Patterns and Dynamics of Users' Behavior and Interaction: Network Analysis of an Online Community

Pietro Panzarasa and Tore Opsahl
Queen Mary University of London, School of Business and Management, London E1 4NS, United Kingdom. E-mail: {p.panzarasa, t.opsahl}@qmul.ac.uk

Kathleen M. Carley
Institute for Software Research International, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: kathleen.carley@cs.cmu.edu

This research draws on longitudinal network data from an online community to examine patterns of users' behavior and social interaction, and infer the processes underpinning dynamics of system use. The online community represents a prototypical example of a complex evolving social network in which connections between users are established over time by online messages. We study the evolution of a variety of properties since the inception of the system, including how users create, reciprocate, and deepen relationships with one another, variations in users' gregariousness and popularity, reachability and typical distances among users, and the degree of local redundancy in the system. Results indicate that the system is a "small world" characterized by the emergence, in its early stages, of a hub-dominated structure with heterogeneity in users' behavior. We investigate whether hubs are responsible for holding the system together and facilitating information flow, examine first-mover advantages underpinning users' ability to rise to system prominence, and uncover gender differences in users' gregariousness, popularity, and local redundancy. We discuss the implications of the results for research on system use and evolving social networks, and for a host of applications, including information diffusion, communities of practice, and the security and robustness of information systems.

Introduction

The conceptualization of many social systems as networks made of nodes (or vertices) joined together by ties (or edges) has long facilitated the study of the origins, evolution, and consequences of patterns of social interaction (Smith-Doerr & Powell, 2005; Wasserman & Faust, 1994). Empirical and theoretical work on social networks has traditionally been an interdisciplinary endeavor, characterized 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 (Bernard, Kilworth, Evans, McCarty, & Selley, 1988; Fararo & Sunshine, 1964; Foster Rappoport, & Orwant, 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 exploring the effects that social networks have on outcomes relevant to individuals and organizations, including information and disease diffusion (Valente, 1995), group and team performance (Guetzkow & Simon, 1954), and trust building and cooperation (Coleman, 1988), to name only a few.

It has now become widely accepted that although 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 (Bernard, Kilworth, Kronenfeld, & Sailer, 1984). For most of the past 50 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

Received June 30, 2008; revised November 21, 2008; accepted November 21, 2008

 • Published online 2 February 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21015

strong social interaction could differ considerably from one person to another (Bernard et al., 1988; Fararo & 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 labor-intensive and onerous to administer, and therefore most of the networks studied were of a fairly limited size, often comprising only a few tens (e.g., Bernard et al., 1988) or hundreds (e.g., Fararo & Sunshine, 1964) of people.

The lack of high quality and large-scale network datasets has delayed progress in the statistical modeling of the network structure and behavior. In recent years, however, two major developments have prompted substantial advances in research on networks. First, more accurate studies of large-scale social networks (Barabási et al., 2002; Newman, 2001; Watts & Strogatz, 1998) 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 behavioral 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 modeling the structure and behavior of complex networks (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004). These converging developments and circumstances have motivated many scholars to study large-scale statistical properties and capture in quantitative terms the organizing principles encoded in the structure of a variety of complex networks.

As noted by a few scholars (Burt, 2000; Smith-Doerr & Powell, 2005; McPherson, Smith-Lovin, & Cook, 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., Ahn, Han, Kwak, Moon, & Jeong, 2007; Barabási et al., 2002; Holme, Edling, & Liljeros, 2004; Kossinets & Watts, 2006; Powell, White, Koput, & Owen-Smith, 2005; Uzzi & Spiro, 2005), recent empirical work has been primarily concerned only with the static structural features of social networks, while the evolution of interaction patterns and their underpinning organizing 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 & Carley, 1996; Burt, 2000; Snijders, 2005). The extent to which individuals' choices at the local level dynamically affect the global network structure 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.

With this article, we take a first step in this direction. Drawing on recent theoretical and methodological advances in network science, our goal is to study patterns and dynamics of social interaction among individuals by examining how an online communication network is assembled over time as users join and leave the network, and ties are created and severed. To this end, we empirically examine an evolving 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 (Baym, Zhang, & Lin, 2004; Boase, Horrigan, Wellman, & Rainie, 2006; Boase, 2008; Stern, 2008; Wellman et al., 1996; Wellman, 1999; Wellman & Haythornthwaite, 2002; Wellman et al., 2006). The nodes of the network are the users, whereas ties between users are established by online messages. Ties 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 crystallized 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 constructed over time as a result of the addition and removal of nodes and ties.

The idea of using electronic information systems to uncover the large-scale statistical properties of communication networks is not new. Recent empirical studies have investigated network datasets drawn from phone records (Aiello, Chung, & Lu, 2000; Onnela et al., 2007), e-mail log-files (Diesner, Frantz, & Carley, 2006; Ebel, Mielsch, & Bornholdt, 2002; Guimerà, Danon Díaz-Guilera, Giralt, & Arenas, 2006; Kossinets & Watts, 2006), and registers of users' activities in online communities (Ahn et al., 2007; Holme et al., 2004). The dataset we analyze in this paper belongs to the last category. However, unlike recent studies of online communities (Ahn et al., 2007; Holme et al., 2004; Rothaermel & 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 (Mok, Wellman, & Basu, 2007; Monge, Rothman, Eisenberg, Miller, & Kirste, 1985; Boyer O'Leary, & Cummings, 2007; Wellman & Haythornthwaite, 2002). In addition, unlike other recent studies of online communication networks (Ebel et al., 2002; Guimerà et al., 2006), our work focuses on network dynamics and thus helps shed light on the social processes that are responsible for the structural properties of the system (Banks & Carley, 1996, Snijders, 2005).

Our unique dataset allows us to study the evolution of an online community since its inception. We can thus follow the trajectory of users' online behavior since the time at which they joined the community and started sending or receiving their first message. The dataset includes all users and all interactions among them that took place in the observation period, thereby overcoming the limitations of the studies based on alternative network sampling methods, such as snowball sampling (Ahn et al., 2007). By making explicit use of the time at which each user joins the community and each tie is formed, and by examining the evolution of users' behavior and interaction patterns, our study paves the way toward a better understanding of the distinctive organizing principles that govern interpersonal dynamics in computer-mediated communication. In so doing, while inspiring the development of network and organization theories that place a special emphasis on dynamics, this article seeks to extend current research on system use (Burton-Jones & Gallivan, 2007; Lee, Kozar, & Larsen, 2003; Szajina, 1993; Tyre & Orlikowski, 1994; Venkatesh, Morris, Davis, & Davis, 2003) and online social interactions (Boase et al., 2006; Boase, 2008; Wellman et al., 1996; Wellman & Haythornthwaite, 2002). We offer an insight into users' social behavior by focusing on how they initiate, reciprocate, deepen their relationships with one another, and create densely connected local environments. Our analysis also sheds a new light on the role of heterogeneity in system use (Amiel & Sargent, 2004; Compeau, Higgins, & Huff, 1999; Taylor, 2004; Zmud, 1979). We challenge the idea of a prototypical user by uncovering substantial variations among users in terms of gregariousness, popularity, and local redundancy. Our study shows how a select minority of users manage to rise to system prominence and take on the role of hubs responsible for holding the system together and reducing distances among other users. We explore the opportunities that users have to acquire such role and how these opportunities vary depending on the time at which users join the system.

Our analysis also aims to provide a platform for drawing a host of practical implications for information diffusion, search, and retrieval, as well as for the management and design of information systems. The lack of a centralized control in most electronic forms of communication may suggest that moderators and managers are deprived of the appropriate measures for presiding over the flow of information and for promoting valuable communication (Bolton & Dewatripont, 1994). Yet our research sheds light on the opportunities for effective control and management that arise precisely as a result of the distributed nature of communication. In particular, we show how the principles that govern the structure and evolution of the system can be used to devise the appropriate strategies for channeling information in the right direction and making sure that it quickly reaches most users in the system, for searching and retrieving information in a timely fashion, and for protecting the system from the negative global effects of users' misuse of communication.

We proceed as follows. First, we describe our research setting and how the longitudinal network dataset was created. We then analyze the structure and dynamics of social interaction among users and focus on a number of properties including the growth patterns followed by the system, the degree of gregariousness and popularity of users, the extent to which users are reachable from others along relatively short paths, and users' tendency to cluster into tightly knit yet interconnected social circles. The article concludes with an interpretation of the findings and a discussion of their implications for theory and practice.

Network Data and Methods

We study the network structure and 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 among 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 (Ahn et al., 2007; Holme et al., 2004). Unlike e-mail 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. The network dataset covers the period from April to October 2004. Our analysis includes all users that sent or received at least one message during that period. To ensure that our data do indeed reflect interpersonal communication, users who simply registered but did not communicate were excluded from this analysis. Furthermore, to filter out spamming activities, two companies that gained access to the online community with the purpose of mass-communication were excluded. Technical support staff, such as moderators, with the only aim of facilitating the smooth functioning of the community were also excluded. A total of 1,899 users was recorded, of whom 58.87% were male.

During the observation period, users sent a total number of 59,835 online messages. To create the longitudinal network dataset, we compiled a register of all these messages, and for each of them we recorded the time-stamp, sender and recipient. To ensure privacy protection, before we received the data, all individual identifiers, such as usernames, e-mail and IP addresses, were removed. Each user was randomly assigned an identification number. For each online message, the information we obtained thus included only the identification numbers of sender and receiver, and the time at which the message was sent. Moreover, to circumvent the legal requirement of obtaining a written consent from every user, the content of all messages was not made available to us.

In constructing the network, we adopted a weak notion of social relationship (Constant, Sproull, & Kiesler, 1996; Granovetter, 1973). A directed tie is established from one user to another if one or more messages have been sent from the former to the latter. A total of 20,296 directed ties was recorded during the observation period. Our definition of a directed tie thus implies that a social relationship between two users exists even when only one message has been sent from one user to the other. Our choice of setting a low threshold of one message for representing social relationships was motivated by the fact that the average strength of a directed

DOI: 10.1002/asi

tie (i.e., the ratio between number of messages sent and number of directed ties) is 2.95, and approximately 80% of all directed ties have a strength (i.e., number of messages sent along them) of 3 or less. In this case, setting higher thresholds for social relationships would have added isolates and reduced the connectivity of the network. For the same reason, and to maintain directionality of ties, we did not follow Kossinets and Watts (2006) in focusing only on reciprocated ties and then transforming them into undirected ones.

We analyzed the longitudinal network dataset in two steps. First, at any time point we constructed the instantaneous cumulative network reflecting all the social interactions that took place before that point, *since the time at which the community started*. In what follows, we report measurements only from day 7 as before that time statistics are poor. The network is then measured with a daily frequency. This choice was motivated by the fact that sampling with a daily frequency produced a reasonable solution to 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 minimizing the bias resulting from errors in time measurements (by using a sampling period that is not too small). On average, the number of messages sent shows a significant drop in the early hours of the morning, with the minimum level at 7:00 am, 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 7:00 am instead of midnight so as to minimize the interruption of the ongoing flow of social interaction.

Second, we constructed the network by imposing lifespans to social relationships. In fact, one limitation of measuring a cumulative network stems from the assumption that social relationships, once established, never decay. To overcome this limitation, we introduced lifespans to make sure that, if two users do not continue to communicate, their connection will be severed. The length w of the lifespan thus determines which past events are taken into account to generate the instantaneous structure of the community at any point in time. More generally, the choice of w should be motivated by the analysis of which past events are relevant to the current state of the community (Kossinets & 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.

In our analysis, we adopted a lifespan of 3 weeks and, in analogy with the cumulative network, we measured the network created with this lifespan with a daily frequency. Our choice of the 3-week lifespan was motivated by a number of considerations. First, a weekly period seems to be an appropriate temporal unit of analysis as users' 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 (Howard, Rainie, & Jones, 2002). Second, among the various multiples of the one-week unit, we chose the 3-week lifespan because 3 weeks represent the best approximation of the time at which the rates of increase in messages and in new tie formation stabilize, while the system is still rapidly growing. In addition, this choice finds further support in the observation of the distribution of dyadic response time: about 96.86% of all reciprocated ties are in fact reciprocated within 3 weeks.

