Network Evolution and Interpersonal Dynamics: An Empirical Investigation of Online Communication

Pietro Panzarasa and Tore Opsahl
School of Business and Management
Queen Mary College, University of London
p.panzarasa@qmul.ac.uk; t.opsahl@qmul.ac.uk

Abstract

This research draws on longitudinal network data from an online community to examine patterns of social interaction and infer the mechanisms shaping interpersonal dynamics. The online community represents a prototypical example of a complex evolving network in which connections between users are established over time by online messages. Results indicate that the network topology is compact, densely connected, with a higher clustering than would be expected by chance, and dominated by a minority of users that preside over a disproportionately large amount of connections. The network follows a power-law scaling behaviour that is governed by preferential linking mechanisms. Connections between users are not established at random, but are more likely to originate from users that have already created many connections, and to be directed towards users that have already received many connections. The effects of preferential linking on interpersonal dynamics, however, are mediated by the degree to which users share the same demographic characteristics and affiliation to the same groups. We discuss the implications of the findings for research on evolving social networks. We also describe the managerial implications of our results for a host of applications, from information diffusion to the security and robustness of electronic information systems.

1 Introduction

The conceptualisation of many social systems as networks made of vertices (or nodes) joined together by edges (or connections) has long facilitated the study of the origins, evolution, and consequences of patterns of social interaction (e.g., see Smith-Doerr and Powell 2005, Wasserman and Faust 1994). Empirical and theoretical work on social networks has traditionally been an interdisciplinary endeavour, characterised by a rich research agenda concerned with a broad range

of social phenomena and crosscutting multiple levels of analysis. A substantial body of work has concentrated on the topological properties of networks, with a view to uncovering the global structural patterns that emerge from the ways in which individuals behave at a local level (e.g., Bernard et al. 1988, Fararo and Sunshine 1964, Foster et al. 1963). Milgram (1967) conducted one of the first empirical investigations of social networks, and his "six-degrees-of-separation" experiment is usually taken as evidence of the "small-world effect" hypothesis that most pairs of individuals in a population can be connected by a short chain of intermediate acquaintances, even when the size of the population is very large (Kochen 1989). Scholars have also been interested in the mechanisms governing network evolution and shaping interpersonal dynamics. Early studies have investigated how individuals may benefit from participating in locally dense clusters (Coleman 1988), for example by choosing new acquaintances that are already connected to current acquaintances, a process known as triadic closure (Davis 1970, Holland and Leinhardt 1970, Rapaport 1953). In addition, a longstanding tradition of research has focussed on the effects that sharing a significant demographic characteristic has on tie strength (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 network evolution, 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).

It has now become widely accepted that, while directly probing the network structure and functioning, many of the early studies of social networks suffered from two fundamental weaknesses (for a review, see Marsden 1990). The first one pertains to the much studied problem of "informant inaccuracy": whenever data collection relies on asking people for information, findings become highly sensitive to subjective bias (cf. the review article by Bernard et al. 1984). For most of the past fifty years, network datasets have been collected mainly through survey instruments, typically associated with the uncertainties arising from the inaccurate and subjective responses of subjects. For example, when interviewees were asked to name contacts and indicate the interaction strength, what was considered to be an "acquaintance" and a strong social interaction could differ considerably from one person to another (e.g., Bernard et al. 1988, Fararo and Sunshine 1964, Foster et al. 1963, Kochen 1989). A second problem associated with past research on networks is concerned with the size of the dataset. Survey instruments and direct observation methods are typically labour-intensive and onerous to administer, and therefore the size of most networks mapped turned out to be fairly limited, often comprising only a

few tens (e.g., Bernard et al. 1988) or hundreds (e.g., Fararo and Sunshine 1964) of people.

The lack of high quality and large-scale network datasets has delayed progress in the statistical modelling of the network topology and behaviour. In recent years, however, two major developments have prompted substantial advances in research on networks, and contributed to the birth of what has been called the "new science of networks" (Watts 2004). First, more accurate studies of large-scale social networks (e.g., Barabási et al. 2002, Newman 2001b) have been made possible by the advent of new technological resources, such as more powerful computers and the growing availability of electronic databases. Second, a noticeable breakdown of boundaries between the social and behavioural sciences on the one hand, and physics and applied mathematics on the other, has spurred scholars to make substantial progress in the study of networked systems. Many new theoretical developments have explored a number of network-related problems and proposed a variety of analytical tools for modelling the structure and behaviour of complex networks (for a recent update on network research, see Dorogovtsev and Mendes 2003).

These converging developments and circumstances have motivated many scholars to study large-scale statistical properties and capture in quantitative terms the organising principles encoded in the topology of a variety of complex networks. These principles, however, and their implications for the analysis of interaction patterns and interpersonal dynamics are not to be taken without caution. For instance, recent studies of the network of movie actors (Watts and Strogatz 1998), the network of the artists working in Broadway musicals (Uzzi and Spiro 2005), and scientific collaboration networks (Barabási et al. 2002, Newman 2001b) are paradigmatic examples of modelling efforts that, by using methodologies rooted in statistical physics, aim to characterise the statistical regularities observed in the large-scale topology of networks where vertices are indeed individuals. However, the extent to which an edge between two individuals, as defined in these studies, does represent a genuine instance of a social interaction is not an issue without ambiguities. For example, the appearance of two actors in the same movie may not necessarily imply that these individuals were involved in some form of interaction leading to a social relationship, or that their acquaintance extended beyond the boundaries of the artistic production. In addition, in the networks of movie actors (Watts and Strogatz 1998) and creative artists (Uzzi and Spiro 2005), the decision of interacting with others may not be made entirely by the individuals, but instead it can be delegated to a higher level, such as the casting director. Clearly, this may undermine the attempt to trace the origins of global topological properties to the decisions made by individuals at the local level. Moreover, for scientific collaboration networks findings can be biased by the fact that, due to well-established practices of some

research facilities, a long list of coauthors does not necessarily imply that among all of them there was actual exchange of ideas or any other form of social interaction (Newman 2001b).

As noted by a few scholars (e.g. Burt 2000, Smith-Doerr and Powell 2005, McPherson et al. 2001), an additional problem that has thwarted developments in traditional and more recent empirical studies of networks is the limited use of longitudinal data and the subsequent lack of emphasis on dynamics. With only a few exceptions (e.g., Barabási et al. 2002, Holme et al. 2004, Kossinets and Watts 2006, Powell et al. 2005, Uzzi and Spiro 2005), recent empirical work has been primarily concerned only with the static topological features of social networks, while the evolution of interaction patterns and their underpinning organising principles have been mostly overlooked. Yet, social networks are inherently evolving systems: over time, individuals join and leave the networks, and social relationships are created and severed (Banks and Carley 1996, Burt 2000, Snijders 2005). The extent to which individuals' choices at the local level dynamically affect the global network topology is largely an empirical matter that can only be investigated by using a longitudinal network dataset where the time at which individuals and relationships have been added to, or removed from, the network is explicitly available.

More recently, new research problems have been brought to the surface of the debate on social network evolution by a number of empirical studies that have investigated the structure of large-scale networks of many different kinds. The observation that a variety of networks exhibit unusual fat-tailed distributions of connections (e.g., Albert et al. 1999, Faloutsos et al. 1999) has prompted a surge of interest in understanding how networks evolve and such fat-tailed distributions are generated. One of the most influential attempts to capture and formalise the general organising principles and processes responsible for the way in which connections are established over time is a simple mathematical model of network growth, proposed by Barabási and Albert (1999), and based on the idea that "popularity is attractive" (Dorogovtsev and Mendes 2003). Inspired by this model, extensive studies have shown that many real-world networks evolve not at random, but with vertices connecting preferentially to other vertices that are already well-connected (e.g., see Jeong et al. 2003). However, despite the ubiquity of such preferential linking mechanism in a variety of non-social networks, it still remains unclear whether the attractiveness of popularity can also drive the evolution of social networks in which connections are not costless, information is not free, and social structure is expected to affect network growth (Jin et al. 2001, Watts 2004).

Recent empirical studies have indeed produced mixed results on the mechanisms governing social network evolution. While a few scholars have found evidence that individuals preferentially establish social relationships with other well-connected individuals (e.g., Barabási et al.

2002, Newman 2001a), other scholars have instead reported a tendency of the social network to evolve according to other local ordering principles, such as homophily and focus constraints (e.g., Hinds et al. 2000, Kossinets and Watts 2006, Louch 2000, Powell et al. 2005, Reagans 2005). In light of this, an important research problem that has thus far received scanty attention is whether, in situations where popularity shapes interpersonal dynamics, it does so in isolation or in combination with other principles of network growth (Banks and Carley 1996, Snijders 2005). A further complication arises from the fact that most social relationships, from acquaintance to business alliances, can be formalised as directed edges as they originate from a vertex and terminate at another one. Yet, in most empirical studies of evolving social networks, the assumption is often made that relationships are undirected (e.g., Powell et al. 2005, Uzzi and Spiro 2005), and that the mechanisms explaining how vertices make and receive new connections are the same (e.g., Barabási et al. 2002). If it is reasonable to conjecture that popularity may affect an individual's chance to *receive* further connections, nonetheless it is not clear how that could also affect the individual's tendency to *generate* new connections. Further empirical investigation of the mechanisms governing social network evolution is clearly needed.

With this paper, we take a first step in this direction. Drawing on recent theoretical and methodological advances in network science, our goal is to study the evolution of interaction patterns and uncover mechanisms of interpersonal dynamics. To this end, we empirically examine a dynamic communication network in which users of an online community communicate and engage in social interactions by sending or receiving online messages. Our use of online communication to study the underlying network of social relationships is supported by recent studies indicating that online communication serves as much social function as other kinds of social interaction, including face-to-face and telephone conversations (Wellman and Haythornthwaite 2002). The vertices of the network are the users, whereas edges between users are established by online messages. Edges thus originate from decisions to initiate a social interaction that are made independently by individual users at the local level. In turn, as these local decisions are dynamically formed and crystallised into a series of messages, they shape the global structure of the network and its evolution over time. We use the series of time-stamped messages to reproduce the steps through which the network is assembled over time as a result of the addition and removal of vertices and edges.

The idea of using electronic databases to uncover the large-scale statistical properties of communication networks is not new. Recent empirical studies have investigated network datasets drawn from phone records (e.g., Aiello et al. 2000), email log-files (e.g., Ebel et al. 2002, Kossinets and Watts 2006), and registers of users' activities in online communities (e.g., Holme et al. 2004).

