the observed coefficient was replicated in a large proportion of the random networks. If the coefficient was above or below the interval, we argued that the prominent nodes preferentially directed their strongest ties towards or away from each other, respectively.

In Chapter 4, we offered a new approach to the study of the evolution of networks. Instead of limiting ourselves to conventional single snapshots of the network structure, we proposed to apply a regression framework often used in epidemiological studies to investigate the evolution of networks where the exact chronological order of ties is known. We applied this framework to an online social network and tested six growth mechanisms that might guide people's communication choices. These were: triadic closure, preferential attachment, reciprocity, homophily, focus constraints, and reinforcement. Most of the results were in line with expectations;

however, we found that the number of common contacts (triadic closure) was not a significant predictor of future ties in a multivariate analysis. This might be a reflection of the one-to-one online communication where an individual's contacts do not observe each other as is the case in offline social settings. Moreover, we found that popularity (preferential attachment) was mitigated when tested with other mechanisms, and further mitigated when ties could be reinforced (i.e., when considering the weighted network). These findings have critical implications for understanding the structure and function of networks.

It is my hope and intention that the work within this thesis has an impact on the community of researchers interested in networks. Currently, there is a lack of software programmes that can handle weighted and longitudinal networks. Therefore, the fourth project was devoted to providing researchers with a platform called *tnet* for easily conducting an analysis of these types of networks. Functions to calculate the methods proposed in the previous chapters as well as others proposed in the literature (e.g., the generalisation by Newman, 2001c, of Freeman's, 1978, closeness measure) have been programmed in the statistical software *R*. These functions have been incorporated into software package. Moreover, this package is now a publicly available open-source package. Other researchers will thus have the opportunity to easily implement generalisations of measures to weighted and longitudinal networks. In turn, this might prompt the development of interesting new measures.

The work within this thesis forms part of a wider research agenda concerned with the development of more sophisticated methods for analysing network data. Due to the fact that most of the existing methods are often defined for binary static networks only, it is necessary to reduce data so that it fits these methods. In this process, some of the richness contained within the data is removed. While there is abundance of network features that can be analysed in conventional binary static networks, this thesis covered only a few structural properties of weighted networks, and only

one method for studying networks' evolution. Thus, two areas of future research is concerned with the development of methods of studying weighted networks and methods of analysing networks where the exact sequence of ties is known.

In addition, there are other directions within this research agenda. First, directed networks have been an integral part of methods developed by sociologists. However, the same is not the case for most methods developed by physicists (Albert and Barabási, 2002). Thus, future research is likely to be concerned with this issue as the two disciplines within network research converge (see Chapter 1 for more details on the different groups of network researchers). Second, two-mode networks, such as the scientific collaboration network used in Chapter 3, are often projected onto one-mode networks. This introduces a number of biases in the network, which might invalidate the results. For example, the clustering coefficient (Chaper 2) of the one-mode projection of a randomly reshuffled two-mode network is generally higher than the coefficient found in a random one-mode network with the same number of nodes and ties as the one-mode projection (Newman, 2001b). This is due to the fact that the two-mode structure introduces clusters in the one-mode projection. Although some elements of the two-mode structure can be maintained by creating a weighted one-mode network (Newman, 2001c)[31], it would be of interest to redefine the clustering coefficient for two-mode networks. In turn, this might yield a coefficient with less biases. Similarly, it might be more appropriate to reshuffle the two-mode structure before projecting it onto a one-mode network when applying the weighted rich-club effect as suggested in Section 3.2.1. Thus, two main areas of future research within this research agenda is involved with extending and developing methods for analysing datasets where ties are directed and where the two-mode structure is maintained.

Furthermore, there are an endless number of possible empirical applications of

[31]For more details: `http://toreopsahl.com/2009/05/01/projecting-two-mode-networks-onto-weighted-one-mode-networks/`

the methods proposed within this thesis. In fact, a key element within all the methods is generality and the ability for researchers to tune the methods. More specifically, conditioned on the context in which data is collected and defined, the clustering coefficient enables researchers to define triplet values in multiple ways, any prominence parameter and level can be used in the weighted rich-club effect, and the framework for analysing tie formation allows for multiple growth mechanisms to be included. As the topological rich-club effect has been applied to networks varying from the Italian interbank network (De Masi et al., 2006) to protein-protein interaction networks (Colizza et al., 2006), the methods proposed within this thesis can be applied to any network given the appropriate data being collected (i.e., tie weights for the weighted clustering coefficient and the weighted rich-club effect, and the exact sequence of ties for the evolution framework). In fact, the weighted rich-club has already formed part of a study on the world trade network (Zlatic et al., 2008).