To test the robustness of the results with respect to the length of the lifespan, we also constructed the network (and measured it with a daily frequency) by using lifespans of 2 and 6 weeks. Our choice was motivated as follows. On the one hand, the 2-week lifespan was justified by the fact that there are spells of reduced activity (e.g., between day 62 and 73) that could be hardly detectable using longer lifespans. On the other, we also considered a 6-week lifespan because the sixth week represents the end of a phase of significant network growth, followed by a phase in which users and acquaintances grow more smoothly toward their cumulative asymptotic levels. In fact, after 6 weeks the system already includes 77% of all users and 67% of all acquaintances. Moreover, the rate of increase in new messages shows a significant drop around the same time, when about 67% of all messages have already been sent.

Users, Acquaintances, and Interaction Patterns

Each user has an an out-degree k^O, defined as the number of ties originating from the user, and an in-degree k^I, defined as the number of ties terminating at the user (Wasserman & Faust, 1994). We define acquaintances in terms of user's out-degree. This ensures that users choose who their acquaintances are. 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. Because users can send more than one message to the same contact, the total number of messages (59,835) is different from the total number of acquaintances (20,296). The average number of acquaintances is 10.69.

Figure 1 shows the evolution of users and average number of acquaintances when social relationships cannot dissolve (i.e., in the cumulative network) and when they can (i.e., with lifespans). A new user is recorded when he or she sends or receives a message for the first time. Moreover, a user gains one new acquaintance as soon as he or she sends a message to another user for the first time.

Unlike what recent models of growing networks suggest (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004), Figure 1 indicates that the online community is not constructed uniformly over time. By contrast, the community 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 6 weeks, 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 community

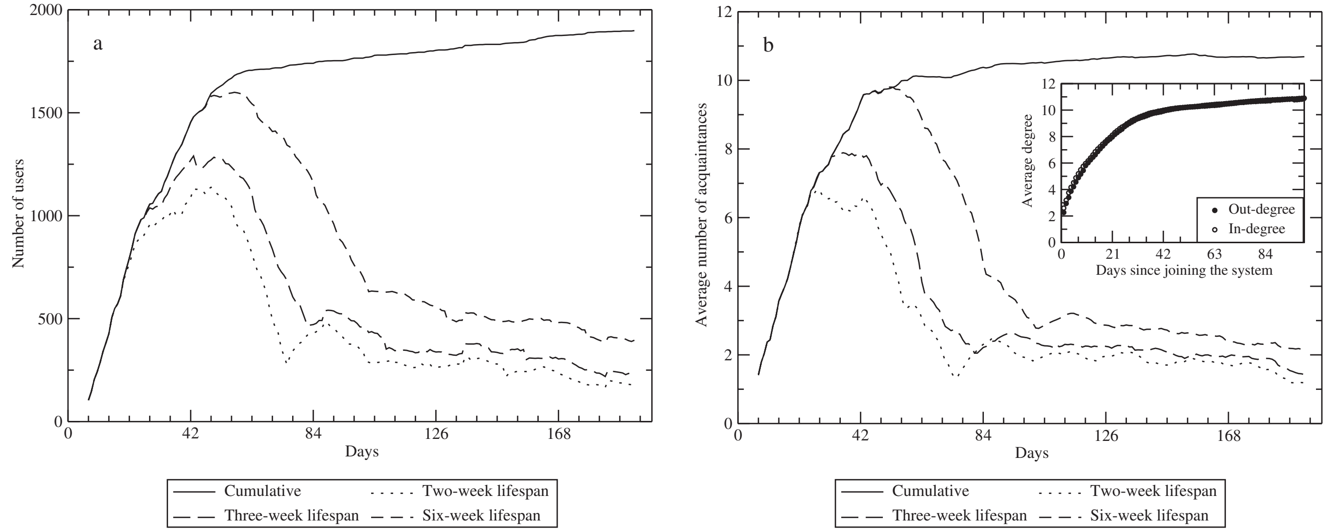

FIG. 1. (a) Evolution of users; (b) Evolution of average number of acquaintances. Inset refers to average out- and in-degrees as a function of the number of days since users joined the system.

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 & Mendes, 2003; Faloutsos, Faloutsos, & Faloutsos, 1999). The main reason for "accelerated" growth is that users create new acquaintances not only as they join the community, but also in subsequent stages. In agreement with recent theoretical and empirical studies of network growth (Albert & Barabási, 2000; Barabási et al., 2002; Guimerà, Uzzi, Spiro, Amaral, 2005), the community thus evolves as a result of not only the contributions of newcomers but also the continuing activities of incumbents. Moreover, in our specific setting where users are geographically co-located, 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 (Boase et al., 2006; Boase, 2008; Stern, 2008; Wellman, 1999; Wellman et al., 2006). As the system grows, users become more interconnected and, as shown in Figure 1b, the average number of acquaintances rapidly increases.

In the second phase, including the remaining weeks, the community is no longer driven by "accelerated" growth, but is instead characterized by structural stability. Specifically, this phase is marked by a noticeable decline and convergence of the daily rates of increase in users and acquaintances. This leads to a reduction, for the cumulative network, of the rate of increase in the average number of acquaintances, which then remains almost stable for the rest of the observation period. Since the spring term ended on June 19 (i.e., day 66 in our dataset), the phase of structural stability does not seem to be triggered by the beginning of the summer break. Moreover, the community does not show a noticeable increase in users and communication in correspondence of the start of the fall term (September 20, i.e., day 159 in our dataset). In this respect, the dataset is fairly robust against seasonal biases and is thus likely to reflect evolutionary patterns of genuine social interactions among the users of the community.

So far the analysis has embodied a dimension of time that emphasizes the system's macro perspective: the time elapsed since the system started. This notion of time enables us to study how the overall system evolves, and at what stage certain structural properties emerge. In addition to this macro perspective, there is another dimension of time that helps further unearth the micro processes underpinning the evolution of the system: the time elapsed since users joined the system. Looking at this micro dimension of time enables us to follow users' behavior since the very first day they started using the system. In so doing, we can investigate whether on average users acted in the same way over time or, conversely, showed a tendency to concentrate the creation of their social relationships during any specific period of time with respect to the day they joined the community. In the inset of Figure 1b, we plotted users' average out- and in-degrees as a function of the number of days they have remained in the system. We selected all users that remained in the system for at least 100 days, and for these users we measured their system use since the day they joined the community. As indicated by the figure, both out- and in-degree show a rapid increase in the first 6 weeks, and then they stabilize in the remaining weeks.

This result points to an interesting parallel between our two dimensions of time, and especially between the length of the phase in which the overall system developed and the length of the phase in which users' personal networks expanded. Not only does the rapid growth of the system occur during the first 6 weeks since the day the system started, but also users tend to concentrate their activities of creation of new acquaintances in their first 6 weeks of system use, *regardless of the time at which they joined the system.* On the one

DOI: 10.1002/asi

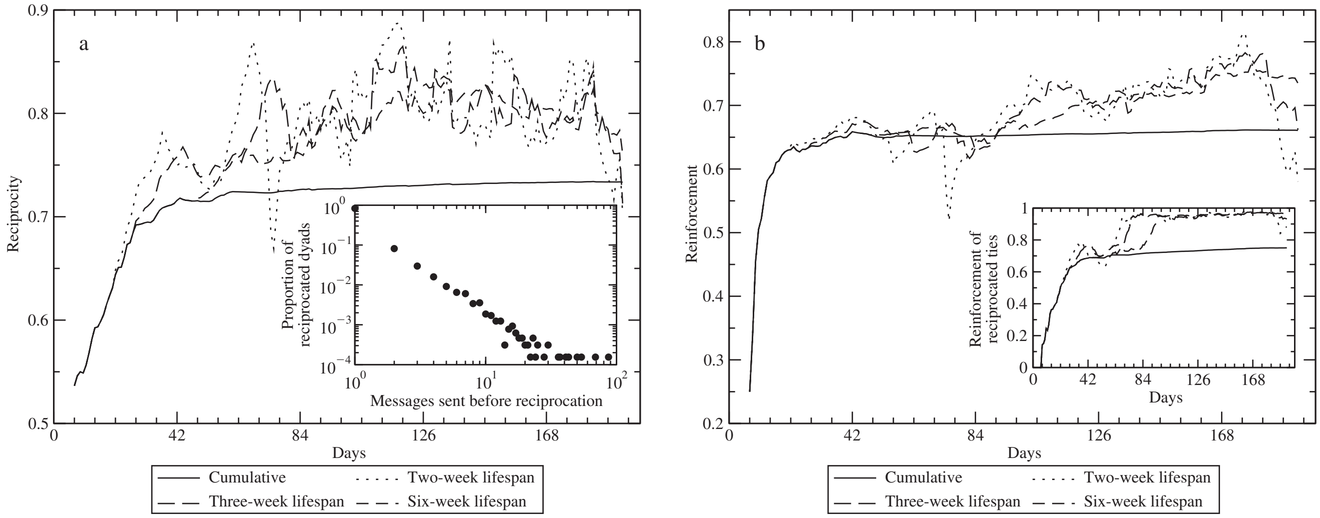

FIG. 2. (a) Evolution of the proportion of relationships that are reciprocated; inset shows the proportion of reciprocated relationships plotted against the number of messages sent before reciprocation occurs; (b) Evolution of the proportion of messages used to reinforce existing relationships; inset refers to the evolution of reinforcement of reciprocated relationships.

hand, most users joined the system in the first few weeks, during which the rapid increase of their personal networks contributed toward the "accelerated" growth of the overall system. On the other, a minority of users joined the system in subsequent weeks. These users also expanded their networks quickly soon after joining the system; however, this increase in network activities has been offset by the reduced activities of the majority of users that joined the system in the previous weeks, which resulted in the observed stability of the macro properties of the overall system.

The macro dimension of time, embodied in our analysis of the two-fold regime of network evolution, can be further investigated by looking at variations in interaction patterns since the first day the system started. This allows us to highlight the contribution to network evolution and dynamics of system use of two processes often neglected in modeling efforts (Burton-Jones & Gallivan, 2007; Lee et al., 2003; Szajina, 1993; Tyre & Orlikowski, 1994): "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 toward the rapid growth in average number of acquaintances as users return the attention they receive from others by replying to their messages (Gouldner, 1960; Plickert, Côté, & Wellman, 2007). We find it interesting that the vast majority of reciprocated relationships are reciprocated as soon as a single message has been sent. As shown in the inset of Figure 2a, in only a minority of cases does it take many messages to get a contact to reply and reciprocate a relationship. At the same time, the divergence between the rate of increase in total messages per user and the rate of increase in average number of acquaintances becomes more pronounced as time goes by. This is a clear indication of another 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; Onnela et al., 2007). While reciprocity, by creating new ties, 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 their existing ties to communicate with old acquaintances. In this sense, Figure 2b suggests that (a) most activities of "network construction" occur in the first phase, and (b) 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.

We also examined how reciprocity combined with reinforcement over time. To this end, we tested users' tendency to invest preferentially in relationships with contacts that have already replied to their messages (Plickert et al., 2007). Of the total number of messages sent to existing acquaintances, we measured the fraction sent to acquaintances that have already reciprocated. As shown in the inset of Figure 2b, we found that this fraction for the cumulative network increases during the initial growth period until it reaches its asymptotic value of about 0.7. Users value reciprocation and show a tendency to strengthen their relationships with those from whom they receive attention (Gould, 2002; Gouldner, 1960). The deepening of social relationships thus stems from a combination of reinforcement and reciprocity (Plickert et al., 2007).

These results on the non-uniform nature of network evolution show an interesting similarity with what has been found in mainstream research on small (offline) groups (Levine & Moreland, 1990). In particular, the observed two-fold regime of network evolution, organized into two distinct stages of "network construction" and "network use," is in agreement with, and provides further support in favor of, established stage models of group formation and development (e.g., Tuckman, 1965; Tuckman & Jensen, 1977). These models suggest that social relationships are created primarily at the onset of group formation when individuals learn about one another and establish expectations about one another's behavior. In subsequents phases, individuals tend to reinforce their relationships and develop a strong sense of group identity. Although we could not test for the emergence among users of a sense of belonging and group commitment, our results nonetheless point to a similar pattern of evolution in which the creation of ties with new acquaintances takes place predominantly in the first few weeks, while the remaining period is devoted mainly to reinforcing ties with old acquaintances.