The dataset we analyse in this paper belongs to the last category. However, unlike recent studies of online communities (e.g., Holme et al. 2004, Rothaermel and Sugiyama 2001), we examine online communication between users that belong to the same institution and are in spatial proximity with one another. A clear implication of this feature is that online communication is more likely to be strongly correlated with face-to-face interaction than would be the case if users were geographically distant (Monge et al. 1985, Wellman and Haythornthwaite 2002). In addition, unlike other recent studies of dynamic communication networks (e.g., Holme et al. 2004, Kossinets and Watts 2006), our work delves into the growth mechanisms and social processes that are responsible for the observed topological properties of the network (Banks and Carley 1996, Snijders 2005).

By making explicit use of the time at which each vertex joins the network and each edge is formed, and by taking the direction of edges into account, we study the evolution of interaction patterns and infer the distinctive ordering principles that govern interpersonal dynamics in an electronic environment. In so doing, this paper seeks to extend current research on complex interactive systems and inspire the development of network theories that place a special emphasis on dynamics. Our analysis also aims to provide a platform for drawing a host of practical implications for the management of online communities and, more generally, electronic information systems. The lack of a centralised control in most electronic forms of communication may suggest that managers are deprived of the appropriate measures for presiding efficiently over the flow of information and for promoting valuable communication (cf. Bolton and Dewatripont 1994). Yet this research sheds light on the opportunities for effective management and control that arise precisely as a result of the distributed nature of communication. By taking advantage of the principles that govern the structure and evolution of the network, managers can channel information in the right direction, make sure that it quickly reaches most individuals in the network, search and retrieve the information they need in a timely fashion, and protect the whole network from the negative global effects of individuals' misuse of communication.

We proceed as follows. First, we describe our research setting and how the longitudinal network dataset was created. Following this, we present results on network evolution. To this end, we examine a number of statistical properties characterising the network structure, and show how these properties evolve over time. We then investigate the growth mechanisms governing network evolution and describe their implications for interpersonal dynamics. The paper concludes with an interpretation of the findings and a discussion of their implications for theory and practice.

2 The Data

We study the network evolution of an online community in which the users are students at the University of California, Irvine. The community was aimed to sustain social interaction between students and help them enlarge their circles of friends. To join the community, each user was asked to create a profile providing a number of personal details (cf. Holme et al. 2004). Unlike email communication, the online community allowed each user's profile to be searched by others who then could make their decisions to communicate on the basis of the information offered by the profile. This included the user's demographic characteristics and details about his or her (online) popularity, such as the number of times the profile was visited, the user's list of friends, personal blogs, and forum postings. This feature makes the online community an appropriate research setting for uncovering network growth mechanisms as it enables us to test which pieces of information (e.g., users' popularity or demographic characteristics) guided individuals' choices of communication and the way connections were forged. The network dataset covers a period of 194 days, from April to October 2004. To create the longitudinal network dataset, we compiled a register of all online messages sent during the observation period, and for each message we recorded the time-stamp, sender and recipient.[1] For each user, we also collected the following attributes: age, gender, year of study, region of origin, marital status, year in college, and affiliation to academic unit.[2]

We analysed the longitudinal network dataset in two steps. First, at any time point we constructed the instantaneous cumulative network reflecting all the events that took place before that point, *since the beginning* of the observation.[3] The network is measured with a daily frequency: as time goes by and daily measures are recorded, the network will include one additional day until all 194 days are taken into account at the end of the observation period.[4]

[1]To ensure privacy protection, all individual identifiers, such as email and IP addresses were removed, and usernames were anonymised before we received the data. Moreover, the content of messages was not made available.

[2]Academic units include: Arts; Biological Sciences; Education; Engineering; Humanities; Information and Computer Science; Interdisciplinary Studies; Management; Physical Sciences; Social Ecology; and Social Sciences.

[3]Measurements start from day 7 as before that time statistics are poor.

[4]The choice of the appropriate sampling frequency reflects the trade-off between ensuring that sequential events are not wrongly classified as simultaneous (by using a sampling period that is not too large), and minimising the bias resulting from errors in time measurements (by using a sampling period that is not too small). We estimated that sampling for structural changes with a daily frequency produced a reasonable solution to this trade-off, by also taking into account the cyclic daily trend of the number of messages sent. In fact, on average this number shows a significant drop in the early hours of the morning, with the minimum level at 7:00am, whereas during the rest of the day it shows a constant increase until it reaches its peak at midnight. For this reason, we chose to sample every day at 7:00am instead of midnight so as to minimise the interruption of the ongoing flow of social

We also constructed the network by using smoothing windows of fixed widths. The width of a smoothing window determines which past events are taken into account to generate the instantaneous topology of the network at any point in time. More precisely, with a fixed window width w, the network at any time point t is taken to be the result of all events that took place *only* in the period of length w that ends at t. In our analysis, we use a smoothing window of 21 days. We also check that results are robust for $w = 14$ and $w = 42$.[5] This means that at any day t, the network reflects all the social interactions that took place only within the period including the preceding 13, 20 or 41 days, respectively, and terminating at the end of day t. Therefore, with smoothing windows the effective span of the data varies from 181, to 174, and 153 days. Moreover, like the cumulative network, the network created with smoothing windows is measured with a daily frequency.

Our analysis includes all users that sent or received at least one message. To ensure that our data do indeed reflect interpersonal communication, users who simply registered, but did not communicate were excluded from this analysis. Furthermore, two companies that gained access to the online community with the purpose of mass-communication were excluded. A total of $1,899$ users was recorded at the end of day 194 for the cumulative network. A directed edge is established from one user to another if one or more messages have been sent from the former to the latter. Therefore, each vertex has an in-degree k^I (i.e., the number of edges terminating

interaction.

[5]There are a number of smoothing methods for constructing a longitudinal network dataset (Cortes et al. 2003). More generally, the choice of w should be motivated by the analysis of which past events are relevant to the current state of the network (e.g., see Kossinets and Watts 2006). Too small or too large values of w will have the effect of, respectively, breaking ongoing social interactions into two independent sets of interactions, or conflating two separate interactions into a single one. Our choice of w is motivated by a number of considerations. First, a 7-day period seems to be an appropriate temporal unit of analysis as students' online activity follows a weekly pattern. On average, the number of messages sent increases in the first few days of the week, until it reaches a peak on Thursday, and then shows a significant drop during the weekend. Second, we take various multiples of the one-week unit. We use the 21-day window because three weeks represent the best approximation of the time at which the rates of increase in messages and in new edge formation stabilise while the network is still rapidly growing. This choice finds further support in the observation of the distribution of dyadic response times: about 96.86% of all reciprocated edges are reciprocated within 21 days. We also used $w = 14$ and $w = 42$ to check robustness for the following reasons. On the one hand, the two-week window was motivated by the fact that there are spells of reduced activity (e.g., between day 62 and 73) that could be hardly detectable using smoothing windows of large width. On the other, we chose a six-week window because week 6 represents the end of a phase of significant network growth, followed by a phase in which users and acquaintances grow more smoothly towards their cumulative asymptotic levels. In fact, at day 42 the network already includes 77% of all users and 67% of all acquaintances. Also, the rate of increase in new messages shows a significant drop around day 42, when about 67% of all messages have already been sent.

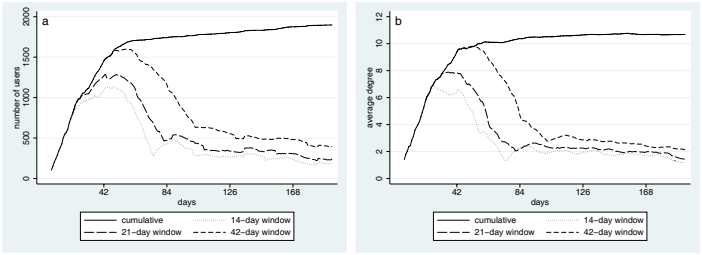

Figure 1: (a) Evolution of users; (b) Evolution of average degree.

at the vertex) and an out-degree k^O (i.e., the number of edges originating from the vertex) (Wasserman and Faust 1994). We define acquaintances in terms of vertex out-degree. This ensures that users are aware of their acquaintances. In other words, if user i sends a message to user j, then j becomes i's acquaintance, but i is not j's acquaintance unless another message is sent back from j to i. Since users can send more than one message to the same contact, the total number of messages is different from the number of acquaintances. At the end of day 194, out of a total of $59,835$ messages, there are only $20,296$ acquaintances. The mean number of acquaintances (or average out-degree) is thus $\bar{k^O} = 10.69$. Since $\bar{k^O}$ is equal to $\bar{k^I}$, the average in-degree (Wasserman and Faust 1994), to avoid confusion, and when the distinction does not add clarity, in what follows we will simply refer to the mean number of acquaintances as the average degree $\bar{k}$. Furthermore, in addition to directed edges, undirected edges are established between users when at least one message has been sent from one user to the other. The average degree based on undirected edges is $\bar{z} = 14.57$.

3 Network Evolution

Figure 1 shows the evolution of users and average degree $\bar{k}$ for the cumulative network and when smoothing windows are used. A new user is recorded as soon as he or she sends or receives a message for the first time. Moreover, one new acquaintance is added to the system as soon as a user sends a message to some other user for the first time. Unlike what recent models of growing networks suggest (see Dorogovtsev and Mendes 2003, Chapter 5), Figure 1 indicates that the network topology is not constructed uniformly over time. By contrast, the network evolves according to a two-fold regime, whereby an initial phase of rapid network growth is followed by a phase of structural stability.

The first phase, including the first six weeks (days 1-42), witnesses rapid increase in the number of users and acquaintances. The total number of acquaintances always exceeds the total number of users, and the former grows faster than a linear function of the latter. In this sense, the evolution of the network can be described as a non-linear process of "accelerated"

growth, also documented in other growing networks, such as the WWW, the Internet, networks of citations in the scientific literature, collaboration networks, networks of metabolic reactions and of software components (Barabási et al. 2002, Dorogovtsev and Mendes 2003, Faloutsos et al. 1999). The main reason for "accelerated" growth is that, after joining the network, users proceed to create new acquaintances. In agreement with recent theoretical and empirical studies of network growth (e.g., Albert and Barabási 2000, Barabási et al. 2002, Guimerà et al. 2005), the network topology thus evolves as a result not only of the contributions of newcomers, but also of the continuing activities of incumbents.[6] As the network grows, users become more interconnected and, as shown in Figure 1b, the average degree rapidly increases. The beginning of a new phase of structural stability (at day 43 and including the remaining 152 days) is marked by a noticeable decline and convergence of the daily rates of increase in users and acquaintances. This results in a reduction of the rate of increase in the average degree, which then remains almost stable for the rest of the observation period.