References

Aboud, F. E., Mendelson, M. J., 1996. Determinants of friendship selection and quality: Developmental perspectives. In: Bukowski, W. M., Newcomb, A., Hartup, W. W. (Eds.), The Company They Keep: Friendship in Childhood and Adolescence. Cambridge University Press, New York, NY, pp. 87–112.

Ahuja, G., 2000. Collaborative networks, structural holes, and innovation: a longitudinal study. Administrative Science Quarterly 45, 425–455.

Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Reviews of Modern Physics 73, 47–97.

Albert, R., Jeong, H., Barabási, A. L., 1999. Diameter of the world-wide web. Nature 401, 130–131.

Amaral, L. A. N., Guimerà, R., 2006. Complex networks: Lies, damned lies and statistics. Nature Physics 2, 75–76.

Amaral, L. A. N., Scala, A., Barthélémy, M., Stanley, H. E., 2000. Classes of small-world networks. Proceedings of the National Academy of Sciences 97, 11149–11152.

Ash, G., 1997. Dynamic Routing in Telecommunication Networks. McGraw-Hill, New York, NY.

Balcan, D., Erzan, A., 2007. Content-based networks: A pedagogical overview. Chaos 17 (026108).

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Barabási, A.-L., Jeonga, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T., 2002. Evolution of the social network of scientific collaborations. Physica A 311, 590–614.

Barrat, A., Barthélémy, M., Pastor-Satorras, R., Vespignani, A., 2004. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences 101 (11), 3747–3752.

Batagelj, V., Mrvar, A., 2007. Pajek: Program for Large Network Analysis: version 1.20. `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`.

Bernard, H. R., Killworth, P. D., Kronenfeld, D., Sailer, L. D., 1984. The problem of informant accuracy: the validity of retrospective data. Annual Review of Anthropology 13, 495–517.

Bernard, H. R., Kilworth, P. D., Evans, M. J., McCarty, C., Selley, G. A., 1988. Studying social relations cross-culturally. Ethnology 27 (2), 155–179.

Bollobás, B., 1998. Modern Graph Theory. Springer, New York, NY.

Boost Library developers, 2008. Boost Library. `http://www.boost.org`.

Borgatti, S. P., Carley, K., Krackhardt, D., 2006. Robustness of centrality measures under conditions of imperfect data. Social Networks 28 (2), 124–136.

Borgatti, S. P., Everett, M. G., Freeman, L. C., 2002. Ucinet for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard, MA.

Breslow, N. E., 1996. Statistics in epidemiology: The case-control study. Journal of the American Statistical Association 91 (433), 14–28.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Romkins, A., Wiener, J., 2000. Graph Structure of the Web. Preprint IBM Almaden.

Burt, R. S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R. S., 2005. Brokerage and Closure. An Introduction to Social Capital. Oxford University Press, New York, NY.

Burt, R. S., Lin, N., 1977. Network time series from archival records. Sociological Methodology 8, 224–254.

Butts, C. T., 2006. sna-package: Package for Social Network Analysis. R package version 1.4.

Butts, C. T., 2008. A relational event framework for social action. Sociological Methodology 38 (1), 155–200.

Butts, C. T., Handcock, M. S., Hunter, D. R., 2008. network: Classes for Relational Data. `http://statnet.org/`.

Caldarelli, G., 2007. Scale-Free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford, UK.

Coleman, J. S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Colizza, V., Flammini, A., Serrano, M. A., Vespignani, A., 2006. Detecting rich-club ordering in complex networks. Nature Physics 2, 110–115.

Cosslett, S. R., 1981. Efficient estimation of discrete-choice models. In: Manski, C. F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 467–492.

Cox, D. R., Hinkley, D. V., 1974. Theoretical Statistics. Chapman & Hall, London, UK.

Cross, R., Parker, A., 2004. The Hidden Power of Social Networks. Harvard Business School Press, Boston, MA.

Davis, J. A., 1970. Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. American Sociological Review 35 (5), 843–851.

De Masi, G., Iori, G., Caldarelli, G., 2006. Fitness model for the italian interbank money market. Physical Review E 74 (066112).

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271.

Doreian, P., 1969. A note on the detection of cliques in valued graphs. Sociometry 32 (2), 237–242.

Dorogovtsev, S. N., Mendes, J. F. F., 2003. Evolution of Networks. From Biological Nets to the Internet and WWW. Oxford University Press, New York, NY.

Drucker, P. F., 1993. The Post-Capitalist Society. HarperBusiness, New York, NY.

Ebel, H., Mielsch, L.-I., Bornholdt, S., 2002. Scale-free topology of e-mail networks. Physical Review E 66, 035103.