We now introduce a refinement to our macro perspective of time by imposing a lifespan to social relationships (Kossinets & Watts, 2006). In this case, when social relationships do not last forever, past events lose relevance for the current state of the system, depending on when they took place. Elapsed time is thus measured not since system inception, but since the day at which the oldest events that are still relevant for the current state of the system occurred. This helps integrate the analysis of the online community in a number of ways. When relationships may dissolve, some users and acquaintances are lost, while others are gained as the community evolves. In particular, network properties will result from the interplay between this gain and loss in users and acquaintances. An equilibrium state is thus reached when the gain and loss offset each other. As Figure 1 indicates, regardless of the length of the lifespan, users and average number of acquaintances increase in the first phase, whereas at the beginning of the second phase they rapidly decrease until they reach an equilibrium state. This decline is chiefly due to the fact that, as time goes by, the loss in users and acquaintances is greater than the gain. Furthermore, Figure 1 suggests that with lifespans stability is reached at a later stage than in the case where relationships never decay, with fluctuations becoming smoother as the lifespan gets longer.

The length of the lifespan also affects interaction patterns. As shown in Figure 2, the fluctuations of reciprocity and reinforcement around their asymptotic values become smoother as relationships are assumed to persist for a longer period of time. However, results indicate that reinforcement continues to increase over time, regardless of the length of the lifespan, with users gradually devoting more time and resources to communicating with old acquaintances. We find it interesting that when a lifespan is used, the ratio between messages sent to old acquaintances that have already reciprocated and total number of messages to old acquaintances increases to almost 100%. While this indicates that late adopters during the stability phase tend to reinforce all reciprocated relationships, it also shows that not all early adopters do so, but only those that keep using the system over time. As users that were active only in the growth phase are excluded from the analysis, the system only reflects the activities of early adopters with repeated interactions over time. Thus, when users that are active in the stability phase, regardless of when they joined the system, communicate with an old acquaintance, it is almost always the case that their communication occurs over a reciprocated relationship.

Gregariousness and Popularity

One of the crucial and often neglected issues in the analysis of system use and, more generally, of the organizing principles underpinning a system is concerned with the degree of heterogeneity in the way users behave and interact (Amiel & Sargent, 2004; Compeau et al., 1999; Taylor, 2004; Zmud, 1979). Some users may be more gregarious and establish more social relationships than others, thus creating a higher-than-average number of acquaintances. Similarly, some users may be more popular and be contacted more often than others, thus becoming acquaintances of a higher-than-average number of other users. If differences among users were pronounced, it would be more difficult to use average properties for describing how a typical user behaves. For instance, when communication is heavily dominated by few very gregarious users acting as hubs that send messages to many other users, the actual average number of acquaintances is in fact considerably higher than the number of acquaintances the typical user has (Barabási & Albert, 1999; Ebel et al., 2002; Holme et al., 2004). Moreover, the degree of heterogeneity in users' gregariousness and popularity has important implications in terms of the robustness of the system with respect to unexpected events. For instance, when few large hubs coexist with many poorly connected users, these hubs would keep the system from falling apart when users randomly leave (Albert, Jeong, & Barabási, 2000). Heterogeneity in users' behavior also affects how widely and quickly information can be spread and retrieved across a system. Hubs have an influence and information-spreading capacity that is disproportionate to their number, and can easily retrieve a large amount of information by leveraging on the abundance of their communication partners (Adamic, Lukose, Puniyani, & Huberman, 2001). Thus, the degree of variation in users' gregariousness and popularity directly translates into a number of structural properties that, in turn, affect how dynamic processes, such as information flow, unfold within the system.

We define $p(k^O)$ and $p(k^I)$ to be the fraction of users in the system that have, respectively, out-degree k^O and in-degree k^I. Thus, $p(k^O)$ is the fraction of users that have k^O acquaintances, whereas $p(k^I)$ is the fraction of users that are the acquaintances of k^I other users. The out- and in-degree distributions are frequency distributions giving $p(k^O)$ and $p(k^I)$, respectively, for each k^O and k^I (Dorogovtsev & Mendes, 2003, p.10). In networks in which ties are randomly generated, any two nodes are connected with equal probability, and therefore the out- and in-degree distributions are binomial or Poisson in the limit of a large network size (Erdös & Rényi, 1960). More specifically, as the total number of nodes becomes large, the random network shows a Poisson distribution that is sharply peaked around the mean and decays rapidly on both sides. This means that, on a random network, most nodes have the same number of ties, whereas the likelihood of generating and attracting a higher-than-average

DOI: 10.1002/asi

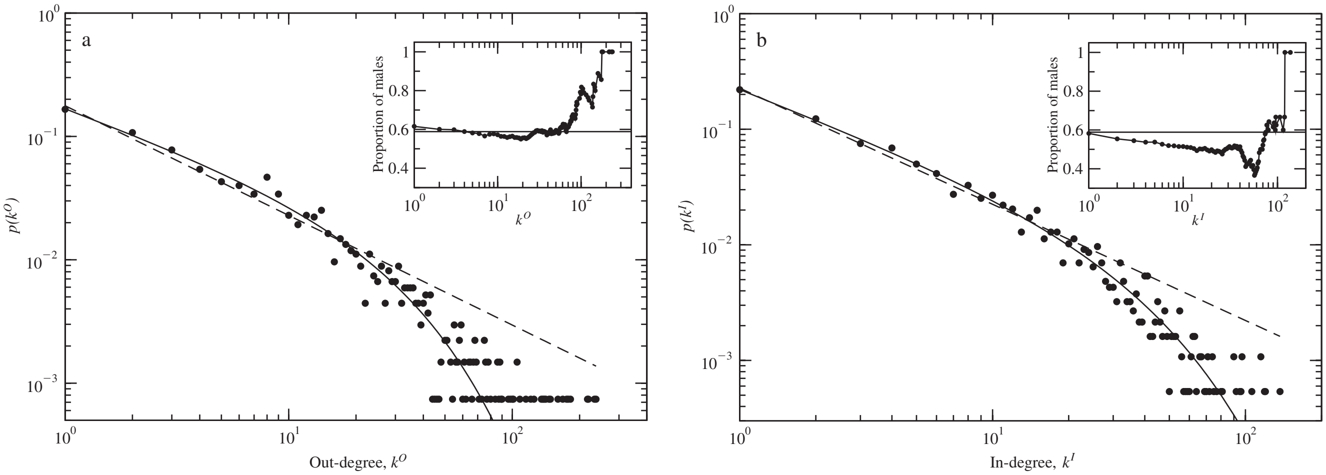

FIG. 3. (a) Out-degree distribution; (b) In-degree distribution. Insets refer to the proportion of males within subsets of users characterized by an out-degree (a) and in-degree (b) equal to or higher than a certain value.

number of ties is quite limited. However, this property is quite different from what has been observed in most real-world networks (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004). Empirical research has shown that, regardless of their nature, many real-world networks share the same structural property: they all exhibit out- and in-degree distributions that are fat-tailed (Barabási & Albert, 1999). This implies that in these networks there are nodes acting as hubs, with out- and in-degrees far above the mean. This common property can be seen as evidence of a universal organizing principle that governs the way in which many different real-world networks evolve over time.

In recent literature, there has been a significant discussion as to how to formally represent the fat-tailed degree distributions observed in real-world networks. For example, a number of authors (e.g., Albert, Jeong, & Barabási, 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 following form:

$$p(k^O) \sim (k^O)^{-\tau} \quad (1)$$

$$p(k^I) \sim (k^I)^{-\tau}, \quad (2)$$

where τ is a constant exponent. Barabási and Albert (1999) have suggested that similar degree distributions apply to a great variety of other real-world networks. When the degree distribution of a network can be formalized by a power-law function, the network is said to be "scale-free" (Barabási & Albert, 1999). The hallmark of "scale-free" networks is the high level of heterogeneity in the way nodes generate or attract connections: while the majority of nodes 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 out- and in-degree distributions of our cumulative network. These are not Poisson distributions peaked around their mean, but they exhibit a fat-tailed form. This means that there are a minority of users that are far more gregarious and popular than the average. Both distributions are well approximated by the linear behavior on the double logarithmic scale, as indicated by the straight lines in Figure 3. More precisely, they can be fitted by power-law functions with an exponent τ of 0.889 for the out-degree distribution, and 1.005 for the in-degree distribution. The power-law fit for the out-degree distribution has an R^2 of 0.9735, whereas the one for the in-degree distribution has an R^2 of 0.9895, both with p values of less than 0.001.

Moreover, as shown by the curved lines in Figure 3, the out- and in-degree distributions can also be well fitted by power-law functions with an exponential cut-off[1] (Amaral, Scala, Barthélémy, & Stanley, 2000). A number of possible explanations of the origins of a cut-off have been proposed (Barabási et al., 2002; Krapivsky, Redner, & Leyvraz, 2000). For example, Newman (2001), suggested that the cut-off was produced by the finite time frame of his analysis that prevented nodes from connecting to a large number of others. The same argument could also help explain the better fit we obtained with a cut-off. As a result of the limited period of time in which the network was measured, even the most gregarious users may have created fewer acquaintances than would be the case if the observation period were longer. Similarly, in our limited time frame, even the most popular users may have been contacted by fewer users than in a longer observation period. Time constraints thus translate into limits on communication behavior, with obvious implications in terms of how gregarious or popular users can be.

---

[1] The estimated power-law functions with exponential cut-off for the out- and in-degree distributions are as follows:

$$p(k^I) \sim (k^I)^{-0.847} e^{-k^I/30.306}, \quad (5)$$

$$p(k^O) \sim (k^O)^{-0.665} e^{-k^O/28.01}. \quad (6)$$

The fit has an R^2 of 0.9901 for the out-degree distribution, and 0.996 for the in-degree distribution, both with p values of less than 0.001.

DOI: 10.1002/asi

The findings thus lend support to the conjecture that the system belongs to the class of "scale-free" networks. Interestingly, both estimated power-law exponents are lower than two. This is the distinctive signature of a network structure that is dominated by the few highly connected nodes, and not by the majority of poorly connected ones (Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2001). This property has profound implications for the connectivity and organizing principles of the system that will be explored in the following sections. Here we focus our attention on users and their online behavior, and especially on those with high levels of out- and in-degrees.

First, we measured users' average out-strength (in-strength) as the average number of messages sent to (received from) others (Opsahl, Colizza, Panzarasa, & Ramasco, 2008). We expected hubs to be weakly connected to others, based on the conjecture that all users are homogeneously limited by the same constraints of resources and time. In this case, having more contacts should reduce the amount of resources and time spent on each of them (Burt, 1992). We were surprised to find a positive and significant ($p < 0.001$) Pearson's pairwise correlation coefficient between average out-strength (in-strength) and out-degree (in-degree) of 0.28 (0.44). This signals that hubs spend more time and resources with each of their contacts than the less connected users. Clearly, users show different dispositions toward using the system. Those with many contacts are also comfortable in using the system and thus devote a large amount of their time and resources to deepen their online relationships (Wellman & Giulia, 1999), whereas users that are reluctant to make many contacts online also tend to invest little in their relationships. Interestingly, this finding is in qualitative agreement with the "rich get richer" model that Kraut et al. (2002) suggested to explain the relationship between users' extraversion and the benefits gained from Internet use. According to that empirical study, users who are highly sociable and make many contacts are likely to obtain more social benefits (e.g., increase in social involvement and self esteem, and decrease in loneliness) from using the Internet than those who are introverted and have fewer contacts. In a similar vein, we found that extraverted users tend to exchange more messages with their contacts and are, therefore, more likely to amplify their involvement in online relationships than introverted users.

Second, we further examined heterogeneity in system use (Amiel & Sargent, 2004; Compeau et al., 1999; Tyre & Orlikowski, 1994; Zmud, 1979). To assess how unevenly gregariousness and popularity are distributed across users, for both out- and in-degree distributions we measured the normalized entropy and Pearson's skewness coefficients. Entropy is typically used to study how diverse the degrees of nodes are, and takes values ranging from zero to one (Wang, Tang, Guo, & Xiu, 2006). We found that the entropy coefficients of the out- and in-degree distributions are 0.40 and 0.42, respectively. The Pearson's skewness coefficient, based on the difference between mean and median of a distribution, measures the extent to which the majority of nodes have a degree above or below the mean. It therefore gives an indication of how asymmetric a distribution is around the mean, and ranges between -3 and 3. We found values of 1.05 and 1.12, for the out- and in-degree distributions, respectively. The findings therefore further confirm that users differ from one another in the way they communicate, and the difference is more pronounced for popularity than gregariousness (Gould, 2002). In other words, the way attention is paid to others is less unevenly distributed across users than the way attention is received.