The two-fold regime of network evolution can be further investigated by looking at the changes in interaction patterns. This allows us to highlight the contribution to network evolution of two processes often neglected in network modelling efforts: "network construction" and "network use". As indicated by Figure 2a, reciprocity, here defined as the proportion of dyads in which the two users are each other's acquaintances, initially increases. Clearly, this contributes towards the rapid growth in average degree. At the same time, the divergence between the rate of increase in total messages per user and the rate of increase in average degree becomes more pronounced as time goes by. This is a clear indication of an ongoing change in interaction patterns. In fact, the difference between the two rates gives evidence of the extent to which users send more than one message to the same acquaintances, thereby reinforcing existing relationships (Guimerà et al. 2005, Reagans 2005). While reciprocity, by creating new edges, operates as a social mechanism of "network construction", reinforcement can be seen as a form of "network use", in that it documents the extent to which users rely on the existing network structure to communicate. In this sense, Figure 2b suggests not only that most activities of "network construction" occur in the first phase, but also that, as time goes by, more activities of "network use" set the stage for a network structure with relatively stronger relationships, and for a subsequent new phase of stability in interaction patterns.

The use of smoothing windows allows us to shed light on the interplay between "network construction" and "network use". In the first phase (for $w = 14$ and $w = 21$), both users

[6] In the case of the online community, incumbents' activities are further amplified by the likely tendency of users to contact others they have already met face-to-face, which swiftly replicates online the dyads that already exist offline.

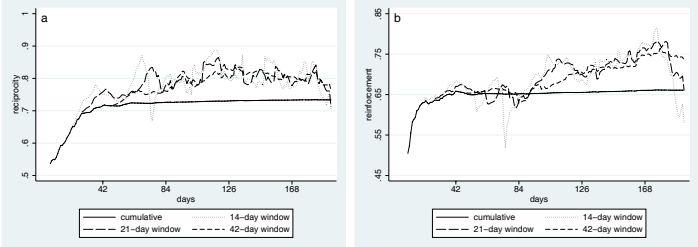

Figure 2: (a) Evolution of the proportion of relationships that are reciprocated; (b) Evolution of the proportion of messages used to reinforce existing relationships.

and acquaintances rapidly grow, yet for the rest of the observation, after only a few days of less pronounced increase, they rapidly decline towards their asymptotic values. This is chiefly due to the fact that "network construction" is mostly concentrated in the first few days, while "network use" increases over time as users reinforce their existing relationships. This means that, as time goes by, the marginal contribution to "network construction" of one additional day is not enough to compensate for the contribution lost as one day is taken out of the relevancy horizon. Moreover, this net loss in "network construction" is primarily the result of a more pronounced loss in acquaintances than users. This explains why, with windows, the average degree declines soon after the first few days of network growth cease to remain relevant.

With smoothing windows, stability is reached at a later stage than for the cumulative network, with fluctuations becoming smoother as the window width gets larger. In particular, the larger the width of the window, the longer it takes for users and acquaintances to reach their asymptotic values. In fact, stability occurs when the marginal contribution towards "network construction" becomes equal to the contribution lost as the window moves forward of one day. The time at which this occurs depends on how soon the events that took place at an earlier stage (and are richer in "network construction") become obsolete and are replaced by events that took place at a later stage (and are richer in "network use"). The more initial events are included in the window and the longer they remain in it, the longer it will take for the network to reach stability. Window width also affects communication patterns. As shown by Figure 2, the fluctuations of reciprocity and reinforcement around their asymptotic values become smoother as window width gets larger. Regardless of window width, however, results indicate an increasing trend in reinforcement, with users gradually devoting more time and resources to communicating with old acquaintances.

3.1 Degree Distribution

We define $p(k^I)$ and $p(k^O)$ to be the fraction of vertices in the network that have, respectively, in-degree k^I (i.e., with k^I incoming edges) and out-degree k^O (i.e., with k^O outgoing edges).

Equivalently, $p(k^I)$ and $p(k^O)$ are the probability that a vertex uniformly chosen at random has, respectively, in-degree k^I and out-degree k^O. The in- and out-degree distributions are the probability functions giving $p(k^I)$ and $p(k^O)$, respectively for each k^I and k^O (see Dorogovtsev and Mendes 2003, p.10). In random networks, any two vertices are connected with equal probability, and therefore the degree distribution is binomial, or Poisson in the limit of a large network size (Erdős and Rényi 1960). More specifically, as the total number of vertices N becomes large, the random network shows a Poisson distribution, sharply peaked, with a tail that decays as $1/k!$. This rapidly decreasing distribution is quite different from what has been observed in most real-world networks (for a review, see Dorogovtsev and Mendes 2003, Chapter 3). In fact, the empirical degree distributions of real-world networks show unusual fat tails of values that are far above the mean.

In recent literature, there has been a significant discussion as to how to formally represent this tail. For example, a number of authors (e.g., Albert et al. 1999, Barabási and Albert 1999, Faloutsos et al. 1999) have made the case that the distributions of connections on the Internet and the WWW closely obey power-law functions. For directed networks, these take the form: $p(k^I) \sim (k^I)^{-\tau}$ and $p(k^O) \sim (k^O)^{-\tau}$, where τ is a constant exponent. Barabási and Albert (1999) have suggested that similar power-law degree distributions apply to a great variety of real-world networks. When the degree distribution of a network can be formalised by a power-law function, the network is said to be "scale-free" as it lacks any intrinsic characteristic scale for the degree fluctuations, unlike, for instance, the Poisson or exponential distributions (Barabási and Albert 1999). The hallmark of "scale-free" networks is the high level of heterogeneity in the connectivity properties: while the majority of vertices are relatively poorly connected, there is an appreciable probability of finding a select minority of hubs that are many times better connected than average.

Figure 3 reports the empirical in- and out-degree distributions of the cumulative network at the end of the observation. A fat-tailed form of these distributions is clearly visible. Both distributions are well approximated by the linear behaviour on the double logarithmic scale, as indicated by the straight lines in Figure 3. More precisely, they can be fitted by the power-law functions: $p(k^I) \sim \left(k^I\right)^{-1.005}$ and $p(k^O) \sim \left(k^O\right)^{-0.889}$, where $p(k^I)$ is the probability that a user is the acquaintance of k other users, whereas $p(k^O)$ is the probability that a user has k acquaintances.[7] The power-law fit for the in-degree distribution has an R^2 of 0.9895, whereas

[7]For the in-degree distribution, the probability $p(k^I = 0)$ of being nobody's acquaintance is taken to be zero as users who did not receive any message were excluded from the analysis. Similarly, for the out-degree distribution the probability $p(k^O = 0)$ of having no acquaintance is taken to be zero as users who did not send messages were excluded from the analysis.

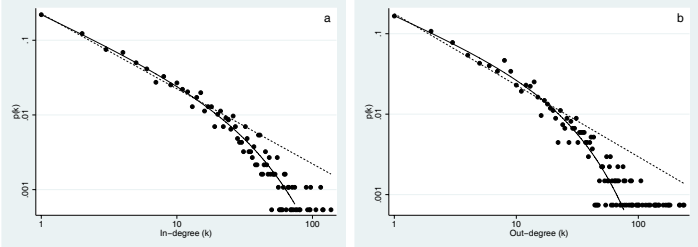

Figure 3: (a) In-degree distribution; (b) Out-degree distribution.

the one for the out-degree distribution has an R^2 of 0.9735, both with p values of less than 0.001. Moreover, as shown by the curved lines in Figure 3, the in- and out-degree distributions can also be well fitted (respectively, with R^2 of 0.996 and 0.9901, and p values of less than 0.001) by power-law functions with an exponential cut-off, following $p(k^I) \sim \left(k^I\right)^{-0.847} e^{-k^I/30.306}$ and $p(k^O) \sim \left(k^O\right)^{-0.665} e^{-k^O/28.01}$.[8]

The findings thus lend support to the conjecture that the network belongs to the class of "scale-free" networks: while the majority of users appear to communicate with only a few others, there is a small fraction of hubs that collect a disproportionately large amount of incoming and outgoing connections. Interestingly, both estimated power-law exponents are lower than two. This is the distinctive signature of a network structure dominated by the hubs, and not by the majority of poorly connected vertices (Dorogovtsev and Mendes 2003). In the infinite network limit, and when vertices do not exit the network and edges do not decay, the first moment of the in- and out-degree distributions (i.e., the average degree) must therefore diverge.

When network growth is non-linear, the degree distribution does not need to be stationary (Dorogovtsev and Mendes 2003). Indeed one way to investigate the effects of "accelerated" growth on the network structure is to analyse how the degree distribution varies over time. We measured the empirical in- and out-degree distributions at the end of each day, with and without smoothing windows. For each day, the theoretical in- and out-degree distributions are power-law dependences, and in each case the fit has an R^2 ranging from 0.9617 to 0.9995 for the in-degree distribution, and from 0.9879 to 0.9998 for the out-degree distribution, both with p values lower than 0.001.

[8] A number of possible explanations of the origins of an exponentially truncated power-law function have been proposed (e.g., see Barabási et al. 2002, Krapivsky et al. 2000). For example, Newman (2001b) suggested that the fitted values of the cut-off were produced by the finite time frame of his analysis that prevented vertices from connecting to a large number of others. The same argument could also help explain the better fit we obtained with an exponential cut-off: the limited period of time covered by our dataset may have prevented users from interacting with a larger number of others than would be the case with a longer observation period (see also Section 4.3).

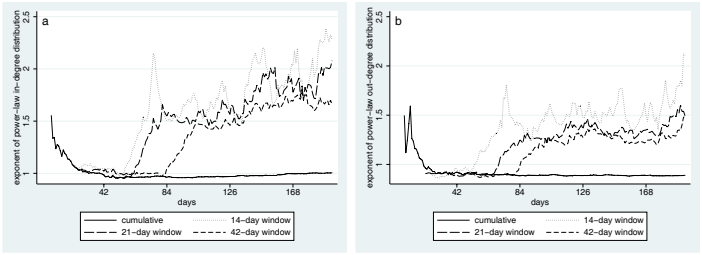

Figure 4: Evolution of the exponent τ of the power-law in-degree (a) and out-degree (b) distribution.