Erdős, P., Rényi, A., 1959. On random graphs. Publicationes Mathematicae 6, 290–297.

Erdős, P., Rényi, A., 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61.

Fararo, T. J., Sunshine, M., 1964. A Study of a Biased Friendship Network. Syracuse University Press, Syracuse, NY.

Feld, S. L., 1981. The focused organization of social ties. American Journal of Sociology 86, 1015–1035.

Fisek, H., Norman, R., Nelson-Kilger, M., 1992. Status characteristics and expectation states theory: a priori model parameters and test. Journal of Mathematical Sociology 16 (4), 285–303.

Foster, C. C., Rapoport, A., Orwant, C. J., 1963. A study of a large sociogram: Elimination of free parameters. Behavioural Science 8, 56–65.

Frank, O., Strauss, D., 1986. Markov graphs. Journal of the American Statistical Association 81, 832–842.

Freeman, L. C., 1978. Centrality in social networks: Conceptual clarification. Social Networks 1, 215–239.

Freeman, L. C., 1992. The sociological concept of "group": An empirical test of two models. American Journal of Sociology 98 (1), 152–166.

Freeman, L. C., 2004. The Development of Social Network Analysis: A Study in the Sociology of Science. BookSurge, North Charleston, SC.

Freeman, L. C., Borgatti, S. P., White, D. R., 1991. Centrality in valued graphs: A measure of betweenness based on network flow. Social Networks 13 (2), 141–154.

Friedkin, N. E., 1984. Structural cohesion and equivalence explanations of social homogeneiety. Sociological Methods and Research 12, 235–261.

Gouldner, A. W., 1960. The norm of reciprocity: A preliminary statement. American Sociological Review 25 (2), 161–178.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Guimerà, R., Mossa, S., Turtschi, A., Amaral, L. A. N., 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proceedings of the National Academy of Sciences 102, 7794–7799.

Guimerà, R., Sales-Pardo, M., Amaral, L. A. N., 2007. Classes of complex networks defined by role-to-role connectivity profiles. Nature Physics 3, 63–69.

Gulati, R., Gargiulo, M., 1999. Where do interorganizational networks come from. American Journal of Sociology 104, 1439–1493.

Hall, B. H., Jaffe, A. B., Tratjenberg, M., 2001. The NBER patent citations data file: Lessons, insights, and methodological tools. NBER Working Paper No. 8498.

Hallinan, M. T., 1974. A structural model of sentiment relations. American Journal of Sociology 80, 364–378.

Hallinan, M. T., Kubitschek, W. N., 1988. The effects of individual and structural characteristics on intransitivity in social networks. Social Psychology Quarterly 51, 81–92.

Handcock, M. S., 2003. Statistical models for social networks: degeneracy and inference. In: Breiger, R., Carley, K. M., Pattison, P. E. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp. 229–240.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Morris, M., 2003. statnet: Software Tools for the Statistical Modeling of Network Data. `http://statnetproject.org`.

Hansen, M. T., 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. Administrative Science Quarterly 44, 232–248.

Harris, R. G., 2001. The knowledge-based economy: intellectual origins and new economic perspectives. International Journal of Management Reviews 3, 21–40.

Heider, F., 1946. Attitudes and cognitive organization. Journal of Psychology 21, 107–112.

Hinds, P. J., Carley, K. M., Krackhardt, D., Wholey, D., 2000. Choosing work group members: Balancing similarity, competence, and familiarity. Organizational Behavior and Human Decision Processes 81 (2), 226–251.

Hinds, P. J., Kiesler, S., 1995. Communication across boundaries: Work, structure, and use of communication technologies in a large organization. Organization Science 6 (4), 373–393.

Holland, P. W., Leinhardt, S., 1970. A method for detecting structure in sociometric data. American Journal of Sociology 76, 492–513.

Holland, P. W., Leinhardt, S., 1971. Transitivity in structural models of small groups. Comparative Group Studies 2, 107–124.

Holland, P. W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association 76, 33–65.

Holme, P., Edling, C. R., Liljeros, F., 2004. Structure and time-evolution of an internet dating community. Social Networks 26, 155–174.

Hosmer, D. W., Lemeshow, S., 2000. Applied Logisitc Regression, 2nd ed. John Wiley & Sons, New York, NY.

Hunter, D. R., Goodreau, S. M., Handcock, M. S., 2008. Goodness of fit of social network models. Journal of the American Statistical Association 103, 248–258.

Jeong, H., Néda, Z., Barabási, A.-L., 2003. Measuring preferential attachment for evolving networks. Europhysics Letters 61, 567–572.

Kalmijn, M., Flap, H., 2001. Assortative meeting and mating: Unintended consequences of organized settings for partner choices. Social Forces 79 (4), 1289–1312.