Third, to understand how users respond to the attention they receive from others, and in particular to investigate whether the most popular users are also the most gregarious ones, we examined the relationship between users' gregariousness and popularity. To this end, we measured the Pearson's correlation coefficient between nodes' out- and in-degrees. We found a pairwise correlation of 0.83 ($p < 0.001$), thereby suggesting that users tend to balance their popularity with the attention they direct to others. Thus, not only are gregariousness and popularity unevenly distributed across users, but the imbalance seems to favor the same users on both dimensions. Users that rise to system prominence by being socially active and making contact with many others are also the most attractive targets of others' attention within the system. Interestingly, the observed correlation between high degrees of gregariousness and popularity seems to provide support in favor of recent studies on effective managerial practices and leadership that underlie the importance of not only directing, monitoring, exercising influence, and doing the talking, but also more mundane non-directive dimensions, such as listening and being receptive and well-disposed toward others (Alvesson & Sveningsson, 2003). In addition, these results cast a new light on the role of reciprocity as a driving force behind heterogeneity in system use (Plickert et al., 2007). On the one hand, as users become very popular, they also become very gregarious because they tend to return the attention they receive. Reciprocity thus transforms heterogeneity in popularity into heterogeneity in gregariousness. On the other, not all relationships in the system are reciprocated and users tend not to fully repay the attention they receive. This partial reciprocity is responsible for a less pronounced heterogeneity in gregariousness than in popularity (Gould, 2002).

Finally, we carried out a comparative analysis of male and female users across various levels of gregariousness and popularity. Recent research on gender difference in behavior and attitudes toward computer-mediated communication has indicated that, even though the proportion of women using the Internet is nearly equal to that of men, women are nonetheless heavier users of e-mail than men (Boneva & Kraut, 2002; Cummings & Kraut, 2002; Fallows, 2005). On average, we found that men tend to send messages and create new acquaintances to a greater extent than women, but women are more successful in attracting other users' attention than men. In fact, women's average in-degree is 32% higher than men's in-degree. In addition to this, because we found heterogenity in system use, we can draw on our data to uncover gender differences for various degrees of system use. To this end,

DOI: 10.1002/asi

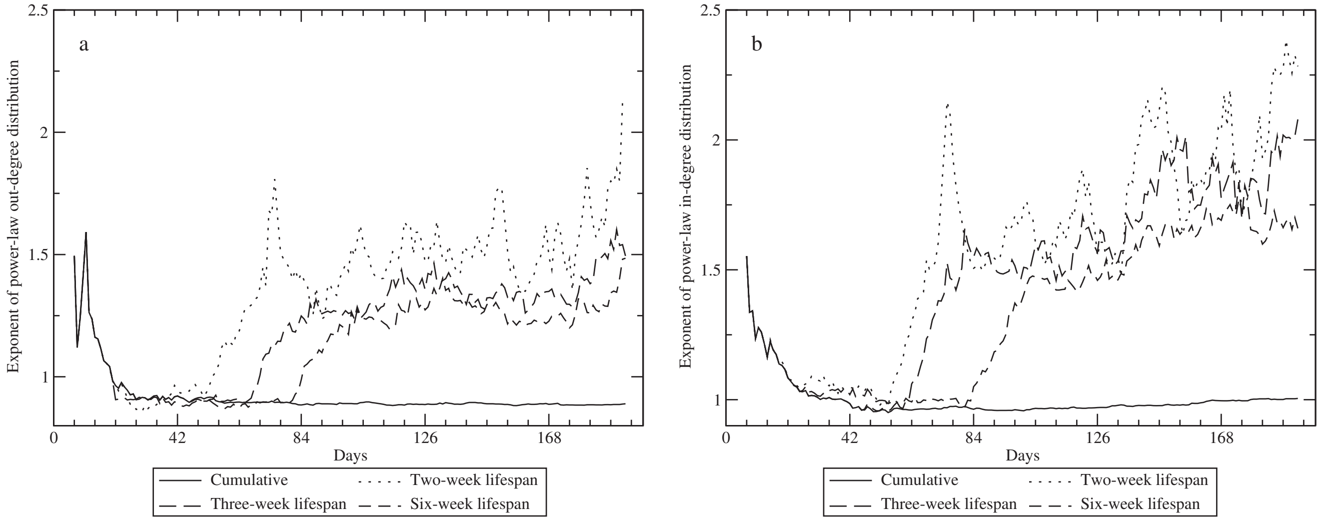

FIG. 4. Evolution of the exponent τ of the power-law out-degree (a) and in-degree (b) distributions.

for each value of out-degree (in-degree), we created subsets of users that have an out-degree (in-degree) of at least that value (Opsahl et al., 2008), and measured the proportion of men in these subsets compared to the overall proportion in the whole system (insets in Figure 3a and b). We found that men outnumber women for the highest values of gregariousness and popularity. For these values, the proportion of male users is well above the baseline proportion found in the whole system. Thus, in qualitative agreement with other studies suggesting that men tend to dominate conversations in a variety of settings, including face-to-face meetings and classrooms (Crawford, 1995; Crawford & Kaufman, 2006), we found that, on average, men are more gregarious and communicate more than women, and that, among all users, men are the ones with the most pronounced levels of system use.

The hub-dominated network structure reflects how the system is organized at the end of our observation period. However, the heterogeneity in users' gregariousness and popularity may have changed over time, and different users might have taken on the role of hubs at different times. To investigate variations in heterogeneity in users' behavior, we analyzed how the out- and in-degree distributions changed over time. To this end, we measured the empirical in- and out-degree distributions each day, both for the case when social relationships never decay and when they have an imposed lifespan. For each day, these distributions are well fitted by power-law functions.[2] Figure 4 shows the evolution of the exponents of these functions. 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 (Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2001). Moreover, when social relationships may dissolve, the out- and in-degree distributions remain fat-tailed, even though their exponents fluctuate due to the relatively small number of active users, especially after the initial period of "accelerated" growth and when relationships decay quickly. The initial growth period also exhibits an increasing correlation between users' out- and in-degree until it reaches its asymptotic value. Users tend to correct their imbalance between gregariousness and popularity. On the one hand, users that are gregarious, but not popular, become more visible and start receiving more messages from new acquaintances. On the other, popular and non-gregarious users tend to reciprocate and thus increase their acquaintances.

Clearly, the choice of the day at which the network is sampled and of the length of the lifespan affects the observed empirical degree distributions. As the lifespan becomes shorter, the distributions become more unstable and dependent on the time at which the observation is taken. Other studies of evolving social networks (Kossinets & Watts, 2006) have made the case that average network measures remain stable over time and therefore measurements taken within a given observation period can be generalized to represent the network over a longer period of time with reasonable fidelity. Our results, however, suggest that, when social relationships may dissolve, the out- and in-degree distributions become vulnerable to measurement assumptions such as the length of the lifespan. In particular, when this length is relatively short, the out- and in-degree distributions show considerable variations depending on the time at which measurements are taken. Therefore, it becomes difficult to infer from network snapshots the shape of the out- and in-degree distributions that can be assumed with reasonable accuracy to reflect the structure of the network over a relatively long period of time.

While the system remains dominated by hubs throughout the whole observation period and irrespective of varying decay thresholds for social relationships, the opportunity for

[2]For each day, these distributions are fitted by power-law functions, and, in each case, the fit has an R^2 ranging from 0.9879 to 0.9998 for the out-degree distribution, and from 0.9617 to 0.9995 for the in-degree distribution, both with p values lower than 0.001.

DOI: 10.1002/asi

a user to become a hub may depend on the time at which the user joins the system. In particular, is it possible for a user to become a hub when the phase of "accelerated" growth is over? Or are there first-mover advantages that enable only those that joined the system in its early days to take on the role of hubs? Understanding who can become a hub and at what stage helps explain whether a newcomer still retains the opportunity to rise to system prominence and how effectively incumbents can compete with one another to attract attention or disseminate information across the system.

To investigate first-mover advantages, we measured the Pearson's correlation coefficient between users' out-degrees (in-degrees) at the end of the observation period and the number of days spent online since they joined the system. We obtained a value of 0.18 (0.24 for in-degree, both with $p < 0.001$), which suggests that, on average, early entrants are likely to become more gregarious (popular) than latecomers. On average, a user's degree is thus an increasing function of the days the user spent in the system. Despite this first-mover advantage, however, different users managed to rise to system prominence at different points in time. For each week, we selected the subsets including the n most gregarious (popular) users for various values of n. We found that the number of unique users in these subsets across the whole observation period is never equal to n, for all values of n. This implies that, although only a select subset of users took on the role of hub, this role was played by different users over time. Moreover, as n gets larger, the number of users that were included in the top n increasingly diverges from n. Thus, rising to system prominence becomes more difficult and selective as the subset of the users classified as hubs becomes smaller. For example, 264 (269) users managed to be among the top 100 most gregarious (popular) ones at some stage across all weeks, while only 26 (32) managed to be among the top 10. Interestingly, while no user that joined the system after the sixth week ever managed to rank higher than the 19th (18th) most gregarious (popular) user, those who ranked among the top 18 (17) joined the system at various stages during that 6-week period. Thus, although there is clearly some advantage associated with early entrance, newcomers still retain the potential to overtake incumbents in becoming hubs. Moreover, users that joined the system in the same week competed to rise to system prominence and became hubs at different points in time.

Reachability and Robustness: The Existence and Emergence of a Giant Component

One fundamental property that contributes to the success of information diffusion in communication networks such as our online community is the possibility for nodes to reach as many other nodes as possible along some path (Kiesler & Cummings, 2002; Boyer O'Leary & Cummings, 2007; Wasserman & Faust, 1994). Clearly, if the system were fragmented into multiple disconnected groups of users, information could circulate only locally within, but not between, individual groups. Barriers against information diffusion, however, disappear when users within a group begin to communicate and establish ties with other users in a different group. This will lead ultimately to the creation within the system of a very large connected subset of users each of whom can reach the others along some path. In these circumstances, most users would be able to communicate with any other user either directly or through other users that would act as intermediaries along a chain of ties. In the modern terminology of networks, when a network is connected in such a way that a large proportion of nodes are reachable from each other, the network is said to contain a *giant component* (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004). In this case, the nodes not included in the giant component belong to other isolated components of a small size that is independent of the total number of nodes in the network.

As a first step toward understanding how freely and widely information flows in the system, we then investigated the existence of a giant component and measured its size at the end of the observation period. The system shows a clear large component, encompassing 1, 893 users out of a total of 1, 899. The second-largest component is made up of just two users. This is clearly evidence of a connected system in which information can travel between almost any two users along some path, in one direction or in the other. In other words, the giant component is *weakly connected* in that, for any pair of nodes i and j in it, either i can reach j or j can reach i. Moreover, with such pronounced differences between largest and second-largest components, it is clear that, by the end of the observation period, the system is well connected and shows no risk of fragmentation into small isolated components that could reduce its communication capabilities[3] (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004).

Our analysis of reachability among users can now be refined to take into account the dynamic nature of the online community. The existence of a giant component at the end of the observation period does not ensure that information could circulate extensively over the entire duration of the system. In fact, the system might have evolved as a collection of disconnected groups of users until, toward the end of the observation, some user in one group created crucial new acquaintances in other groups, thereby bringing all these groups together. Clearly, the point in time in which a giant component emerges is not a trivial subject of analysis, and has

[3]While ensuring reachability among most users along some path, the existence of a giant weakly connected component therefore does not imply that information can flow between almost any two users *in both directions*. To examine whether a substantial fraction of users could contact and be contacted by others, we also measured the size of the giant *strongly connected* component, defined as the largest fraction of nodes in a network that are mutually reachable by a directed path (Dorogovtsev & Mendes, 2003). That is, for any pair of nodes i and j in the giant strongly connected component, i can reach j and j can reach i, either directly or through indirected ties. In the community, the strongly connected component includes 1, 266 users, whereas the second-largest strongly connected component is again made of only two users.