Figure 4 shows the evolution of the exponent of both power-law dependences. For the cumulative network, the estimated exponents always remain below two. Therefore, for the whole period, the average structural properties of the network are dominated by the small fraction of hubs (Dorogovtsev and Mendes 2003). By contrast, with smoothing windows the in- and out-degree distributions remain fat-tailed, yet they exhibit regions with distinct power-law behaviour separated by crossover points at which the exponent of the power-law dependence crosses the value of two. Clearly, the choice of the day at which the network is sampled and of the window width affects the observed empirical degree distribution. As fewer events are taken into account, the distribution becomes more unstable and dependent on the time at which the observation is taken. Unlike what has been found in other studies of evolving social networks (e.g., Kossinets and Watts 2006), our results suggest that, when vertices join and leave the network and edges are formed and severed in a relatively short period of time, the vulnerability of the degree distribution to measurement assumptions makes it difficult to draw generalisations from network snapshots with reasonable accuracy.

3.2 Giant Component

The theory of percolation and random networks suggests that, as the density of edges increases, there is a continuous phase transition at which a giant weakly connected component is generated (Bollobás 1985, Erdős and Rényi 1960). Clusters of connected vertices develop and then generate links with each other until a connected subset of reachable vertices is created whose size scales extensively. The emergence of this subset is usually referred to as percolation transition (Guimerà et al. 2005, Newman 2001b, Newman et al. 2001). Above the transition, the giant component includes a large proportion of the network, whereas the size of all other components is small and independent of the number of vertices in the network. Our data provide support for the predicted percolation transition. At the end of the observation period, the cumulative network shows a clear large component, encompassing $1,893$ vertices out of a total of $1,899$. The second-largest component is made up of just two vertices. We also measured the size of the

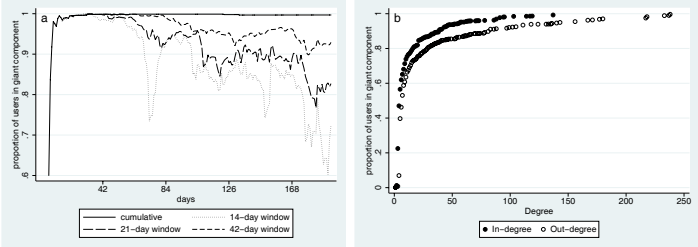

Figure 5: Size of the giant component: (a) Evolution; (b) Sensitivity analysis.

largest strong component in which vertices are mutually reachable by a directed path (Dorogovtsev and Mendes 2003). This component includes 1,266 vertices, whereas the second largest strong component is again made of only two vertices. With such pronounced differences between largest and second-largest components, it is clear that, by day 194, the network is well within the percolating regime (Dorogovtsev and Mendes 2003, Newman 2001b). In other words, the network is not on the borderline of connectedness and there is no risk of fragmentation.

To study the emergence of the giant component, for each day we measured the relative size of the largest weakly connected component as a ratio between the number of vertices in it and the total number of vertices in the network. As shown by Figure 5a, a phase transition takes place at a very early stage, and the majority of vertices (98%) soon become part of a single giant cluster (by day 9).[9] The network always remains well inside the percolating regime, regardless of the measurement assumptions. However, after the initial growth, the relative size of the giant component varies depending on whether smoothing windows are used and on their width. In fact, while for the cumulative network the relative size of the giant component remains almost stationary, with smoothing windows it decreases over time. In particular, the smaller the width of the window, the smaller the relative size of the giant component and the more vulnerable the measurements become to the timing of sampling.[10]

Since we noticed that the network belongs to the class of "scale-free" networks, it is interesting to investigate whether a giant component emerges as a result of the structural role played by the highly connected vertices in holding the whole network together. Figure 5b shows the size of the largest weakly connected component when all vertices with a specific in- and out-degree or higher are removed from the network. It is clear that the network remains connected through-

[9]Our results are in line with other empirical findings (e.g. the movie actor network (Watts and Strogatz 1998)). When the database includes the very early moments of the network, the majority of vertices quickly form a large percolating cluster at an early stage of the network (Barabási et al. 2002).

[10]Despite this, the relative size of the giant component never goes below 60.22% of the whole network with $w = 14$, 83.58% with $w = 28$, and 89.78% with $w = 42$. In addition, the size of the second-largest weakly connected component never exceeds 7.18% of the network with $w = 14$, 1.74% with $w = 28$, and 1.28% with $w = 42$.

out the tail of the in- and out-degree distributions. The size of the giant component never goes below 70% of the whole network until all vertices with 8 or more incoming edges (i.e., 16.32% of all vertices), or 13 or more outgoing edges (i.e., 12.37% of all vertices), are removed. From that point, and as additional vertices with lower in- and out-degree are removed, the network quickly disintegrates. Being vulnerable to the removal of a relatively small subset of vertices, the network shows indeed one of the defining features of "scale-free" networks: a minority of high-degree vertices are responsible for creating a giant component and connecting the network (Albert et al. 2000).

3.3 Average Distances

The geodesic distance between a pair of vertices is the shortest path length between them. The average geodesic distance g is the mean shortest path length between all pairs of vertices that have a connecting path (Wasserman and Faust 1994, p.110). For the cumulative network at day 194, we found: $g = 3.055$. This value was calculated using undirected edges included only in the giant connected component, as distances between vertices in disconnected components are infinite.

To investigate the "small-world effect", we now compare our findings with g_{rand}, the geodesic distance of a random network with same size and average degree as the giant component of our network. We found: $g_{rand} = 3.0845$.[11] More generally, for a given number N of vertices with a given mean degree $\bar{z}$, the geodesic distance in a random network scales logarithmically with N and can be approximated by $logN/log\bar{z}$ (Bollobás 1985). Watts and Strogatz (1998) defined a social network as being "small" if typical distances are comparable with those on a random network, i.e., if they grow logarithmically or slower with network size for fixed mean degree. Formally, when a network is "small", the ratio $\gamma = \frac{\text{measured geodesic distance}}{\frac{\log N}{\log \bar{z}}}$ equals one. For the giant component of the cumulative network at day 194, we found: $\gamma = 1.086$.

We also measured $g^{'}_{rand}$, the geodesic of a random network with same size and degree distribution as the giant component of our network (Newman et al. 2001). We found: $g^{'}_{rand} = 2.9004$.[12] Thus, g mirrors fairly closely both g_{rand} and $g^{'}_{rand}$. However, while being larger than $g^{'}_{rand}$ as more edges are used to create local clusters (see Section 3.4), g is smaller than g_{rand} because of

[11] This value is the average of geodesics from 100 simulations of random networks with $N = 1,893$ and $\bar{z} = 14.62$.

[12] $g^{'}_{rand}$ is the average of geodesics from 100 simulations of random networks with $N = 1,893$ and same degree distribution as the giant component of the undirected real network. We used the giant component of the undirected real network because edges are reciprocated in the real directed network but not in a corresponding directed random network, which would produce an undirected random network with more edges than the real undirected network.

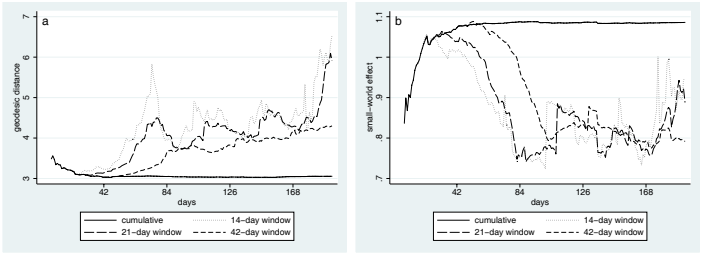

Figure 6: Evolution of geodesic distance (a) and "small-world effect" (b).

the role played by the highly connected vertices in reducing distances. Altogether, these findings indicate that the network is indeed compact and, like other documented social networks, can be regarded as a "small world" (e.g., Ebel et al. 2002, Davis et al. 2003, Newman 2001b, Uzzi and Spiro 2005, Watts and Strogatz 1998).

Figure 6a shows the evolution of geodesic. For the cumulative network, after an initial decline, geodesic reaches its asymptotic level at a time in which vertices and average degree are still rapidly increasing (Barabási et al. 2002, Holme et al. 2004). This indicates that the network obtains a stationary topology before the rate of growth in its size slows down. Note that the observed reduction in geodesic runs counter to what extant literature suggests (e.g., see Albert et al. 1999, Bollobás 1985). In fact, according to all network models, average distances should increase with network size. However, since in our case the network follows a non-linear process of "accelerated" growth, the structure of connections becomes denser over time, and average separation between vertices decreases. Results also indicate a negative correlation between distances and network size when smoothing windows are used: while geodesic shows an increasing trend, network size declines over time (Figure 1a). This can be explained by the fact that, with windows, the decay of edges drives the average degree down (Figure 1b), thereby negatively affecting network compactness.

Moreover, as shown in Figure 6b, there is always convergence between the measured distance and the expected $logN/log\bar{z}$ behaviour (see also Davis et al. 2003, Uzzi and Spiro 2005). However, the "small-world effect" does not remain stable over time and is vulnerable to measurement assumptions. When windows are used, network compactness declines, and the convergence of average distances with those on a classical random network becomes more unstable and less pronounced.

3.4 Clustering

As recent studies have pointed out, social networks differ from random networks in one fundamental way: they exhibit local communities in which a higher-than-average number of vertices

are connected to one another (Jin et al. 2001, Newman and Park 2003). If there are three individuals, i, j and l, and i is acquainted with j and l, how likely is it that j is acquainted with l? This symmetry among triples of vertices is often referred to as transitivity (Davis 1970, Holland and Leinhardt 1970, Levine and Kurzban 2006, Rapaport 1953, Wasserman and Faust 1994). Transitivity is the probability of j being acquainted with l. For undirected networks, it can be measured with the following clustering coefficient C (cf. Newman et al. 2001):

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{1}$$

where, $0 \leq C \leq 1$.[13] In a completely connected network, C is equal to one. By contrast, in a random network with N vertices and average degree $\bar{z}$, we have: $C \approx \bar{z}/N$; hence, C goes to zero in the limit of large network size (Dorogovtsev and Mendes 2003, p.16). In the sociology literature, this measure is often referred to as "fraction of transitive triples" (Wasserman and Faust 1994).

Empirical evidence has shown that, in most real-world networks, clustering takes a non-zero value (e.g., Barabási et al. 2002, Davis et al. 2003, Ebel et al. 2002, Holme et al. 2004, Ingram and Roberts 2000, Newman 2001b, Uzzi and Spiro 2005, Watts and Strogatz 1998). For social networks, this means that, even when the size of the network is very large, there is a finite probability that two individuals will be acquainted if they share an acquaintance. This hypothesis can be tested with our data by using undirected edges for the cumulative network at day 194. We obtained: $C = 0.0568$. This value can be compared to the clustering coefficient C_{rand} of a random network of same size and average degree as the real network. We found: $C_{rand} = 0.0077$. C is also larger than $C'_{rand} = 0.0139$, the clustering coefficient of a corresponding random network of identical size and same degree distribution as the real network (Newman et al. 2001, Watts and Strogatz 1998).[14] The way in which edges are established across the network thus differs from what would be expected with random connections.