Karlberg, M., 1997. Testing transitivity in graphs. Social Networks 19 (4), 325–343.

Karlberg, M., 1999. Testing transitivity in digraphs. Sociological Methodology 29, 225–251.

Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18, 39–43.

Katz, L., Proctor, C. H., 1959. The configuration of interpersonal relations in a group as a time-dependent stochastic process. Psychometrika 24, 253–287.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political Analysis 9 (2), 137–163.

Korte, C., Milgram, S., 1970. Acquaintance linking between white and negro populations: application of the small world problem. Journal of Personality and Social Psychology 15, 101–108.

Kossinets, G., Watts, D. J., 2006. Empirical analysis of an evolving social network. Science 311, 88–90.

Krackhardt, D., 1992. The strength of strong ties: The importance of philos in organizations. In: Nohria, N., Eccles, R. (Eds.), Networks and Organizations: Structure, Form, and Action. Harvard Business School Press, Boston MA, pp. 216–239.

Lazarsfeld, P. F., Merton, R. K., 1954. Friendship as social process: A substantive and methodological analysis. In: Berger, M., Abel, T., Page, C. (Eds.), Freedom and Control in Modern Society. Van Nostrand, New York, NY, pp. 18–66.

Lazega, E., 2001. The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership. Oxford University Press, Oxford, UK.

Leskovec, J., Kleinberg, J., Faloutsos, C., 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Levin, D. Z., Cross, R., 2004. The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. Management Science 50 (11), 1477–1490.

Long, J. S., Freese, J., 2003. Regression Models for Categorical Dependent Variables using Stata, rev. ed. Stata Press, College Station, TX.

Louch, H., 2000. Personal network integration: Transitivity and homophily in strong-tie relations. Social Networks 22, 45–64.

Luce, R. D., Perry, A. D., 1949. A method of matrix analysis of group structure. Psychometrika 14 (1), 95–116.

Luczkowich, J. J., Borgatti, S. P., Johnson, J. C., Everett, M. G., 2003. Defining and measuring trophic role similarity in food webs using regular equivalence. Journal of Theoretical Biology 220, 303321.

Lumley, T., 2008. survival: Survival analysis, including penalised likelihood. `http://cran.r-project.org/web/packages/survival/`.

Marsden, P. V., 1990. Network data and measurement. Annual Review of Sociology 16, 435–463.

Maslov, S., Sneppen, K., 2002. Specificity and stability in topology of protein networks. Science 296, 910–913.

MathWorks, Inc., 2007. Matlab Software: Version 7.4 (2007a). Natick, MA.

Matthew, 25:29. The Bible. BibleGateway.com: English Standard Version.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Structural Analysis of Discrete Data: with Econometric Applications. MIT Press, Cambridge, MA, pp. 197–272.

McPherson, J. M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444.

Merton, R. K., 1968. The Matthew effect in science. Science 159, 56–63.

Milgram, S., 1967. The small world problem. Psychology Today 2, 60–67.

Molloy, M., Reed, B., 1995. A critical point for random graphs with a given degree sequence. Random Structures and Algorithms 6, 161–180.

Monge, P., Rothman, L., Eisenberg, E., Miller, K., Kirste, K., 1985. The dynamics of organizational proximity. Management Science 31, 1129–1141.

Mood, A. M., Graybill, F. A., Boes, D. C., 1974. Introduction to the Theory of Statistics, 3rd ed. McGraw-Hill, Singapore.

Moody, J., 2004. The structure of social science collaboration network: Disciplinary cohesion from 1963 to 1999. American Sociological Review 69, 213–238.

Moreno, J. L., 1938. Who Shall Survive? Foundations of Sociometry, Group Psychotherapy, and Sociodrama. Nervous and Mental Disease Publishing Co., Washington, DC.

Newman, M. E. J., 2001a. Clustering and preferential attachment in growing networks. Physical Review E 64, 016131.

Newman, M. E. J., 2001b. Scientific collaboration networks: I. Network construction and fundamental results. Physical Review E 64, 016131.

Newman, M. E. J., 2001c. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64, 016132.

Newman, M. E. J., 2001d. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98, 404–409.

Newman, M. E. J., 2003. The structure and function of complex networks. SIAM Review 45, 167–256.

Newman, M. E. J., 2004a. Analysis of weighted networks. Physical Review E 70, 056131.

Newman, M. E. J., 2004b. Who is the best connected scientist? a study of scientific coauthorship networks. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (Eds.), Complex Networks. Springer, Berlin, Germany, p. 337370.

Newman, M. E. J., 2006. Finding community structure in networks using the eigenvectors of matrices. Physical Review E 76 (036104).