DOI: 10.1002/asi

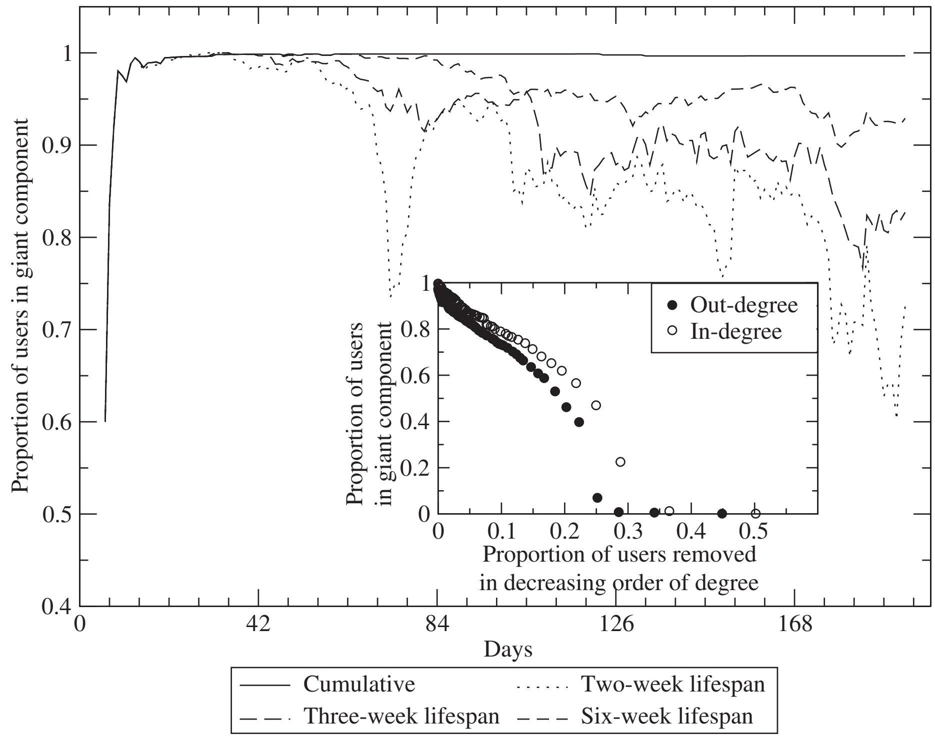

FIG. 5. Evolution of the size of the giant component. Inset refers to sensitivity analysis.

profound implications for the feasibility of communication over time.

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 users in it and the total number of users in the system (Guimerà et al., 2005; Newman, 2001; Newman, Strogatz, & Watts, 2001). As shown in Figure 5, a giant component is generated at a very early stage: at day 9, the majority of users (98%) were reachable along some path. The "accelerated" growth experienced by the system is certainly responsible for the early emergence of a giant component. After joining the system, users keep communicating and creating new acquaintances, which brings about a rapid increase in the density of ties and in reachability among users in the very early days of the observation period. This prompts the system to undergo an abrupt change in its global structural properties with the onset of a connected giant component (Bollobás, 1985; Erdös & Rényi, 1960). Information can thus flow freely over most of the observation period, since almost the inception of the system.

We also examined whether the system remained connected when social relationships have a lifespan. We found that, after the initial growth, the relative size of the giant weakly connected component varies depending on the length of the lifespan. In fact, while for the cumulative network the relative size of the giant component remains almost stationary, with lifespans it decreases over time. In particular, the shorter the length of the lifespan, the smaller the relative size of the giant component, and the more vulnerable it is to the time at which measurements are taken. Despite this, however, the relative size of the giant component never goes below 60.22% of the whole network across all lifespans, whereas the size of the second-largest component never exceeds 7.18% of the network. Thus, while having an impact on its relative size, imposing different lifespans to social relationships does not seem to fragment the system into multiple disconnected components. Even if relationships were to dissolve after a certain period of time, the system would still remain well connected and communication would not suffer. Information could flow among a fraction of connected users that would still be fairly large, albeit smaller than what was found under the assumption of everlasting social relationships.

The notions of reachability and giant component can also help answer a fundamental question that has been the focus of considerable recent research in network science: How robust is a network to removal of nodes? Alternatively, in terms of a communication network, one may ask: Would the exit of some users hinder communication within the network? The Internet, for example, is a highly robust network in that there are multiple paths between nodes (the routers) along which information can flow (Pastor-Satorras & Vespignani, 2004). Even if a router failed, the system would still be able to divert information flow by using the alternative paths without substantially obstructing communication. This kind of robustness to failure crucially depends on how a network is connected, and can be investigated by looking at the largest

DOI: 10.1002/asi

fraction of nodes that would remain reachable after the failure (Albert et al., 2000). If, in the surviving network, the majority of the remaining nodes were still reachable along some path and thus belonged to a giant component, then the network would show resilience to failure.

This argument can be further extended by asking the following questions: What is the maximum level of damage a network can suffer without compromising its connectivity? What is the maximum fraction of nodes that could be removed from a network without having the effect of disintegrating the network into multiple disconnected components? Or, alternatively, is it possible to identify the nodes that, if removed, would destroy the giant component of a network? Clearly, these nodes play a pivotal role in holding a network together (Albert et al., 2000; Pastor-Satorras & Vespignani, 2004). Being able to identify such nodes would then have profound implications for the understanding of how a network is organized and functions.

In the context of our online community, a natural way of addressing these problems would be to cast them in terms of users' communication behavior. In particular, we were interested in examining whether the highly gregarious and popular users played any critical role in preserving communication within the system. To this end, we performed a sensitivity analysis and measured the relative size of the giant component that would survive after the removal of varying fractions of users. Since we noted that the main structural feature of the system resides in the co-existence of users that are highly heterogeneous in terms of their gregariousness and popularity, it is interesting to investigate whether the system is held together precisely by the most gregarious and popular users. To this end, we ranked all users in decreasing order of gregariousness and popularity, and measured the effects on the size of the largest weakly connected component of removing users from the system, starting from the most gregarious and popular ones.

The inset of Figure 5 shows the relative size of the largest component as a percentage of the size of the surviving system when varying proportions of users selected in decreasing order of gregariousness and popularity are removed from the system. It is clear that the system remains connected only when fairly limited fractions of users are removed. The size of the giant component never goes below 70% of the whole system until all users with 25 or more outgoing ties (i.e., 12.37% of all users), or 21 or more incoming ties (i.e., 16.32% of all users), are removed. From that point, and as additional users with lower levels of gregariousness and popularity are removed, the system quickly disintegrates. Being vulnerable to the removal of a relatively small subset of highly gregarious and popular users, the system shows indeed one of the defining features of "scale-free" networks: a minority of well-connected nodes are responsible for creating a giant component and holding the network together (Albert et al., 2000). Gregarious and popular users thus hold the crucial role of enabling information to flow within the system, in that they create paths along which others can communicate. If the gregarious and popular users were to exit the system or if communication with them became unfeasible or even slowed down, the system would suffer from a serious communication breakdown.

Average Distances and the "Small-World Effect"

In addition to reachability among users, another fundamental property that contributes to the successful functioning of a communication network is the speed at which information can flow from one part of the network to another (Rogers, 2003; Valente, 1995). Assuming that information travels at the same speed along all paths of the same length, one could reasonably argue that the time it takes for a piece of information to be transmitted from one user to another is proportional to the length of the shortest path that connects the two users (Wasserman & Faust, 1994). A general question one may ask is: How many intermediaries are needed on average for one user to reach another one? Clearly, the fewer the intermediaries, the shorter the distance between the two users and the more quickly information can travel between them (Lazer & Friedman, 2007; Valente, 1995). While reachability and the existence of a giant component ensure that information can flow widely within the system, short distances between reachable users ensure that information can flow in a timely and accurate fashion.

This idea of distance between nodes in a network has inspired one of the fundamental problems that has driven network research for the last few decades, the concept known today as the "small-world effect" or "six degrees of separation" (Milgram, 1967; Watts & Strogatz, 1998). In this section, we will focus on distances between users and measure them against an appropriate benchmark. In so doing, we will investigate the degree to which the system can be regarded as a tightly woven "small world" linking users seemingly far removed from one another in social space.

We define the geodesic distance between a pair of users as the length of the shortest path between them. For example, if the minimum number of intermediaries required to link users i and j were two, the geodesic distance between i and j would be three. The average geodesic distance g is the mean shortest path length between all pairs of users that are reachable through a connecting path (Wasserman & Faust, 1994, p. 110). For the cumulative network at the end of the observation period, we found: $g = 3.055$. Thus, on average it takes only two intermediaries for two randomly selected users to pass information to each other.

Figure 6a shows the evolution of the average geodesic distance. For the cumulative network, when social relationships are assumed to persist indefinitely, after an initial decline, the geodesic distance reaches its asymptotic level at a time in which users and the average number of acquaintances are still rapidly increasing. This indicates that the system obtains a stationary topology before the rate of growth in its size slows down. We find it interesting to note that the observed reduction in geodesic distance 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

DOI: 10.1002/asi

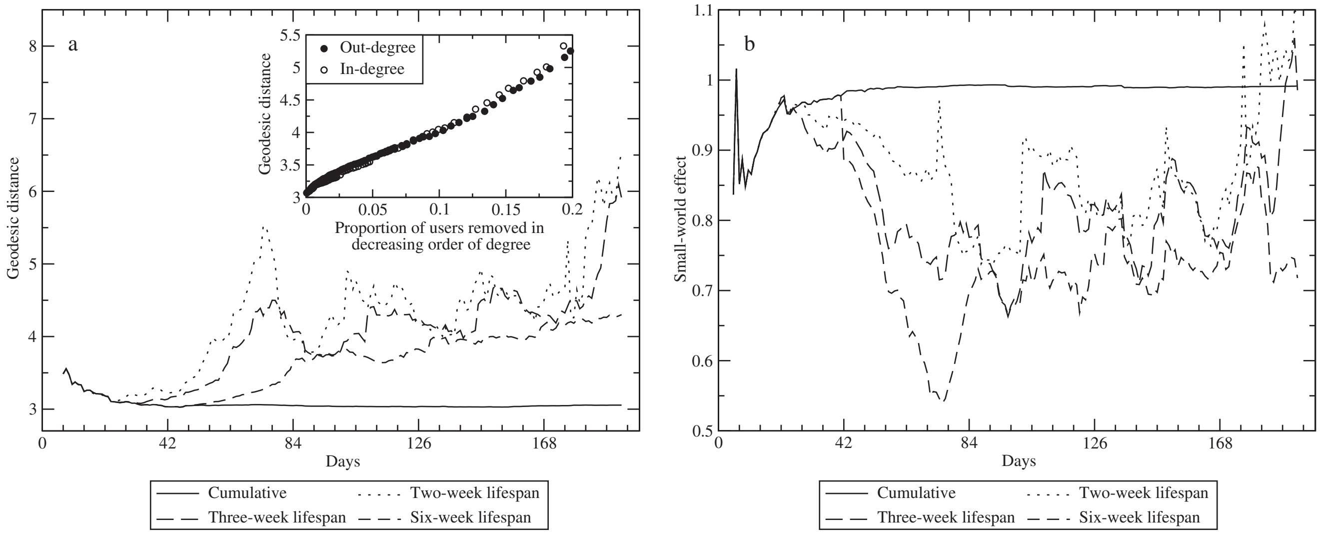

FIG. 6. Evolution of: (a) geodesic distance; and (b) the "small-world effect". Inset refers to the effect of hubs on geodesic.

should increase with network size. However, because, in our case, the system follows a non-linear process of "accelerated" growth, the structure of ties becomes denser over time, and thus average separation between users decreases (Ahn et al., 2007; Barabási et al., 2002; Holme et al., 2004; Leskovec, Kleinberg, & Faloutsos, 2005). As newcomers join the system and create new ties, and incumbents continue creating new ties with one another, the number of ties grows at a higher rate than the number of users. This increasing densification creates new bridges connecting disparate parts of the system, thus bringing distances between users down.

Results also indicate a negative correlation between distances and system size when social relationships have a lifespan. In this case, while the average geodesic distance shows an increasing trend, the number of users declines over time (Figure 1a). This can be explained by the fact that, when relationships dissolve, the decay of ties drives the average number of acquaintances down (Figure 1b), which negatively affects the compactness of the system.

Despite suggesting in some sense that on average the number of intermediaries required to link two randomly chosen users is not that large, at least in absolute terms, these findings however cannot tell us much as to whether the observed distances are short enough to make the system a "small world." Clearly, the same distances we found here would have a different meaning if they were observed in another system with many more users, or in a system with the same number of users as our system but in which ties between users were fewer or established in a different way. To properly evaluate whether the distances we observed are a genuine signature of the "small-world effect," we thus need to assess them against an appropriate benchmark.

In recent years, Watts and Strogatz (1998) defined a social network as being "small" if typical distances between nodes (not counting the nodes that are not connected at all) are comparable with those on a corresponding network in which ties are established at random. In fact, it can be proved rigorously that a network in which ties are randomly created exhibits distances that are much smaller than its size (i.e., the number of nodes; Bollobás, 1985). The "small-world effect" refers precisely to the situation in which a network shows distances that can be approximated by those on a corresponding random network. Formally, a network is "small" when, given a random network with the same number of nodes and undirected ties as the actual network, the ratio

$$\gamma = \frac{\text{measured geodesic distance}}{\text{geodesic on the corresponding random network}} \tag{3}$$

equals one.