Figure 7 shows the evolution of the clustering coefficient for the cumulative (undirected) network and with smoothing windows. For the cumulative network, C always remains greater than would be expected by chance: for each day, we measured C_{rand} and C'_{rand}, and both coefficients are lower than the corresponding C for that day. In particular, the deviation between actual and predicted clustering is most pronounced during the initial phase of "accelerated" growth, while it attenuates and becomes more stable in the remaining period. It is interesting to note that,

[13] Note that the factor of three in the numerator of Eq. 1 compensates for the fact that each complete triangle of three vertices contributes three connected triples, centered on each of the three vertices, and ensures that $C = 1$ on a completely connected network.

[14] Both C_{rand} and C'_{rand} are based on 100 simulations of random networks. C'_{rand} is measured using the same degree distribution as the real undirected network.

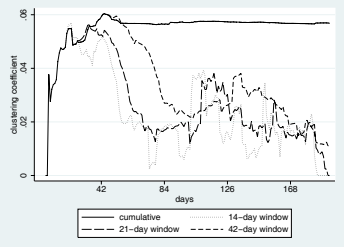

Figure 7: Evolution of clustering.

while results are vulnerable to window width and timing of sampling, the clustering coefficient of the cumulative network exhibits an increasing trend that is at variance with other empirical findings (e.g., Holme et al. 2004, Kossinets and Watts 2006) and with the behaviour predicted by some recently developed network models (e.g., Barabási and Albert 1999, Barabási et al. 2002). Effects arising from social and spatial proximity can help explain the findings (Feld 1981, Monge et al. 1985). When individuals all belong to the same circles, they may independently start separate interactions in pairs, and in this way contribute also to clustering. In our case, being part of the same campus or even the same academic unit may play a fundamental role in triggering independent dyadic interactions (cf. Newman 2001b). Users may communicate "offline" with one another in pairs, and then replicate their interactions online. Affiliation to the same groups, and not referrals from existing acquaintances, may then be partly responsible for introducing users and creating "offline" loops that are rapidly replicated online as soon as users join the network.

4 Network Growth Mechanisms

In this section, we propose and evaluate three organising principles of network growth: (1) new edges from a newcomer to an incumbent are more likely to be directed towards a "popular" vertex (i.e., with a high in-degree) than an "unpopular" one; (2) new edges from an incumbent to a newcomer are more likely to be created by an "active" vertex (i.e., with high out-degree) than a "dormant" one; (3) new edges between incumbents are more likely to be created by an "active" vertex than a "dormant" one, and directed towards a "popular" vertex than an "unpopular" one.

4.1 Preferential Attachment

Classical network models assume that the attachment of edges in a growing network takes place randomly, without any preference (Bollobás 1985, Erdős and Rényi 1960). This means that the probability that an edge becomes attached to a vertex is taken to be independent of the

degree of this vertex. In recent years, however, the study of the topology of real-world networks and the discovery of fat-tailed degree distributions has induced a paradigm shift and prompted the development of a series of new models that provide a more realistic representation of the growth mechanisms underpinning evolving networks (for a review, see Dorogovtsev and Mendes 2003, Chapter 5). One of the most influential growing network models, proposed by Barabási and Albert (1999), is premised on the assumption of *preferential attachment*. Drawing on the "Gibrat principle" implemented in the 1950s by Simon (1955) and better known in sociology as the "Matthew Effect" (Merton 1968), preferential attachment formalises the notion that newly added vertices are more likely to connect to well-connected vertices than poorly connected ones. More specifically, the probability that new edges will become attached to a vertex is assumed to be proportional to the degree of that vertex. Unlike random networks, where the probability of attracting new edges remains independent throughout the whole growth process, networks evolving under the mechanism of preferential attachment show a rise in connectivity in the direction of vertices that are already highly connected. The combination of growth and preferential attachment can thus explain the emergence of an inhomogeneous network structure and a fat-tailed connectivity distribution (Barabási and Albert 1999). Moreover, a series of studies have demonstrated that the functional form of the dependence of the probability of attracting edges on the degree of a vertex plays a crucial role in determining the form of the stationary degree distribution (e.g., Dorogovtsev and Mendes 2003, Krapivsky et al. 2000). In particular, only linear preferential linking produces "scale-free" distributions, whereas non-linear dependence generates deviations from a power-law.

Using our data, we can test for preferential attachment. First, we study the tendency of a new vertex to establish an edge with an incumbent vertex that receives a higher-than-average number of connections. To this end, we define the "preferential attachment propensity" P_T as follows:

$$P_T = \frac{1}{T} \times \sum_{t=1}^{T} \frac{k^I_{j,t} - \bar{k}^I_t}{k^I_{max,t} - \bar{k}^I_t}, \tag{2}$$

where $k^I_{j,t}$, $\bar{k}^I_t$ and $k^I_{max,t}$ are, respectively, the in-degree of an incumbent vertex j to which a new vertex becomes attached, the average in-degree, and the maximum in-degree that can be found in the network, immediately *before* the addition of a new edge at t. Here t does not refer to real time, but to the series of edges created by new vertices and directed towards incumbent ones.[15] If network growth were governed by random linking, Eq. 2 would approach zero in

[15]Between two subsequent additions of such edges, the network has evolved as a result of the addition of new internal edges. In order to take into account the most recent addition of *any* edge just before the new edge was created at t, $k^I_{j,t}$, $\bar{k}^I_t$ and $k^I_{max,t}$ cannot be measured simply at $t-1$, as this would exclude edges created between $t-1$ and t.

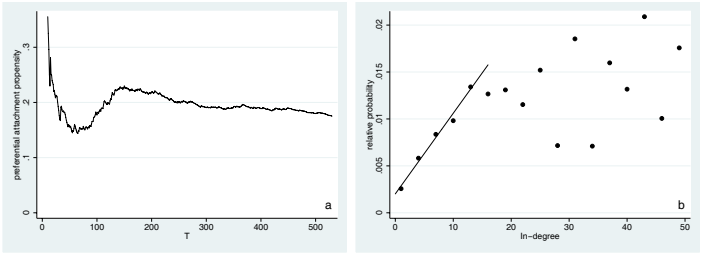

Figure 8: Edges connecting a newcomer to an incumbent: (a) Preferential attachment propensity; (b) Relative probability that a newcomer will connect to an incumbent with a given in-degree.

the limit $T \to \infty$. By contrast, if linking were preferential and proportional to the in-degree of the vertex, Eq. 2 would take a value higher than zero. Finally, if linking were preferential and inversely proportional to vertex in-degree, Eq. 2 would take a value lower than zero. As shown by Figure 8a, after an initial fluctuation, P_T asymptotically approaches the final value of 0.1752 at $T = 530$.[16] The network exhibits a propensity for non-random linking: new vertices preferentially establish connections with incumbent vertices with higher-than-average in-degree.

We now measure the probability of attachment as a function of vertex in-degree. Following Newman (2001a), we define $RP(k^I)$ as the relative probability that a new vertex becomes attached to an incumbent whose in-degree is k^I:

$$RP(k^I) = \frac{1}{T} \times \sum_{t=1}^{T} \frac{P(k^I)_t}{\frac{n_{k^I,t}}{N_t}}, \tag{3}$$

where $P(k^I)_t$ is the probability that at t a new vertex becomes connected to an incumbent with in-degree k^I, $n_{k^I,t}$ is the number of vertices that, immediately *before* the new vertex joins the network at t, have received k connections, and N_t is the number of vertices immediately *before* the new vertex joins the network. Thus, $RP(k^I)$ is the ratio between the actual probability of connection between a new vertex and an incumbent with in-degree k^I and the probability of a connection between the two vertices in a random network where current in-degree does not affect the likelihood of attracting further edges. For growing random networks, we would have $RP(k^I) = 1$ for every k^I; for networks growing under preferential attachment, conversely, $RP(k^I)$ should increase with k^I. Moreover, if it increases linearly, then the resulting stationary

[16] We also evaluated Eq. 2 for a corresponding random network with the same number of edges (530) between a new vertex and an incumbent one as in the real network. At each time step, one new vertex was added and a new edge was randomly established between that vertex and an incumbent one. This produced an exponential in-degree distribution and a value of P_{530} very close to zero. Same result was obtained with a random network similarly constructed, but with the same number of vertices (1,899) as the real network. We then conducted a z-test on the difference between observed value of P and its expected value on a random network, and found a p value lower than 0.001, making it highly unlikely that the reported difference is simply a chance occurrence.

degree distribution will be a power-law (Barabási and Albert 1999, Krapivsky et al. 2000).

In Figure 8b, the normalised values for Eq. 3 are shown. As the figure indicates, $RP(k^I)$ increases with k^I for the initial part of the curve. A vertex with in-degree of 16 is 1.91 times as likely to receive a new connection as a vertex with an in-degree of only 8, and about 9.79 times as likely as a vertex with no connections. For higher values of k^I, the curve first appears to flatten off, and then fluctuates. However, the statistics are relatively poor in the region of large values of k^I given the limited number of vertices with that many connections. The observed behaviour is not surprising: no user can reasonably communicate with an indefinitely large number of others in a finite period of time, and therefore the increasing trend of $RP(k^I)$ must stop at some point. The best linear fit of the data is up to the in-degree of 16, with an R^2 of 0.72 and a p value of less than 0.001. Unlike what was found in Newman (2001a), the connectivity value at which the relative probability starts to become sub-linear falls below the value at which the in-degree distribution starts to deviate from the power-law dependence (see Figure 3). This is understandable: in our analysis $RP(k^I)$ refers to the subset of edges from newcomers to incumbents, whereas the in-degree distribution reflects all edges in the network.

4.2 Preferential Activity

We now study the tendency of incumbents with a higher-than-average out-degree to create new connections with newcomers. We define the "preferential activity propensity" P_T^A as follows:

$$P_T^A = \frac{1}{T} \times \sum_{t=1}^{T} \frac{k_{i,t}^O - \bar{k}_t^O}{k_{max,t}^O - \bar{k}_t^O}, \tag{4}$$

where $k_{i,t}^O$, $\bar{k}_t^O$ and $k_{max,t}^O$ are, respectively, the out-degree of an incumbent vertex i, the average out-degree and the maximum out-degree, measured immediately *before* the addition of the new edge at t. If network growth were governed by random activity, Eq. 4 would approach zero in the limit $T \to \infty$. By contrast, if activity were preferential and directly or inversely proportional to vertex out-degree, Eq. 4 would have a value higher or lower than zero, respectively. As indicated by Figure 9a, after small initial fluctuations, P_T^A declines until a final value of 0.2376 is reached at $T = 1,223$.[17] The network thus shows propensity for non-random activity: new vertices receive connections preferentially by incumbent vertices with higher-than-average out-degree.