Newman, M. E. J., Park, J., 2003. Why social networks are different from other types of networks. Physical Review E 68, 036122.

Nordlund, C., 2007. Identifying regular blocks in valued networks: A heuristic applied to the st. marks carbon flow data, and international trade in cereal products. Social Networks 29 (1), 59–69.

Onnela, J.-P., Saramki, J., Hyvnen, J., Szab, G., Lazer, D., Kaski, K., Kertsz, J., Barabási, A.-L., 2007. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104, 7332–7336.

Opsahl, T., 2008. tnet: Software for Analysis of Weighted and Longitudinal networks, version 0.1.0. `http://opsahl.co.uk/tnet/`, see chapter 5.

Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J. J., 2008. Prominence and control: The weighted rich-club effect. Physical Review Letters 101 (168702).

Opsahl, T., Panzarasa, P., 2008. Mechanisms of network dynamics: Theoretical framework and methodological considerations. Proceedings of the 4$^{\text{th}}$ UK Social Network Conference University of Greenwich.

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Social Networks 31 (2), 155–163.

Panzarasa, P., Opsahl, T., 2007. Scientific collaborations in business and management: The effects of network structure on research performance. Proceedings of the 2007 Annual Meeting of the Academy of Management Philadelphia, PA.

Panzarasa, P., Opsahl, T., Carley, K. M., 2009. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. Journal of the American Society for Information Science and Technology 60 (5), 911–932.

Pareto, V., 1897. Cours d'èconomie politique. Macmillan, Paris, France.

Pastides, H., Kelsey, J. L., LiVolsi, V. A., Holford, T., Fischer, D., Goldberg, I.,

1983. Oral contraceptive use and fibrocystic breast disease with special reference to its histopathology. Journal of the National Cancer Institute 71, 5–9.

Pastor-Satorras, R., Vespignani, A., 2004. Evolution and Structure of the Internet. Cambridge University Press, New York, NY.

Pattison, P. E., Wasserman, S., 1999. Logit models and logistic regressions for social networks. II. multivariate relations. British Journal of Mathematical and Statistical Psychology 52, 169–194.

Pearson, M., Michell, L., 2000. Smoke rings: Social network analysis of friendship groups, smoking, and drug-taking. Drugs: Education, Prevention and Policy 7 (1), 21–37.

Pearson, M., West, P., 2003. Drifting smoke rings: Social network analysis and markov processes in a longitudinal study of friendship groups and risk-taking. Connections 25 (2), 59–76.

Peay, E. R., 1980. Connectedness in a general model for valued networks. Social Networks 2, 385–410.

Plickert, G., Côté, R. R., Wellman, B., 2007. It's not who you know, it's how you know them: Who exchanges what with whom? Social Networks 29, 405–429.

Powell, W. W., White, D., Koput, K. W., Owen-Smith, J., 2005. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110 (4), 1132–1205.

Price, D. J. d. S., 1965. Networks of scientific papers. Science 149, 510–515.

R Development Team, 2008. R: Version 2.7. R Foundation for Statistical Computing, Vienna, Austria.

Ramasco, J. J., 2007. Social inertia and diversity in collaboration networks. European Physical Journal ST 143, 47–50.

Ramasco, J. J., Gonçalves, B., 2007. Transport on weighted networks: when the correlations are independent of the degree. Physical Review E 76 (066106).

Ramasco, J. J., Morris, S., 2006. Social inertia in collaboration networks. Physical Review E 73 (016122).

Rao, A. R., Jana, R., Bandyopadhyay, S., 1996. A markov chain monte carlo method for generating random (0, 1)-matrices with given marginals. Sankhya A 58, 225–242.

Rapaport, A., 1953. Spread of information through a population with socio-structural bias. I. Assumption of transitivity. Bulletin of Mathematical Biophysics 15, 523–533.

Reagans, R., McEvily, B., 2003. Network structure and knowledge transfer: The effects of cohesion and range. Administrative Science Quarterly 48, 240–267.

Robins, G. L., Morris, M., 2007. Advances in exponential random graph (p^*) models. Social Networks 29 (2), 169–172.

Robins, G. L., Pattison, P. E., 2001. Random graph models for temporal processes in social networks. Journal of Mathematical Sociology 25, 5–41.

Robins, G. L., Snijders, T. A. B., Wang, P., Handcock, M., Pattison, P., 2007. Recent developments in exponential random graph (p^*) models for social networks. Social Networks 29 (2), 192–215.

Robins, G. L., Woolcock, J., Pattison, P. E., 2005. Small and other worlds: Global network structures from local processes. American Journal of Sociology 110, 894–936.

Scott, J., 2000. Social Network Analysis: A Handbook. Sage Publications, London, UK.