To investigate the "small-world effect" in our system, we simulated 100 random networks with the same number of nodes (1,893) and undirected ties (13,835) as in the giant component of the system. We then measured g_{rand}, the average geodesic distance from these simulated networks, and compared g_{rand} with the observed distance in the system by using Equation 3. We found: $g_{rand} = 3.0845$. Thus, for the cumulative network at the end of the observation period, we obtain: $\gamma = 0.99$. Moreover, as shown in Figure 6b, we found convergence over time between the measured average geodesic distance and the average distance found in corresponding random networks (Davis, Yoo, & Baker, 2003; Uzzi & Spiro, 2005). However, this convergence does not remain stable over time and appears to be vulnerable to the length of the lifespan imposed to social relationships. Despite this, and even though with lifespans the compactness of the system declines, average distances remain very close to, and most of the time even shorter than, those on corresponding random networks.

The results so far are consistent in pointing to the "small-world effect" in the system. However, there is one fundamental problem in using the simulated random networks, from which g_{rand} was measured, as benchmark. In fact, these networks do not show one crucial feature that we found in the

DOI: 10.1002/asi

system. They implicitly assume that the randomly generated acquaintances are homogeneously distributed across the users. By contrast, we found that users are highly heterogeneous in terms of their gregariousness and popularity: the system is indeed dominated by few very gregarious and popular users. To obtain an appropriate benchmark that takes this structural feature into account, we then simulated 100 networks with the same number of nodes and the same heterogeneity in the distribution of ties across nodes as in the giant component of the system. We then measured g'_{rand}, the average geodesic distance from these simulated networks. We obtained: $g'_{rand} = 2.9004$.

The observed geodesic distance g thus mirrors fairly closely both g_{rand} and g'_{rand}. Altogether, these findings indicate that the system is indeed compact and can be regarded as a "small world" (Ebel et al., 2002; Davis et al., 2003; Newman, 2001; Uzzi & Spiro, 2005; Watts & Strogatz, 1998). In addition to this, our findings also help shed light on two organizing principles that govern the way users in the system interact with one another. First, the fact that the observed average geodesic distance g is smaller than g_{rand} provides further evidence in support of the role played by the highly gregarious and popular users in the system (Ebel et al., 2002; Newman et al., 2001). These are in fact the hubs that, through their disproportionately large amount of ties, help reduce the distances between otherwise disconnected users. This explains why the observed distance g is smaller than the distance g_{rand} measured on simulated random networks with no such hubs as in the system.

To further investigate this, we conducted a sensitivity analysis and measured the geodesics of the networks obtained by removing nodes in decreasing order of their out- and in-degrees. As shown in the inset of Figure 6a, geodesic increases steadily as an increasing proportion of users of decreasing levels of gregariousness and popularity are removed from the system. For example, when all users with out-degree greater than 23 are removed (i.e., 13.79% of all users), geodesic rises up to 5.15, yielding a more pronounced divergence from the geodesic we would obtain on a corresponding random network ($\gamma = 1.23$). It is thus a relatively small proportion of users, namely the most connected ones, that are responsible for bringing down distances among other users and facilitating information flow within the system. Further support to the role of hubs in reducing network distances is provided by the observation that heterogeneity in users' gregariousness and popularity, measured in terms of entropy and skewness of the out- and in-degree distributions, shows the highest rates of increase precisely in the growth period in which geodesic shows a significant decline. While opportunities to further rise to system prominence are discriminatively offered only to a select subset of users, they also serve a collective purpose that benefits the whole system as each user can reach all others at shorter distances.

Our findings also give support in favor of a second organizing principle underpinning the system. When we took hubs into account in the simulation of random networks, we obtained a value of g'_{rand} for these networks that is lower than the observed average distance g. Clearly, this evidence suggests that fewer ties are used in the system to link otherwise disconnected clusters of users than would be the case if the ties, still monopolized by a small number of hubs, were established at random. A possible explanation is that, in the system, ties are also used locally to create tightly knit groups of users (Ebel et al., 2002). While contributing toward a reduction in global connectivity (Watts & Strogatz, 1998), these cohesive groups are rich in locally redundant ties that have traditionally been advocated as a source of social capital (Coleman, 1988). Local redundancy is indeed the signature of a special organizing principle that we will examine in the following section.

Local Redundancy and Clustering

As recent studies have pointed out, while reproducing the short distances and the "small-world effect" typically found on random networks, social networks differ from random networks in one fundamental way. They appear to display an ordered structure organized into well-defined tightly knit groups of connected individuals (Watts & Strogatz, 1998). People are embedded in densely connected local groups based on shared experience, interests, or location, which, in turn, are connected to each other by the shortcuts created when individuals in one group also belong to other groups (Jin, Girvan, & Newman, 2001; Newman & Park, 2003). Building on Coleman's (1988) conception of social capital, a number of sociologists have emphasized the benefits that individuals can gain from forging locally redundant social ties and creating socially cohesive groups (Ingram & Roberts, 2000; Levine & Kurzban, 2006; Reagans & McEvily, 2003; Uzzi, 1997; Uzzi & Spiro, 2005). From this perspective, it is often argued that local redundancy and social cohesion promote a sense of belonging, foster trust, facilitate the enforcement of social norms and the transfer of tacit and complex information, and enable the creation of a common culture. However, as pointed out in the previous section, an increase in local redundancy often comes at the expense of global connectivity (Watts & Strogatz, 1998). In particular, the more an individual's new contacts are already connected to the individual's current contacts, the less use the new contacts are in taking the individual closer to new people he or she cannot reach already. Moreover, local redundancy is likely to hinder access to novel information (Burt, 1992; Granovetter, 1973). If all your friends' friends are also your friends, none of your friends can introduce you to new social circles where you may meet someone you have never met before, and who knows something you do not know already.

More generally, the degree of local redundancy in a social network can be better investigated by asking the following question: If there are three randomly chosen 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 nodes is often referred to as transitivity (Davis, 1970; Holland & Leinhardt, 1970; Levine & Kurzban, 2006; Rapaport, 1953; Wasserman & Faust, 1994). Transitivity is the probability of

DOI: 10.1002/asi

j being acquainted with l. When the direction of ties between nodes is not known, transitivity can be measured with the following clustering coefficient C (Newman et al., 2001):

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}} \qquad (4)$$

where a triangle refers to three nodes that are all connected to each other, whereas a connected triple refers to three nodes connected through at least two ties.

The quantity C defined in Equation 4, also known in the sociology literature as "fraction of transitive triples" (Wasserman & Faust, 1994), takes values that lie in the range from zero to one. In fact, in a network where everybody is connected to everybody else, all connected triples are also triangles, and C is therefore equal to one. By contrast, in a network where ties are established at random, the probability of a tie between any two nodes remains the same and does not depend on whether or not the two nodes share a common neighbor. Hence, in such a network C is equal to the probability of a tie, which becomes negligible as the total number of nodes gets very large, given a constant average degree (Albert & Barabási, 2002; Caldarelli, 2007; Dorogovtsev & Mendes, 2003; Newman, 2003; Pastor-Satorras & Vespignani, 2004). Once again, the value of C on a random network provides the benchmark against which one can evaluate the clustering observed in a real-world network (Watts & Strogatz, 1998). In particular, a value significantly higher than the one found on a corresponding random network is the signature of a special organizing principle. It suggests a departure from randomness in the way nodes generate ties and a shift toward a structure in which the probability of a tie between two nodes depends on the number of common neighbors these nodes share.

Empirical evidence has shown that, in most real-world networks, clustering takes a non-zero value, greater than expected by chance (Barabási et al., 2002; Davis et al., 2003; Ebel et al., 2002; Holme et al., 2004; Newman, 2001; Uzzi & Spiro, 2005; Watts & 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 now be tested with our data. For the cumulative network at the end of the observation period, we obtained: $C = 0.0568$. To assess if this value is indeed evidence of non-randomness in the way ties are generated in the system, we simulated 100 networks in which ties are randomly created and with the same number of users and same average number of acquaintances as in our system. For each of these simulated networks we then measured the clustering, and obtained $C_{rand} = 0.0077$, the average clustering from these simulations.[4] However, this value does not take into account the heterogeneity in gregariousness and popularity observed in the system. To take this into account, as with the geodesic distance, we simulated 100 random networks with the same number of users and the same distribution of acquaintances across users as in our system (Newman et al., 2001; Watts & Strogatz, 1998). Once again, we measured the clustering on each of these simulated networks, and obtained an average value of $C'_{rand} = 0.0139$. Because the observed clustering in the system is larger than the values found in all simulated random networks, these results suggest that the way in which ties are established within the system differs from what would be expected with random connections. Users show a tendency to create densely connected local groups by preferentially sending messages to other users that are the acquaintances of their own acquaintances.

We also wanted to investigate how local redundancy varies across users. To associate each user with a measure of local redundancy, we defined for each user the ratio between the total number of triangles the user is part of (Δ), and the expected number of triangles if ties among the user's nearest neighbors were randomly established (Δ_{rand}).[5] When this ratio takes a value higher than one, a user's nearest neighbors are more connected with one another than randomly expected. This would enhance redundancy within the user's local environment, and create higher degrees of social cohesion than would be the case if the ratio were less than one. Using this measure of local redundancy, we found that in analogy with gregariousness and popularity, redundancy is also heterogeneously allocated across users. As shown in Figure 7a, the distribution of local redundancy is skewed, with a Pearson's skewness coefficient of 0.9: while almost every user is successful in building a densely connected local environment, only very few users are associated with very large values of redundancy, whereas the majority exhibit low to moderate values.

We also used our measure of users' local redundancy to examine gender differences in terms of social cohesion (Moore, 1990). We found that on average men exhibit higher levels of local redundancy than women. However, given that redundancy is heterogeneously distributed within the system, this result may mask important differences in the way various levels of redundancy are allocated between men and women. The inset of Figure 7a plots the fraction of male users, with respect to the proportion found in the overall community, for increasing levels of local redundancy. Women tend to rise above the baseline only for moderate levels of redundancy, whereas the majority of users embedded in many more

[4]Note that the value of clustering obtained from the simulated random networks is equal to the density of these networks (i.e., the probability of a tie).

[5]More formally, the number of triangles in which a user is embedded is equal to the number of ties connecting the user's nearest neighbors. The randomly expected number of such triangles is equal to the maximum possible number of ties among the user's nearest neighbors multiplied by the probability of a tie in a corresponding random network with the same number of nodes and ties as in the real one. In turn, this probability is equal to network density, i.e., the ratio between number of ties and maximum possible number of ties (Bollobás, 1985; Wasserman & Faust, 1994). We also checked for robustness and performed the analysis by using the local clustering coefficient that, unlike our measure, is strongly negatively correlated with node degree (Watts & Strogatz, 1998). We obtained consistent results.

DOI: 10.1002/asi

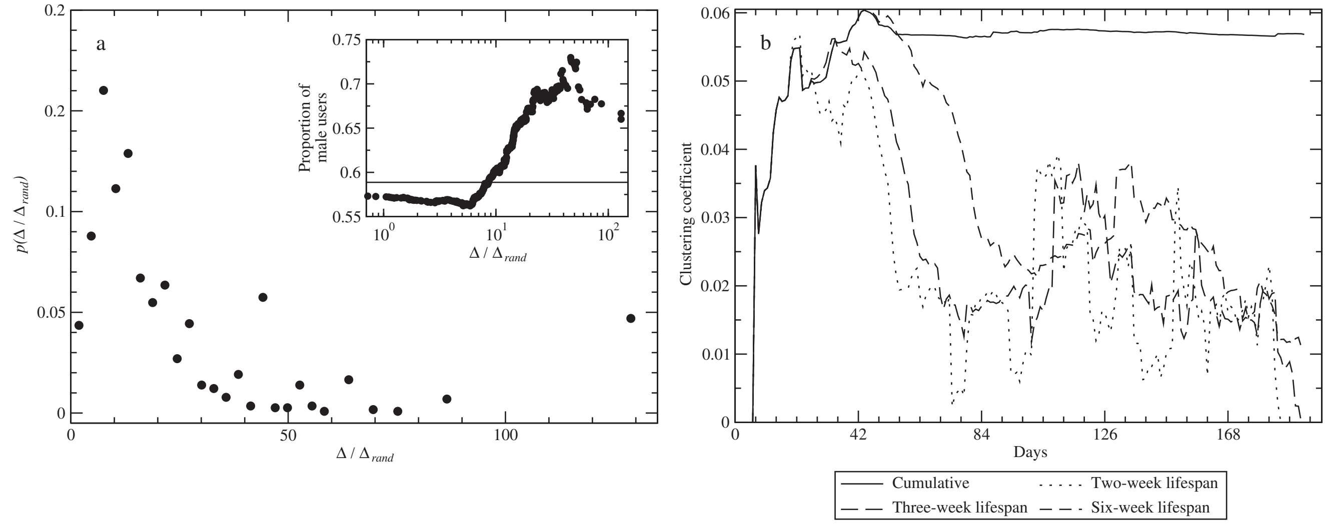

FIG. 7. (a) Distribution of local redundancy among users; (b) Evolution of clustering. Inset refers to the proportion of male users in correspondence to various levels of local redundancy.

triangles than expected by chance are men. Thus, men not only have average larger values of local redundancy than women, but they also preside over the largest values found in the system. While these findings only refer to how embedded users are in their local environments, and not to the social content of relationships (e.g., chatting, social support), they are in qualitative disagreement with what was found by other scholars that reported a tendency of women to invest in social relationships more than men, both offline (Moore, 1990) and online (Boneva & Kraut, 2002; Cummings & Kraut, 2002). Despite this, however, our results so far are theoretically consistent. On the one hand, we found that men not only create more relationships, but also invest in their contacts and develop strong ties to a greater extent than women. On the other, Granovetter (1973) argued that strong ties are more likely to occur within triangles than weak ties. It is therefore not surprising to find that men tend to be more socially embedded than women.