[17] Eq. 4 was evaluated also for a random network with the same number of edges $(1,223)$ from incumbent to new vertices as in the real network. At each time step, one randomly chosen incumbent established an edge with a newcomer. This produced an exponential out-degree distribution and $P_{1,223}^A \approx 0$. Same result was obtained with a random network similarly constructed, but with the same number of vertices $(1,899)$ as the real network: $P_{1,898}^A \approx 0$. Once again, the z-test on the difference between observed value of P^A and its expected value on a random network produced a p value lower than 0.001.

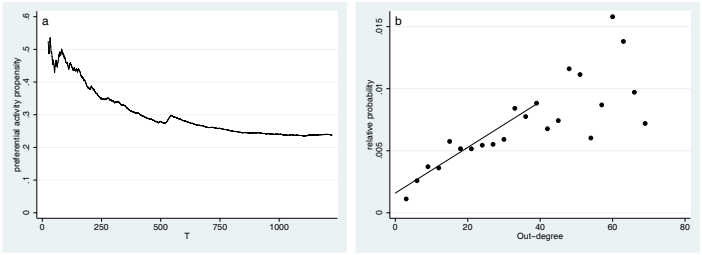

Figure 9: Edges connecting an incumbent to a newcomer: (a) Preferential activity propensity; (b) Relative probability that a newcomer will be contacted by an incumbent with a given out-degree.

We now measure the dependence of the probability that an incumbent creates a new connection with a newcomer on the incumbent's out-degree. We define $RP(k^O)$ as the relative probability that a new vertex receives a connection from an incumbent whose out-degree is k^O:

$$RP(k^O) = \frac{1}{T} \times \sum_{t=1}^{T} \frac{P(k^O)_t}{\frac{n_{k^O,t}}{N_t}}, \tag{5}$$

where $P(k^O)_t$ is the probability that at t a new connection to a new vertex is established by a vertex with out-degree k^O, whereas $n_{k^O,t}$ and N_t are, respectively, the number of vertices with out-degree k^O and the total number of vertices, immediately *before* the new connection is created. For growing networks with no preferential activity, $RP(k^O)$ would be one for every k^O, whereas for networks growing under preferential activity, $RP(k^O)$ should increase with k^O. Once again, if the increase is linear, the resulting stationary out-degree distribution will be a power-law (Barabási and Albert 1999, Krapivsky et al. 2000).

The normalised values for Eq. 5 are shown in Figure 9b. For the initial part of the curve, $RP(k^O)$ increases with k^O: a vertex with out-degree of 40 is 2.26 times as likely to establish a new connection as a vertex with an out-degree of only 20, and about 96.16 times as likely as a vertex with no connections. Then the curve appears to flatten off and fluctuate for higher k^O (as no user can create an indefinitely large number of connections in a finite period of time), although the statistics become poor in the region of very large k^O (as vertices that establish that many connections are few). The best linear fit of the data is up to an out-degree of 40, with an R^2 of 0.53 and a p value of less than 0.001. Once again, as this refers to just a subset of the outgoing edges, sub-linearity begins around a connectivity value that does not coincide with the value at which the out-degree distribution starts to deviate from the power-law dependence (see Figure 3).

4.3 Preferential Involvement

As the network evolves, a number of new edges are established between incumbents (Barabási et al. 2002, Guimerà et al. 2005). These are new connections between vertices that were already

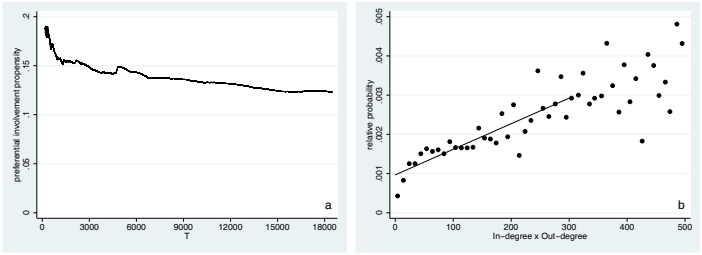

Figure 10: Edges connecting two incumbents: (a) Preferential involvement propensity; (b) Relative probability of a new edge between incumbents as a function of the product between the sender's out-degree and the recipient's in-degree.

part of the system but did not exchange messages before. There have been a few attempts in the literature to propose analytical models of networks that grow as a result of the addition of internal edges between existing vertices. For example, recent studies have formalised growth processes in which one end of the newly added internal edge is chosen randomly whereas the other with probability proportional to vertex degree (Albert and Barabási 2000), or where preferential attachment applies to both ends of the new internal edge (Barabási et al. 2002). However, as here we can distinguish between vertex out- and in-degree, we can separately model the creation and the attachment of internal edges.

First, we study the propensity of an incumbent i with a higher-than-average out-degree to become attached to another incumbent j with a higher-than-average in-degree. We define the "preferential involvement propensity" as follows:

$$P_T^I = \frac{1}{2T} \times \sum_{t=1}^{T} \left[\frac{k_{j,t}^I - \bar{k}_t^I}{k_{max,t}^I - \bar{k}_t^I} + \frac{k_{i,t}^O - \bar{k}_t^O}{k_{max,t}^O - \bar{k}_t^O} \right], \tag{6}$$

where variables are measured as before, with the only exception that now, to avoid one-edge loops, $\bar{k}_t^I$ and $k_{max,t}^I$ are calculated without taking into account $k_{i,t}^I$. If internal edges were added in a non-random fashion, Eq. 6 would take a value different from zero in the limit $T \to \infty$. Specifically, if involvement were preferential and jointly proportional to the out-degree and in-degree of the sending and receiving vertices, Eq. 6 would have a value higher than zero. As shown by Figure 10a, the network does indeed show a non-random linking mechanism for internal edges. P^I remains always positive and asymptotically approaches a final value of 0.1234 at $T = 18,470$.[18]

We now measure the dependence of the probability of an internal edge on the in- and out-degree of, respectively, the sending vertex and the receiving one. We define $RP(k^O, k^I)$ as

[18] Eq. 6 was evaluated also for a random network with the same number of vertices (1,899) and edges (20,296) as the real network. Starting with 1,899 isolated vertices, at each time step we randomly established a new edge between any two vertices until 20,296 edges were created. To replicate internal growth, Eq. 6 was applied to the series of edges connecting non-isolates. We obtained a final value of nearly zero. The z-test on the difference between observed value of P^I and its expected value on a random network produced a p value lower than 0.001.

the relative probability that an incumbent with out-degree k^O connects to an incumbent with in-degree k^I:

$$RP(k^O, k^I) = \frac{1}{T} \times \sum_{t=1}^{T} \frac{P(k^O, k^I)_t}{\frac{d(k^O,k^I)_t}{D_t}}, \tag{7}$$

where $P(k^O, k^I)_t$ is the probability that at t a new edge is established between an incumbent with out-degree k^O and an incumbent with in-degree k^I, and $d(k^O, k^I)_t$ and D_t are, respectively, the number of directed edges that are still to be established from vertices with out-degree k^O to vertices with in-degree k^I, and the total number of directed edges that are still to be established, immediately *before* the addition of the new edge.[19] $RP(k^O, k^I)$ is therefore the ratio between the actual probability of connection between an incumbent with out-degree k^O and an incumbent with in-degree k^I, and the probability of a connection between the two incumbents in a random network where current out-degrees and in-degrees do not affect the likelihood of generating internal edges.

For networks growing with random linking, we would have $RP(k^O, k^I) = 1$ for every combination of k^O and k^I; conversely, for networks growing under preferential involvement, $RP(k^O, k^I)$ should increase as either k^O or k^I increases. Figure 10b shows the normalised values for $RP(k^O, k^I)$, assuming that Eq. 7 factorizes into the product $k^O \times k^I$ (Barabási et al. 2002). Initially, $RP(k^O, k^I)$ shows an increasing trend that can be linearly fitted up to a connectivity value of 300 ($R^2 = 0.24$). Thereafter, the statistics become poor and the curve flattens off with strong fluctuations.[20]

[19] In order to avoid multiple connections in the same direction between pairs of vertices and closed one-edge loops, $d(k^O, k^I)_t$ is not necessarily equal to $n_{k^O,t} \times n_{k^I,t}$ as this product could include pairs that were already connected in that direction as well as loops in case the same vertex is a member of both $n_{k^O,t}$ and $n_{k^I,t}$. Similarly, D_t is not equal to $N_t \times (N-1)_t$ as this product includes directed edges that were already established before the new edge is added.

[20] A similar growth mechanism scaling with the product of vertex degrees for internal edges was also used to describe the evolution of language (see Dorogovtsev and Mendes 2003, p.152) and of scientific collaboration networks (Barabási et al. 2002). For those networks, "accelerated" growth through preferential linking helped explain the existence of a crossover connectivity and two distinct scaling regimes of the degree distribution. A crossover towards a larger power-law exponent can be easily approximated with an exponential cut-off (Barabási et al. 2002). This may partly explain why by truncating the power-law dependence with a similar cut-off we could obtain a better fit to the empirical in- and out-degree distributions. Deviations of these distributions from power-law behaviour in the region of large connectivities are amplified also by sub-linear preferential attachment and activity in that region (cf. Newman 2001a).

5 Implications for Interpersonal Dynamics

In this section, we discuss and compare the three mechanisms of network growth previously outlined. We also investigate how these mechanisms are mediated by two other local ordering principles: homophily, the principle that similarity breeds connection (Lazarsfeld and Merton 1954); and focus constraints, the principle that social associations depend on opportunities for social contact (Feld 1981). A significant body of research on homophily has documented a positive association between sharing an attribute and some baseline level of interpersonal attraction (e.g., McPherson et al. 2001). That attraction could be reflected in a heightened probability of similar people to select each other (Hinds et al. 2000, Kossinets and Watts 2006, Louch 2000), or communicate more frequently and develop a stronger social interaction (Reagans 2005). This suggests that the extent to which the three growth mechanisms of preferential attachment, activity and involvement guide the way connections are forged should be mitigated when interactions occur between socially similar individuals. In particular, we predict that the role of individuals' popularity (in-degree) and past activity (out-degree) in guiding interpersonal dynamics will become weaker as the degree of similarity between the interacting individuals increases. A similar argument can also apply to focus constraints and help formulate predictions of their effects on the three network growth mechanisms. By increasing opportunities for social contacts (Feld 1981, Kossinets and Watts 2006, Louch 2000, Monge et al. 1985), focus constraints are expected to enhance the likelihood of offline communication between individuals, which in turn will mitigate the extent to which preferential attachment, activity and involvement shape online interpersonal dynamics. In other words, individuals in social and physical proximity with each other are more likely than distant individuals to communicate online because they have already met offline. Hence, in the presence of focus constraints, we expect popularity and past activity to become weaker predictors of communication behaviour.