Serrano, M. A., 2008. Rich-club vs rich-multipolarization phenomena in weighted networks. Physical Review E 78 (026101).

Serrano, M. A., Boguñá, M., Vespignani, A., 2007. Patterns of dominant flows in the world trade web. Journal of Economic Interaction and Coordination 2, 111–124.

Simmel, G., 1950. The Sociology of Georg Simmel (KH Wolff, trans.). Free Press, New York, NY.

Simon, H. A., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

Skvoretz, J., 2002. Complexity theory and models for social networks. Complexity 8, 47–55.

Snijders, T. A. B., 1991. Enumeration and simulation methods for 0-1 matrices with given marginals. Psychometrika 56 (3), 397–417.

Snijders, T. A. B., 2001. The statistical evaluation of social network dynamics. Sociological Methodology 31, 361–395.

Snijders, T. A. B., 2002. Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure 3 (2), 361–395.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., Handcock, M. S., 2006. New specifications for exponential random graph models. Sociological Methodology 35, 99–153.

Snijders, T. A. B., Steglich, C. E. G., 2008. Models for analyzing dynamics of valued networks. Proceedings of the 4$^{\text{th}}$ UK Social Networks conference, University of Greenwich, Greenwich, UK.

Snijders, T. A. B., Steglich, C. E. G., Schweinberger, M., Huisman, M., 2007. SIENA: version 3.1. University of Groningen: ICS / Department of Sociology; University of Oxford: Department of Statistics.

Snijders, T. A. B., Steglich, C. E. G., van de Bunt, G. G., 2008. Introduction to actor-based models for network dynamics. Unpublished manuscript, available at `http://stat.gamma.rug.nl/siena_articles.htm`.

Soffer, S. N., Vázquez, A., 2005. Network clustering coefficient without degree-correlation biases. Physical Review E 71 (057101).

Solomonoff, R., Rapoport, A., 1951. Connectivity of random nets. Bulletin of Mathematical Biophysics 13, 107–117.

Sorenson, O., Stuart, T. E., 2001. Syndication networks and the spatial distribution of venture capital investments. American Journal of Sociology 106 (6), 1546–1588.

StataCorp, 2007. Stata Statistical Software: Release 10. StataCorp LP, College Station, TX.

Steglich, C. E. G., Snijders, T. A. B., Pearson, M., 2007. Dynamic networks and behavior: Separating selection from influence. Unpublished manuscript, available at `http://stat.gamma.rug.nl/siena_articles.htm`.

Travers, J., Milgram, S., 1969. An experimental study of the small world problem. Sociometry 32, 425–443.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J. M., 2000. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature 403, 623–627.

Uzzi, B., 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. Administrative Science Quarterly 42, 35–67.

Uzzi, B., Lancaster, R., 2004. Embeddedness and price formation in the corporate law market. American Sociological Review 69, 319–344.

Uzzi, B., Spiro, J., 2005. Collaboration and creativity: The small world problem. American Journal of Sociology 111, 447–504.

Valente, T., 1995. Network Models of the Diffusion of Innovations. Hampton Press, Cresskill, NJ.

Wang, P., Robins, G. L., Pattison, P. E., 2005. PNET: version 1.0. University of Melbourne, Melbourne, Australia.

Wasserman, S., Faust, K., 1994. Social Network Analysis. Cambridge University Press, Cambridge, MA.

Wasserman, S., Pattison, P. E., 1996. Logit models and logistic regression for social networks: I. An introduction to markov graphs and p^*. Psychometrika 61, 401–425.

Watts, D. J., 1999. Small Worlds. Princeton University Press, Princeton, NJ.

Watts, D. J., 2004. The "new" science of networks. Annual Review of Sociology 30, 243–270.

Watts, D. J., Strogatz, S. H., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–442.

Wellman, B., 1999. Living networked on and offline. Contemporary Sociology 28 (6), 648–654.

Wu, Z., Braunstein, L. A., Colizza, V., Cohen, R., Havlin, S., Stanley, H. E., 2006. Optimal paths in complex networks with correlated weights: The world-wide airport network. Physical Review E 74 (056104).

Wuchty, S., 2007. Rich-club phenomenon in the interactome of p. falciparum–artifact or signature of a parasitic life style? PLoS ONE 2007 (e355).

Yang, S., Knoke, D., 2001. Optimal connections: strength and distance in valued graphs. Social Networks 23, 285–295.

Zhou, S., Mondragon, R. J., 2004. The rich-club phenomenon in the internet topology. IEEE Communications Letters, 8, 180–182.

Zipf, G. K., 1935. The Psycho-Biology of Language: An introduction to dynamic philology. Houghton Mifflin, Boston, MA.