Figure 7b shows the evolution of the clustering coefficient when social relationships never dissolve and when they have a lifespan. In the former case, clustering 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 clustering observed that day. In particular, the deviation between observed clustering and that on corresponding random networks is most pronounced during the initial phase of "accelerated" growth, while it attenuates and becomes more stable in the remaining period. While results are vulnerable to the length of the lifespan and the time at which measures are taken, the clustering coefficient of the cumulative network in which relationships never decay exhibits an increasing trend that is at variance with other empirical findings (Holme et al., 2004; Kossinets & Watts, 2006) and with the behavior predicted by some recently developed network models (Barabási & Albert, 1999; Barabási et al., 2002). Effects arising from social and spatial proximity among users can help explain the findings (Feld, 1981; Monge, 1985). Users that belong to the same circles 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 (Newman, 2001). Users may communicate offline with one another in pairs, and then replicate their interactions online (Boase et al., 2006; Boase, 2008; Stern, 2008; Wellman, 1999; Wellman et al., 2006). Affiliation to the same offline social circles, 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 system.

Discussion

In this article, we relied on recent theoretical and methodological advances in network science for studying users' online behavior as well as patterns and dynamics of social interaction in an electronic environment. We used time-stamped online messages to investigate the general regularities governing the initiation and progression of interpersonal communication. The findings have implications for theorizing on evolving social networks and for research on patterns and dynamics of system use and the role of information technology in facilitating social interaction. While clarifying and extending past research on social networks by casting light on critical issues that tend to be overlooked in network modeling efforts, the results also uncover new avenues for theoretical advancement in information systems research. These include: the diversity of growth patterns followed by the system, and the processes of reciprocation and deepening of social relationships that underpin the system's evolution; the role that the direction of ties plays in shaping users' behavior and the structure and evolution of the system; the duality of time

DOI: 10.1002/asi

as seen from the system's and the users' perspectives; and the evolution of the system's structural properties and their vulnerability to measurement assumptions, such as whether or not social relationships among users dissolve, and if they do, the length of their lifespan.

In close analogy with many other social (Davis et al., 2003; Holme et al., 2004; Newman, 2001; Uzzi & Spiro, 2005) and non-social networks (Albert et al., 1999; Faloutsos et al., 1999; Watts & Strogatz, 1998), we found evidence of "small-world" properties, in that the system shows (a) average geodesic distances between users that are comparable to those found in a corresponding random network, and (b) a level of clustering greater than would be expected by chance. In this sense, the system belongs to the class of "small-world networks" (Watts & Strogatz, 1998). In particular, the system exhibits small average geodesic distances throughout the whole observation period, during which they decrease as the size of the system grows (Ahn et al., 2007; Barabási et al., 2002; Holme et al., 2004; Leskovec et al., 2005). Results also indicate that distances remain small regardless of the length of the lifespan imposed on social relationships (Kossinets & Watts, 2006). While lending further support to the idea that most real-world networks are indeed "small worlds," these findings contribute toward a deeper understanding of distance and the effects of system use on dispersed teams (Kiesler & Cummings, 2002; Mok et al., 2007; Monge, 1985; Boyer O'Leary & Cummings, 2007). In particular, our proposed measures of the system's geodesic and the "small-world" effect inspire theory building on the association between dispersion, types of communication, and performance. In this respect, they help researchers conceptualize and measure distances among users by focusing on more dimensions than has been the case in previous research on information systems and teamwork. Shifting emphasis from the more traditional notion of geographic distance to that of network distance adds a new dimension to the study of dispersed teams (Mok et al., 2007). A better understanding of the interplay between geographic and network distances has the potential to inspire the development of new theories on the facilitating role of information technology in increasing the sense of connectedness and reachability among otherwise geographically isolated users.

We also found evidence of "scale-free" behavior (Barabási & Albert, 1999). Our analysis of users' gregariousness and popularity indicates that the empirical distributions for the number of incoming and outgoing ties have a fat-tailed form, irrespective of the time at which they are measured. Thus, users remain heterogeneous in the way they communicate throughout the evolution of the system. On the one hand, this result seems to be in qualitative agreement with previous research in the small-group tradition suggesting that the seeds of emerging status hierarchies can be found in the early stages of the life of a group, whereby initial expectations are then reinforced by later interaction leading to a stable hierarchy (Berger, Fişek, Norman, & Zelditch, 1977). On the other, the 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 social relationships or searching for new acquaintances prevents individuals from having a disproportionately large number of relationships with others, the degree distribution 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 (Amaral et al., 2000; Bernard et al., 1988; Fararo & Sunshine, 1964). Our results do not represent the first reported evidence of "scale-free" behavior for social networks (Barabási & Albert, 1999; Barabási et al., 2002). In particular, fat-tailed degree distributions were also found for other online communication networks such as e-mail networks (Ebel et al., 2002; Kossinets & Watts, 2006) and online communities and social networking services (Ahn et al., 2007; 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 may include relationships between users that never meet offline, and very sporadic relationships that do not absorb much of a user'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 (Boase et al., 2006; Boase, 2008; Stern, 2008; Wellman, 1999; Wellman & Haythornthwaite, 2002; Wellman et al., 2006). If this were the case, due to the imposed constraints on communication, connections should be homogeneously distributed among users (Amaral et al., 2000). Our results on the presence of hubs with strong ties to many other users do not support this hypothesis.

Our findings on the system's "scale-free" behavior open up a new direction for research on technology adoption and use. While underscoring the fundamental role of heterogeneity among users, our analysis challenges the plausibility and generalizability of information systems theories based on the idea of a prototypical user's behavior. We found that users differ in not only popularity and gregariousness, but also their tendency to forge strong ties and develop socially cohesive local environments. In addition to this, we also found gender differences intersecting variations of structural properties across users, and further highlighted the fallacy in which researchers may incur when they look at average properties to characterize the overall system. For example, while female users are on average more popular than male ones, it is a select minority of the latter that manage to become the hubs attracting a disproportionate amount of attention from others. While inspiring research on the cognitive, demographic, and dispositional factors that affect users' behavior (McElroy, Hendrickson, Townsend, & DeMarie, 2007), our work thus also integrates and extends previous studies on differences among users and their influence on technology adoption and use (Amiel & Sargent, 2004; Compeau et al., 1999; Taylor, 2004; Zmud, 1979). We also investigated first-mover

DOI: 10.1002/asi

advantages underlying users' ability to become hubs. In this respect, our findings have the potential to inspire future research on users' opportunities to rise to system prominence. More generally, our emphasis on system growth processes and their role in placing time-dependent constraints on communication suggests a new direction of research to scholars interested in the democratic nature of information technologies and the way opportunities and constraints concerning system use may vary between early and late adopters.

Limitations

The empirical analysis we carried out is not without limitations. The major one, shared by other research in this area, comes from the quality of the data. Even though care has been taken to filter out messages from moderators and users that were only testing the system (Kossinets & Watts, 2006), some of the messages included in the analysis might still fail to reflect genuine social interactions. Although we believe these messages are likely to represent a minority in the dataset, they might have affected the observed "scale-free" behavior of the network. The lack of availability of the content of messages has prevented us from further investigating whether all messages included in the analysis reflected genuine relationships. There are many other issues that we could not study due to lack of message content, such as users' experience of being part of the system, the type and quality of the knowledge transferred and shared, and social influence. Our analysis also draws on a weak notion of social relationship, in that we assumed that just one message was sufficient to create a relationship (Constant et al., 1996; Granovetter, 1973). While this enabled us to include all users into the analysis, and not only those with a tendency to reinforce their relationships, it might have brought about an overestimation of the number of actual relationships. This might also have affected some of the system's observed structural properties.

Even though we followed the evolution of the system since inception, we could not control for prior offline relationships among users. In this sense, the dataset is left-censored. The evolution of the network is thus a combination of two effects: (a) interactions among users that have already met offline; and (b) actual new relationships. This conflation implies that one must be careful in interpreting the results. Since we could not control for prior relationships, it is quite likely that our analysis may capture adoption and diffusion processes through (and hence a reflection of) the existing offline social network. Furthermore, as in all empirical studies, the robustness and generalizability of findings largely depend on the size and quality of the dataset used. In our case, a larger and more detailed dataset, extended over a longer period of time, would have certainly improved the quality of our study.

An important issue that we did not investigate and that future work should address is concerned with the growth mechanisms underpinning the evolution of the system (Banks & Carley, 1996; Snijders, 2005). The evidence of a fat-tailed degree distribution makes it plausible to suspect that the evolution of the system is governed by cumulative advantage mechanisms (Barabási & Albert, 1999; Gould, 2002; Jeong, Néda, & Barabási, 2003; Merton, 1968). However, these mechanisms may drive evolution in combination with other social processes arising from relational content and spatial constraints (Powell et al., 2005). Future research should investigate, for example, the mediating effects of homophily on cumulative advantage, and evaluate how the probability of a tie varies as a function of similarity between users. For example, using our data, this can be done by examining the extent to which the probability of two users being connected by a tie is proportional to the number of demographic characteristics shared by the two users, or the degree to which they share the same disposition toward technology adoption and use (McElroy et al., 2007). In addition, since we found an unexpected increasing trend for clustering, it may be conjectured that users sharing a common acquaintance are more likely to develop a social relationship than users with no common acquaintance (Davis, 1970; Holland & Leinhardt, 1970; Rapaport, 1953). This hypothesis of triadic closure can be tested by assessing how the probability of connections between users varies with the number of acquaintances they have in common.

Implications for Practice

In addition to providing a platform for further theorizing on evolving social networks and patterns and dynamics of system use, our study also has implications for practice in such areas as information diffusion, search, and retrieval, as well as the development, deployment, and security of information systems. The observed "small-world" properties have implications for information diffusion, social capital, and communities of practice. The self-organization of the system into a compact structure with small distances between users suggests that information can travel rapidly and reach most parts of the system accurately without requiring additional investments (Davis et al., 2003; Newman, 2001; Uzzi & Spiro, 2005). Although the effects of network compactness on the system's long-term performance still remain to be empirically investigated, especially when exploitation is facilitated at the expense of exploration (Lazer & Friedman, 2007), nonetheless the effects on short-term performance are likely to be beneficial, chiefly in terms of velocity and fidelity of information transmission. At the same time, while some long-range connections serve as shortcuts that shrink distances, other connections are consumed locally to create cohesive social 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 behavior, and make misuse of communication more likely to be detected and punished (Coleman, 1988; Davis et al., 2003; Ingram & Roberts, 2000; Levine & Kurzban, 2006; Louch, 2000; Reagans & McEvily, 2003; Uzzi, 1997; Uzzi & 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

DOI: 10.1002/asi

forces would help expose the structure of topically organized local communities of practice, and facilitate the design of the appropriate methods for charting interests, accessing the available information, and devising more effective marketing campaigns (Guimerà et al., 2006; Kleinberg & Lawrence, 2001).