Table 1 reports the values for Eq. 2, Eq. 4, and Eq. 6, assessed for various groups of edges. We used the following criteria to create these groups. Edges were initially assigned to three distinct groups depending on whether they connected: a) a newcomer to an incumbent; b) an incumbent to a newcomer; or c) two incumbents.[21] First, from each of these groups we extracted further subsets of edges based on vertex similarity. In agreement with a considerable amount of studies of homophily (McPherson et al. 2001), we used the following five demographic characteristics as indicators of social similarity: age: gender; year of study; region of origin; and marital status. Edges were then grouped according to how many of those attributes the interacting vertices

[21] When Eq. 2 is applied to edges between incumbents, to avoid one-edge loops $\bar{k}_t^I$ and $k_{max,t}^I$ are calculated without taking into account $k_{i,t}^I$.

had in common. Second, in analogy to other empirical studies of the effects of foci on patterns of social relationships (e.g., Feld 1981, Kossinets and Watts 2006, Monge et al. 1985), we used academic unit as an indicator of sources of opportunities for individuals to develop joint activities and interactions. From each of the initial three groups of edges, we then extracted a subset in which the interacting vertices shared academic unit.

	Number of edges			P		P^A		P^I
sender i	new	old	old	new	old	old	old	old
receiver j	old	new	old	old	old	new	old	old
All edges	530	1223	18470	0.1752***	0.1266***	0.2376***	0.1201***	0.1234***
$SA_{i,j} = 0$	119	367	5048	0.1798*	0.1243***	0.2797***	0.1494***	0.1368***
$SA_{i,j} \geq 1$	411	856	13422	0.1738***	0.1275***	0.2196***	0.1091***	0.1183***
$SA_{i,j} \geq 2$	256	512	7650	0.1725**	0.1269***	0.1986***	0.0979***	0.1124***
$SA_{i,j} \geq 3$	125	212	3169	0.1345	0.1191***	0.1382*	0.0874***	0.1032***
$SA_{i,j} \geq 4$	31	59	796	0.0973	0.1054**	0.1418	0.0812*	0.0933**
Same academic unit	76	171	2255	0.1392	0.1233***	0.2021**	0.1008***	0.1121***

Note: $SA_{i,j}$ = number of shared attributes between sender and recipient. Statistical significance of the difference between observed values and expected values on a random network: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 1: Preferential linking mechanisms, homophily and focus constraints.

Popularity is attractive. Users who are more popular have a better chance to be contacted by others. In particular, popularity plays a stronger role in guiding newcomers' choices than incumbents', as a t-test conducted on the difference between the values of P when the senders are newcomers and when they are incumbents indicates ($p < 0.001$). This is not surprising. Newcomers are likely to be less familiar with the system and possess less information about other current users than incumbents, and therefore they rely heavily on popularity to gauge the attractiveness of potential acquaintances. Moreover, our results lend support to the predicted effects of homophily on preferential attachment. We performed a regression analysis using P as the dependent variable and number of shared attributes (from one to five) between interacting users as the independent variable. We found that both when the sender is a newcomer ($\beta = -0.021$; bootstrap $SE = 0.013$; $p < 0.10$) and when he or she is an incumbent ($\beta = -0.004$; bootstrap $SE = 0.002$; $p < 0.05$), homophily exerts mitigating effects on preferential attachment. As the number of the attributes shared by the interacting users increases, users care less about the popularity of their acquaintances. In this case, a less popular recipient can have the same degree of attractiveness as a more popular one as long as he or she shares with the sender a larger number of attributes that offset the negative effects of the lack of popularity. On the other hand, the findings do not support our prediction of the effects of focus constraints on preferential attachment. We conducted a t-test on differences between the values of P for edges between

users that share academic unit and edges between users that do not, and these differences are not statistically significant. Therefore, there is no evidence suggesting that the way popularity guides users' behaviour depends on whether or not they share academic unit.

Activity breeds more activity. Users who have been more active have a better chance to contact others. This is the case especially when the recipients are newcomers, as a t-test conducted on the difference between the values of P^A when recipients are newcomers and when they are incumbents indicates ($p < 0.001$). Thus, the more connections a user has started, the more likely a newcomer is to be contacted by that user. By contrast, the effects of past activity on the likelihood to start communication are mitigated when the recipients are incumbents. In this case, the sender is likely to have more information for assessing the potential acquaintance's attractiveness, and out-degree becomes a weaker predictor of future activity. In addition, the results lend support to the predicted effects of homophily on preferential activity. Estimated regression coefficients for P^A on number of shared attributes between interacting users indicate that when the recipient is either a newcomer ($\beta = -0.036$; bootstrap $SE = 0.006$; $p < 0.001$) or an incumbent ($\beta = -0.019$; bootstrap $SE = 0.001$; $p < 0.001$), homophily attenuates the effects of preferential activity on interpersonal dynamics. Sharing attributes with others has a motivational effect on the sender's behaviour. As the degree of similarity between sender and recipient increases, the sender's out-degree becomes a weaker predictor of his or her future activity, and the choice to communicate is likely to be more informed and based on the recipient's attributes. In this sense, a less active user can be as likely as a more active one to forge a connection when the recipient has more attributes in common with the former than with the latter. The results also support the predicted effects of focus constraints on preferential activity. We conducted a t-test on the difference in the values of P^A when users share academic unit and when they do not, and found a p value less than 0.10 when interaction occurs between an incumbent and a newcomer, and a p value of less than 0.001 when two incumbents interact. Thus, when a user has the academic unit in common with another user, the likelihood that the former will contact the latter will depend less on past activity and more on the opportunities for offline interactions offered by the common affiliation.

Popular and active users attract each other. Connections between incumbents are driven by the combined effects of past activity and popularity. The more connections an incumbent has started, and the more connections another incumbent has received, the more likely the two users are to become, respectively, the sender and the recipient of a new connection. We conducted a t-test on the difference between values of P and P^A when only incumbents interact, and found that P is statistically significantly larger than P^A ($p < 0.01$). Thus, users' popularity exerts a greater

effect on interpersonal dynamics between incumbents than users' past activity. Moreover, the joint role of popularity and past activity varies as a function of the degree of similarity between sender and recipient. The estimated regression coefficient for P^I on number of shared attributes between incumbents ($\beta = -0.010$; bootstrap $SE = 0.001$; $p < 0.001$) indicates that homophily has mitigating effects on preferential involvement. Popularity and past activity have weaker effects on the generation of new connections between incumbents as they become more socially similar. Connections between incumbents are likely to be forged even between a less active sender and a less popular recipient. The attributes of a fairly unpopular recipient can amplify his or her attractiveness to the sender, while at the same time offsetting the negative effects that a limited past activity can have on the sender's likelihood to communicate. Thus, our prediction that homophily attenuates the effects of preferential involvement on interpersonal dynamics is supported. The results also add support in favour of our claim that focus constraints exert mitigating effects on preferential involvement. A t-test conducted on the difference between the values of P^I when users share academic unit and when they do not produced a p value less than 0.001. Sharing the same academic unit attenuates the joint effects of popularity and past activity on the addition of new connections between incumbents. While the results do not lend support to the idea that users may prefer to join the online community by contacting someone they already met offline due to the shared focus, we found evidence that being already part of the online community amplifies the perceived potential benefits that an individual may gain from replicating online an interaction that already exists offline.

6 Discussion and Conclusion

In this paper, we aimed to explore the potential of recent theoretical and methodological advances in network science for studying patterns and dynamics of communication in an electronic environment. We were motivated by the expectation that the use of time-stamped online messages would allow us to infer the general regularities governing the initiation and progression of interpersonal communication. The findings clarify and extend past research by casting light on critical issues that tend to be overlooked in network modelling efforts, including the vulnerability of network properties to measurement assumptions, the role played by edge directedness in network growth, and the effects of social constraints on growth mechanisms. In particular, our results can make scholars more critical of theoretical frameworks that assume certain properties as the common features of many types of networks (Barabási and Albert 1999, Jeong et al. 2003, Watts and Strogatz 1998), while others as the distinctive signature of social networks (Jin et al. 2001, Newman and Park 2003).

One the one hand, in close analogy to many other social (e.g., Davis et al. 2003, Holme et al. 2004, Ingram and Roberts 2000, Newman 2001b, Uzzi and Spiro 2005) and non-social networks (e.g., Albert et al. 1999, Faloutsos et al. 1999, Watts and Strogatz 1998), we found evidence of "small-world" properties, in that the network shows i) average distances between vertices that are comparable to those found in a corresponding random network, and ii) a level of clustering greater than would be expected by chance. In this sense, the network belongs to the class of "small-world networks" (Watts and Strogatz 1998). In particular, our network, like much larger ones (e.g., Albert et al. 1999), exhibits small average distances. Moreover, results indicate that distances remain small regardless of window width and timing of sampling (cf. Kossinets and Watts 2006). This lends further support to the idea that most real-world networks are indeed "small worlds".

On the other, we found evidence of "scale-free" behaviour (Barabási and Albert 1999). The findings indicate that the empirical distributions for the number of incoming and outgoing connections have a fat-tailed form. This result is somewhat surprising. As a few scholars have pointed out (e.g., Jin et al. 2001, Watts 2004), when the cost for creating and maintaining additional connections or searching for new contacts prevents individuals from having an indefinitely large number of social relationships with others, the distribution of vertex degrees is expected to be truncated and yield a scaling region with a characteristic cut-off, or even become strongly peaked around a well-defined mean (see e.g., Amaral et al. 2000, Bernard et al. 1988, Fararo and Sunshine 1964). Our results do not represent the first reported evidence of "scale-free" behaviour for social networks (see e.g., Barabási and Albert 1999, Barabási et al. 2002). In particular, fat-tailed degree distributions were also found for other online communication networks such as email networks (Ebel et al. 2002, Kossinets and Watts 2006) and online communities (Holme et al. 2004). However, compared to the online community we examined here, in which users were in spatial proximity with one another, other forms of online communication are likely to cover a broader range of social interactions. These include relationships between individuals who never meet offline, and very sporadic weak-tie relationships that do not absorb much of an individual's time nor require affective investment. The surprising side of our results comes from the fact that, in networks like ours, online communication is intended primarily to reproduce, integrate, and foster face-to-face interaction. When online communication is strongly correlated with offline social interaction, the cost of handling connections online is expected to reflect closely the non-zero cost associated with offline interaction (Wellman and Haythornthwaite 2002). If this were the case, due to the imposed constraints on communication, connections should be homogeneously distributed among individuals (Amaral et al. 2000). Our results do not support

this hypothesis.