Zlatic, V., Bianconi, G., Díaz-Guilera, A., Garlaschelli, D., Rao, F., Caldarelli, G., 2008. On the rich-club effect in dense and weighted networks. arXiv:0807.0793.

Appendix

A Presented and Published Papers

During my PhD, a number of the projects I have worked on have been presented at conferences. The first presentation was on the structure of the online social network used in most of the chapters of this thesis. This was at the Social Network Analysis Forum 2005 (Centre for Criminology, Oxford University, Oxford, UK). Improved versions of this project have subsequently been presented at the 26$^{\text{th}}$ International Sunbelt Social Network Conference (Vancouver, Canada), the European Conference on Complex Systems 2006 (Saïd Business School, Oxford University, Oxford, UK), and the UK Social Network Conference 2007 (Queen Mary College, University of London, London, UK). This project has now resulted in a publication (see Panzarasa et al., 2009).

Second, the generalisation of the clustering coefficient presented in Chapter 2 was first presented at the Applications of Social Network Analysis 2006 (Institute of Mass Communication and Media Research, University of Zurich, Zurich, Switzerland). By taking into consideration the feedback received, a new version of the paper was presented at the 27$^{\text{th}}$ International Sunbelt Social Network Conference (Corfu, Greece) and the International Workshop and Conference on Network Science 2008 (NetSci'08; Norwich BioScience Institutes, Norwich, UK). This project has now resulted in a publication (see Opsahl and Panzarasa, 2009).

Third, the weighted rich-club effect presented in Chapter 3 has recently been accepted for publication by *Physical Review Letters*, one of the most influential journals in physics. This project was not presented at any conferences before publication (see Opsahl et al. (2008)).

Fourth, a project about the processes that govern online communication was

presented at both the 27th and 28th International Sunbelt Social Network Conference (Corfu, Greece; Tampa, FL, USA) as well as the UK Social Network Conference 2008 (University of Greenwich, London, UK) during which feedback was received that enabled us to change it to the present form (Chapter 4). A manuscript is currently in preparation.

Fifth, in addition to the projects mentioned in this thesis, I have also work on a project that examines the effects of a scientific collaboration network structure on scientific performance. A preliminary study of this network was presented at the 2nd Social Network Forum (Leeds University Business School, Leeds, UK) in 2006. In this presentation, performance was measured by using an institutional performance score. Since then, the project has been extended and split into two papers. The first paper analyses the global structure of the network and was presented at the American Sociological Association's 102nd Annual Meeting in New York (2007). The second paper focusing on performance was presented at the 2007 Annual Meeting of the Academy of Management in Philadelphia (2007) and the UK Social Network Conference 2007 (Queen Mary College, University of London, London, UK). Subsequently, the second paper has been updated by looking at the weighted citation score of authors instead of an institutional performance measure. The improved version has been presented at 28th International Sunbelt Social Network Conference (Tampa, FL, USA), Cass Business School's Workshop on Scientific and Managerial Knowledge 2008 (City University, London, UK), the International Workshop and Conference on Network Science 2008 (NetSci'08; Norwich BioScience Institutes, Norwich, UK), and the UK Social Network Conference 2008 (University of Greenwich, London, UK).

Sixth, I have also been part of a project related to inventor networks in emerging countries, and a qualitative project related to verifying whether online ties are a product of cognitive similarity. These two projects were presented at the 25th

Danish Research Unit for Industrial Dynamics (DRUID) Conference (Copenhagen Business School, Copenhagen, Denmark) and the UK Social Network Conference 2008 (University of Greenwich, London, UK), respectively.

B Appendix to Prominence and Control: The Weighted Rich-club Effect

B.1 Directed Weight reshuffle when prominence is defined in terms of degree

In Figure 7, for the sake of clarity, we only show the results obtained with the two randomisation procedures. We chose the first two models, the Weight reshuffle and the Weight & Tie reshuffle, as they are the ones that lead to the most randomised networks. However, when prominence is defined in terms of the node degree k, all three randomisation procedures introduced in Chapter 3 are considered to be appropriate since they all preserve the degree distribution $P(k)$ of the observed network. More specifically, in addition to the *Weight* and *Weight & tie* reshuffling procedures, the *Directed weight* reshuffling is also appropriate. This allows us to compare the results obtained with the three randomisation procedures on the empirical networks. In Figure 21 we report results obtained with all three reshuffling procedures. It shows that the three randomisation procedures give similar results in all networks under study. This strengthens the results reported in the chapter when prominence is defined in terms of node degree, and also provides support to the application of the Directed Weight reshuffle in the investigation of the weighted rich-club ordering when prominence is defined in terms of node strength and node average weight.