The findings on the "scale-free" behavior of the system have implications for information diffusion, the security and robustness of information systems, and information searching strategies. If the "scale-free" topological properties we uncovered turn out to be a fundamental feature of communication networks (Ahn et al., 2007; 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" structure of the system makes it vulnerable to the unilateral action of the highly connected users. The propagation of viruses is facilitated, and the potential gain from malfeasance is amplified, by a network structure in which hubs co-exist with poorly connected users (Albert et al., 2000). Along their multiple connections, hubs can spread viruses and inaccurate or even disruptive information, reaching most users within the system. Similarly, since the system is held together by a minority of highly connected users, the effectiveness and pervasiveness of communication are largely dependent on them. Should they decide to leave and sever their connections, the system would break apart and information diffusion would suffer.

On the other, it is precisely by becoming aware of the role played by hubs that moderators and 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 users, the spreading of viruses and disruptive behavior 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 foster interpersonal bonds and common identity among the users of the system (Preece, 2000; Ren, Kraut, & Kiesler, 2007; Rogers, 2003; Valente, 1995). At the same time, the hub-dominated structure makes the system fairly robust against the random removal of users and/or failure of communication channels (Albert et al., 2000). As accidental removals and failures disproportionately affect the poorly connected users, the system will remain resilient and break apart only when a significantly large proportion of users and communication channels have been removed. Even the accidental removal of a single hub will not be fatal for the system. In fact, our findings show that the system has a decentralized structure, self-organized into a number of hubs that share the control and monitoring of connections. This makes the system robust against accidental removal of a single hub as the remaining hubs will still succeed in holding the system together.

Finally, the structural properties of the system can be exploited to devise efficient strategies for searching and retrieving information. By biasing the routing of queries toward the most connected users (Adamic et al., 2001), the system can be navigated and searched with larger efficiency than would be the case if users were asked to contact all their nearest neighbors (broadcast search) or only one randomly selected neighbor (random walk search). Thus, understanding the network hub-dominated structure has profound 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 system.

Acknowledgments

We wish to give very special thanks to Andrew Gelman, Caroline Haythornthwaite, Maxine Robertson, and two anonymous reviewers for their valuable feedback on earlier versions of this article. We are also grateful to workshop participants at the University of Oxford and the I.S.I. Foundation in Turin, and to participants at the European Conference on Complex Systems, the Sunbelt Conference, and the UK Social Network Conference for their helpful comments and suggestions.

References

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

Ahn, Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. Proceedings of the 16th International World Wide Web Conference, 835–844.

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

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

Albert, R., & Barabási, A. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 73, 47–97.

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

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

Alvesson, M., & Sveningsson, S. (2003). Managers doing leadership: The extra-ordinarization of the mundane. Human Relations, 56(12), 1435–1459.

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

Amiel, T., & Sargent, S. (2004). Individual differences in Internet usage motives. Computers in Human Behavior, 20(6), 711–726.

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

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509–512.

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

Baym, N., Zhang, Y., & Lin, M. (2004). Social interactions across media. New Media & Society, 6, 299–318.

Berger, J., Fişek, M., Norman, R., & Zelditch Jr., M. (1977). Status characteristics in social interaction: An expectation states approach. New York, NY: Elsevier.

DOI: 10.1002/asi

Bernard, H., Kilworth, P., Kronenfeld, D., & Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. Annual Review of Anthropology, 13, 495–517.

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

Boase, J. (2008). Personal networks and the personal communication system. Using multiple media to connect. Information, Communication and Society, 11(4), 490–508.

Boase, J., Horrigan, J., Wellman, B., & Rainie, L. (2006). The strength of Internet ties. Washington, DC: Pew Internet & American Life Project.

Bollobás, B. (1985). Random graphs. New York, NY: American Press.

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

Boneva, B., & Kraut, R. (2002). Email, gender, and personal relationships. In B. Wellman & C. Haythornthwaite (Eds.), The Internet in Everyday Life, 372–403. Malden, MA: Blackwell Publishing.

Boyer O'Leary, M., & Cummings, J. (2007). The spatial, temporal, and configurational characteristics of geographic dispersion in teams. Management of Information Systems Quarterly, 31(3), 433–452.

Burt, R. (1992). Structural holes: The social structure of competition. Cambridge, MA: Harvard University Press.

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

Burton-Jones, A., & Gallivan, M. (2007). Toward a deeper understanding of system usage in organizations: A multilevel perspective. Management of Information Systems Quarterly, 31(4), 657–679.

Caldarelli, G. (2007). Scale-free networks. Complex Webs in nature and technology. New York, NY: Oxford University Press.

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

Compeau, D., Higgins, C., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. Management of Information Systems Quarterly, 23(2), 145–158.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. Organization Science, 7(2), 119–135.

Crawford, M. (1995). Talking difference: On gender and language. London: Sage.

Crawford, M., & Kaufman, M. (2006). Sex differences versus social processes in the construction of gender. In K. Dindia & D. Canary (Eds.), Sex Differences and Similarities in Communication, 179–194. New York, NY: Routledge, second edition.

Cummings, J., & Kraut, R. (2002). The quality of online social relationships. Communications of the ACM, 45(7), 103–108.

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

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

Diesner, J., Frantz, T., & Carley, K. (2006). Communication networks from the Enron email corpus. Journal of Computational and Mathematical Organization Theory, 11, 201–228.

Dorogovtsev, S., & Mendes, J. (2003). Evolution of networks. From biological nets to the Internet and WWW. New York, NY: Oxford University Press.

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

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

Fallows, D. (2005). How women and men use the Internet. Washington, DC: Pew Internet & American Life Project.

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

Fararo, T., & Sunshine, M. (1964). A study of a biased friendship network. Syracuse, NY: Syracuse University Press.

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

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

Gould, R. (2002). The origins of status hierarchies: A formal theory and empirical test. American Journal of Sociology, 107(5), 1143–1178.

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

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

Guetzkow, H., & Simon, H. (1954). The impact of certain communication nets upon organization and performance in task-oriented groups. Management Science, 1, 233–250.

Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., & Arenas, A. (2006). The real communication network behind the formal chart: Community structure in organizations. Journal of Economic Behavior and Organization, 61, 653–667.

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

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

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

Howard, P., Rainie, L., & Jones, S. (2002). Days and nights on the Internet. In B. Wellman & C. Haythornthwaite (Eds.), The Internet in Everyday Life, 45–73. Malden, MA: Blackwell Publishing.

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

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

Jin, E., Girvan, M., & Newman, M. (2001). The structure of growing social networks. Physical Review E, 64, 046132.

Kiesler, S., & Cummings, J. (2002). What do we know about proximity and distance in work groups? A legacy of research. In P. Hinds & S. Kiesler (Eds.), Distributed Work, 57–82. Cambridge, MA: MIT Press.

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

Kochen, M. (1989). The small world. Norwood, NJ: Ablex.

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

Krapivsky, P., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. Physical Review Letters, 85, 4629.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. Journal of Social Issues, 58(1), 49–74.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. Administrative Science Quarterly, 52, 667–694.

Lee, Y., Kozar, K., & Larsen, K. (2003). The technology acceptance model: Past, present, and future. Communication of the Association for Information Systems, 12(50), 752–780.

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

Levine, J., & Moreland, R. (1990). Progress in small group research. Annual Review of Psychology, 41, 585–634.

Levine, S., & Kurzban, R. (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. (1990). Network data and measurement. Annual Review of Sociology, 16, 435–463.

McElroy, J., Hendrickson, A., Townsend, A., & DeMarie, S. (2007). Dispositional factors in Internet use: Personality versus cognitive style. Management of Information Systems Quarterly, 31(4), 809–820.

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

DOI: 10.1002/asi

Merton, R. (1968). The Matthew effect in science. Science, 159, 56–63.
Milgram, S. (1967). The small world problem. Psychology Today, 2, 60–67.
Mok, D., Wellman, B., & Basu, R. (2007). Did distance matter before the Internet? Interpersonal contact and support in the 1070s. Social Networks, 29, 430–461.
Monge, P., Rothman, L., Eisenberg, E., Miller, K., & Kirste, K. (1985). The dynamics of organizational proximity. Management Science, 31, 1129–1141.
Moore, G. (1990). Structural determinants of men's and women's social networks. American Sociological Review, 55, 726–735.
Newman, M. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98, 404–409.
Newman, M. (2003). The structure and function of complex networks. SIAM Review, 45, 167–256.
Newman, M., & Park, J. (2003). Why social networks are different from other types of networks. Physical Review E, 68, 036122.
Newman, M., Strogatz, S.H., & Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications. Physical Review E, 64, 026118.
Onnela, J.-P., Saramki, J., Hyvnen, J., Szab, G., Lazer, D., Kaski, K., et al., (2007). Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 104, 7332–7336.
Opsahl, T., Colizza, V., Panzarasa, P., & Ramasco, J.J. (2008). Prominence and control: The weighted rich-club effect. Physical Review Letters, 101, 168702.
Pastor-Satorras, R., & Vespignani, A. (2004). Evolution and structure of the Internet. New York, NY: Cambridge University Press.
Plickert, G., Côté, R.R., & Wellman, B. (2007). It's not who you know, it's how you now them: Who exchanges what with whom? Social Networks, 29, 405–429.
Powell, W., White, D., Koput, K., & Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology, 110(4), 1132–1205.
Preece, J., (2000). Online communities: Designing usability and supporting sociability. Chichester, U.K.: John Wiley & Sons.
Rapaport, A. (1953). Spread of information through a population with socio-structural bias. I. Assumption of transitivity. Bulletin of Mathematical Biophysics, 15, 523–533.
Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. Administrative Science Quarterly, 48, 240–267.
Ren, Y., Kraut, R., & Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. Organization Studies, 28(3), 377–408.
Rogers, E. (2003). Diffusion of innovations. New York, NY: Free Press.
Rothaermel, F., & Sugiyama, S. (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.
Smith-Doerr, L., & Powell, W. (2005). Networks and economic life. In N. Smelser & R. Swedberg (Eds.), The Handbook of Economic Sociology, 379–402. Princeton, NJ: Princeton University Press.
Snijders, T.A.B. (2005). Models for longitudinal network data. In P. Carrington, J. Scott, & S. Wasserman (Eds.), Models and Methods in Social Network Analysis, 215–247. New York: Cambridge University Press.
Stern, M. (2008). How locality, frequency of communication and internet usage affect modes of communication within core social networks. Information, Communication and Society, 11(5), 591–616.
Szajina, B. (1993). Determining information systems usage: Some issues and examples. Information & Management, 25, 147–154.
Taylor, W. (2004). Computer-mediated knowledge sharing and individual user differences: An exploratory study. European Journal of Information Systems, 13(1), 52–64.
Tuckman, B.W. (1965). Developmental sequence in small groups. Psychological Bulletin, 63, 384–399.
Tuckman, B.W., & Jensen, M.A. (1977). Stages of small-group development revisited. Group Organization Studies, 2, 419–427.
Tyre, M., & Orlikowski, W. (1994). Windows of opportunity: Temporal patterns of technological adaptation in organizations. Organization Science, 5(1), 98–118.
Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. Administrative Science Quarterly, 42, 35–67.
Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. American Journal of Sociology, 111, 447–504.
Valente, T. (1995). Network models of the diffusion of innovations. Cresskill, NJ: Hampton Press.
Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. Management of Information Systems Quarterly, 27(3), 425–478.
Wang, B., Tang, H., Guo, C., & Xiu, Z. (2006). Entropy optimization of scale-free networks' robustness to random failure. Physica A, 363(2), 591–596.
Wasserman, S., & Faust, K. (1994). Social network analysis. Cambridge, MA: Cambridge University Press.
Watts, D. (2004). The "new" science of networks. Annual Review of Sociology, 30, 243–270.
Watts, D., & Strogatz, S. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440–442.
Wellman, B. (1999). Living networked on and offline. Contemporary Sociology, 28(6), 648–654.
Wellman, B., & Giulia, M. (1999). Net surfers don't ride alone: Virtual communities as communities. In B. Wellman (Ed.), Networks in the Global Village, 331–367. Boulder, CO: Westview Press.
Wellman, B., & Haythornthwaite, C. (2002). The Internet in everyday life. Malden, MA: Blackwell Publishing.
Wellman, B., Hogan, B., Berg, K., Boase, J., Carrasco, J.-A., Côté, R., et al., (2006). Connected lives: The project. In P. Purcell (Ed.), Networked Neighbourhoods, 157–211. Guildford, UK: Springer.
Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework and virtual community. Annual Review of Sociology, 22, 213–238.
Zmud, R. (1979). Individual differences and MIS success: A review of the empirical literature. Management Science, 25(10), 966–979.

DOI: 10.1002/asi