The empirical analysis we carried out is not without limitations. The major one, shared by other research in this area (cf. Wellman and Haythornthwaite 2002), comes from the quality of the data. Even though care has been taken to filter out messages from users who were only testing the system (Kossinets and Watts 2006), some of the messages included in the analysis might still fail to reflect genuine social interaction. While we believe these are likely to represent a minority in the dataset, they might have affected the observed "scale-free" behaviour of the network. Moreover, the quality of the personal details provided by the users as they joined the network represents another source of potential inaccuracy and ambiguity of the data. We did not investigate how truthful and reliable these details are. Finally, the fairly limited size of the dataset is another clear limitation of this research, as this makes it difficult to assess and compare results with generalizations drawn from other empirical studies of much larger-scale networks (Dorogovtsev and Mendes 2003, Chapter 3). A larger and more detailed dataset would have certainly improved the quality of our study.

An important issue that future work should address is concerned with network evolution and its underpinning mechanisms (Banks and Carley 1996, Snijders 2005). In qualitative agreement with other studies (e.g., Hinds et al. 2000, Kossinets and Watts 2006, Louch 2000, Powell et al. 2005, Reagans 2005), our findings suggest that preferential linking mechanisms drive network evolution not in isolation, but in combination with social processes arising from relational content and spatial constraints. These results require further investigation. For example, while we found support for the mediating effects of homophily on preferential linking, how the probability of an edge varies as a function of vertex similarity still remains to be evaluated. Drawing on our data, future research should examine the extent to which the probability of two vertices being connected by an edge is proportional to the number of characteristics shared by the two vertices. In addition, since we found an unexpected increasing trend for clustering, it may be conjectured that individuals sharing a common neighbour are more likely to develop a social relationship than individuals with no common neighbour (Davis 1970, Holland and Leinhardt 1970, Rapaport 1953). This hypothesis of triadic closure can be tested by assessing how the probability of connections between vertices varies with the number of neighbours they have in common.

In addition to providing a platform for further theorising on dynamic social networks, our study also has the potential to inform management practice. The observed "small-world" properties have managerial implications for information diffusion and social capital. The self-organisation of the network into a compact structure with small distances between users suggests

that information can travel rapidly and reach most parts of the network accurately without requiring additional investments (cf. Davis et al. 2003, Newman 2001b, Uzzi and Spiro 2005). While it still remains to be empirically investigated whether network compactness can be detrimental to performance in the long run, especially when exploitation is facilitated at the expense of exploration (cf. Lazer and Friedman forthcoming), nonetheless velocity and fidelity of information transmission are likely to improve network performance in the short term. At the same time, while some long-range connections serve as shortcuts that shrink distances, other connections are consumed locally to create clusters and cohesive groups. Mechanisms of social capital are then activated at the local level as triadic closure is likely to produce reputation effects, encourage trust between users, enhance propensity for cooperative behaviour, and make misuse of communication more likely to be detected and punished (Coleman 1988, Davis et al. 2003, Ingram and Roberts 2000, Levine and Kurzban 2006, Louch 2000, Uzzi and Spiro 2005). In this work we did not attempt to identify the forces promoting triadic closure such as, for example, homophily or the users' common interests in the same topics. Further investigation of these forces would help expose the structure of topically organised local communities, and facilitate the design of the appropriate methods for charting interests, accessing the available information, and devising more effective marketing campaigns (Kleinberg and Lawrence 2001).

The findings on the network "scale-free" behaviour have managerial implications for information diffusion, communication security, robustness, and information searching strategies. If the "scale-free" topological properties we uncovered turn out to be a fundamental feature of electronic communication networks (cf. Ebel et al. 2002), they suggest how fragile and yet at the same time robust communication can be when conducted in an electronic environment. On the one hand, the "scale-free" topology of the network makes it vulnerable to the unilateral action of the highly connected individuals. The propagation of viruses is facilitated, and the potential gain from malfeasance is amplified, by an inhomogeneous connectivity structure (Albert et al. 2000). Along their multiple connections, hubs can spread viruses and inaccurate or even disruptive information, reaching most individuals within the network. Similarly, since the network is held together by a minority of highly connected individuals, the effectiveness and pervasiveness of communication are largely dependent on them. Should they decide to leave and sever their connections, the network would break apart and communication would fail.

On the other, it is precisely by becoming aware of the role played by hubs that managers can devise appropriate strategies for improving the security of communication and enhancing the effectiveness of information diffusion. By targeting prevention efforts at highly connected individuals, the spreading of viruses and disruptive behaviour can be monitored and network

compactness and integrity can be strengthened. Moreover, the role of hubs as "opinion leaders" can be exploited by purposefully directing information campaigns to them in order to sustain the emergence of common practices, aid the spread of ideas and fads, promote the diffusion of innovations, improve sociability, and shape a sense of community. At the same time, the hub-dominated topology makes the network fairly robust against the random removal of individuals and/or failure of communication channels (Albert et al. 2000). As accidental removals and failures disproportionately affect the poorly connected individuals, the network will remain resilient and break apart only when a significantly large proportion of individuals and communication channels have been removed. Even the accidental removal of a single hub will not be fatal for the network. In fact, our findings show that the network has a decentralised topology, self-organised into a number of hubs that share the control and monitoring of connections. This makes the network robust against accidental removal of a single hub as the remaining hubs will still succeed in holding the network together.

Finally, the topological properties of the network can be exploited to devise efficient strategies for searching and retrieving information. By biasing the routing of queries towards the most connected individuals (Adamic et al. 2001), the network can be navigated and searched with larger efficiency than would be the case if individuals were asked to contact all their nearest neighbours (broadcast search) or only one randomly selected neighbour (random walk search). Thus, understanding the network hub-dominated topology has profound managerial implications not only for securing a fast and reliable communication infrastructure, but also for designing the appropriate interventions and search engines that help locate information and resources in a distributed database.

References

Adamic, L.A., R.M. Lukose, A.R. Puniyani, B.A. Huberman. 2001. Search in power-law networks. *Physical Review E* **64** 046135.

Aiello, W., F. Chung, L. Lu. 2000. A random graph model for massive graphs. *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*. Association of Computing Machinery, New York, 171–180.

Albert, R., A.L. Barabási. 2000. Topology of evolving networks: Local events and universality. *Physical Review Letters* **85** 5234.

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

Albert, R., H. Jeong, A.L. Barabási. 2000. Error and attack tolerance of complex networks. *Nature* **406** 378–382.

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

Banks, D. L., K. M. Carley. 1996. Models for network evolution. *Journal of Mathematical Sociology* **21(1-2)** 173–196.

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

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

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

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

Bollobás, B. 1985. *Random Graphs*. New York: American Press.

Bolton, P., M. Dewatripont. 1994. The firm as a communication network. *Quarterly Journal of Economics* **109** 809–839.

Burt, R. S. 2000. Decay functions. *Social Networks* **22** 1–28.

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

Cortes, C., C. Pregibon, J. Volinsky. 2003. Computational methods for dynamic graphs. *Journal of Computational and Graphical Statistics* **12:4** 950–970.

Davis, G.F., M. Yoo, W.E. Baker. 2003. The small world of the American corporate elite, 1982-2001. *Strategic Organization* **1(3)** 301–326.

Davis, J.A. 1970. Clustering and hierarchy in interpersonal relations. *American Sociological Review* **35** 843–851.

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

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

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

Faloutsos, M., P. Faloutsos, C. Faloutsos. 1999. On power-law relationships of the internet topology. *Computer Communication Review* **29** 251–262.

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

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

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

Guimerà, R., B. Uzzi, J. Spiro, L.A.N. Amaral. 2005. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308** 697–702.

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

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

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

Ingram, P., P.W. Roberts. 2000. Friendships among competitors in the Sydney hotel industry. *American Journal of Sociology* **106(2)** 387–423.

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

Jin, E.M., M.. Girvan, M.E.J. Newman. 2001. The structure of growing social networks. *Physical Review E* **64** 046132.

Kleinberg, J., S. Lawrence. 2001. The structure of the web. *Science* **294** 1849–1850.

Kochen, M. 1989. *The Small World*. Norwood, NJ: Ablex.

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

Krapivsky, P.L., S. Redner, F. Leyvraz. 2000. Connectivity of growing random networks. *Physical Review Letters* **85** 4629.

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

Lazer, D.M.J., A. Friedman. forthcoming. Parallel problem solving: The social structure of exploration and exploitation. *Administrative Science Quarterly* .

Levine, S.S., R. Kurzban. 2006. Explaining clustering in social networks: Towards an evolutionary theory of cascading benefits. *Managerial and Decision Economics* **27** 173–187.

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

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

McPherson, J.M., L. Smith-Lovin, J.M. Cook. 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.

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

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

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

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

Newman, M.E.J., S. H. Strogatz, D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E* **64** 026118.

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

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. 2005. Preferences, identity, and competition: Predicting tie strength from demographic data. *Management Science* **51(9)** 1374–1383.

Rothaermel, F.T., S. Sugiyama. 2001. Virtual internet communities and commercial success: Individual and community-level theory grounded in the atypical case of timezone.com. *Journal of Management* **27(3)** 297–312.

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

Smith-Doerr, L., W.W. Powell. 2005. Networks and economic life. N. Smelser, R. Swedberg, eds., *The Handbook of Economic Sociology*. Princeton University Press, Princeton, NJ, 379–402.

Snijders, T.A.B. 2005. Models for longitudinal network data. P. Carrington, J. Scott, S. Wasserman, eds., *Models and Methods in Social Network Analysis*. Cambridge University Press, New York, 215–247.

Uzzi, B., J. Spiro. 2005. Collaboration and creativity: The small world problem. *American Journal of Sociology* **111** 447504.

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

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

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

Wellman, B., C.A. Haythornthwaite. 2002. *The Internet in Everyday Life*. Oxford: Blackwell.