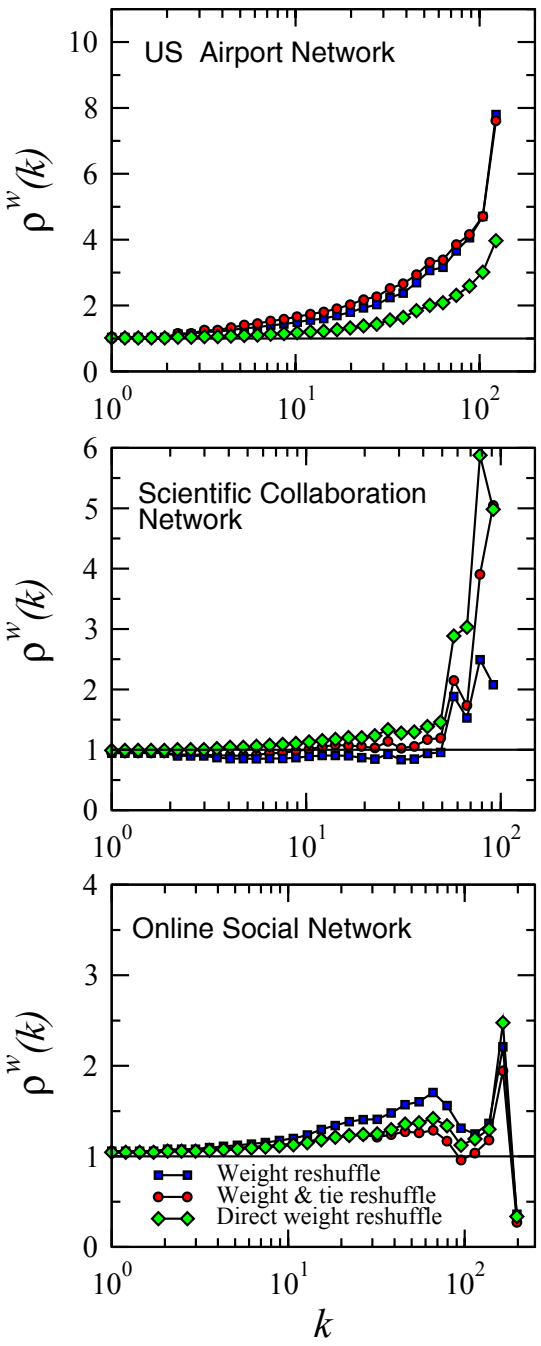


Figure 21: Weighted rich-club ordering among the most connected nodes in: the US Airport Network (top); the Scientific Collaboration Network (middle); and the Online Social Network (bottom). Results obtained with the three null models are shown. We have omitted the confidence interval of the random networks for the sake of clarity.

B.2 Weighted rich-club effect in the *Network Science* collaboration network

Here we report the results of the weighted rich-club effect on the *Network Science* collaboration network (Newman, 2006). To motivate the choice of additional prominence parameters, Figures 8a and b showed the scientists with high degree and strength, respectively, working on networks (theory and experiments). As shown in Figure 22, the smaller collaboration network displays the same behaviour observed in the larger overall scientific collaboration network: a marked positive trend in the topological ordering, as opposed to a random behaviour when the intensity of collaborations is considered.

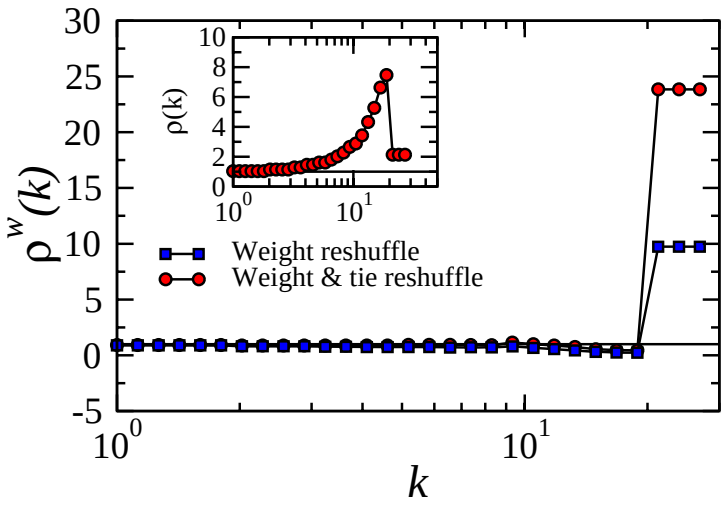


Figure 22: Weighted rich-club ordering among the most connected nodes in the *Network Science* collaboration network. The inset refers to the topological rich-club ordering.