Available online at www.sciencedirect.com
SCIENCE DIRECT®

Social Networks 26 (2004) 155–174

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

Structure and time evolution of an Internet dating community

Petter Holme [a,*], Christofer R. Edling [b], Fredrik Liljeros [b,c]

[a] Department of Physics, Umeå University, S-901 87 Umeå, Sweden
[b] Department of Sociology, Stockholm University, S-106 91 Stockholm, Sweden
[c] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, S-171 77 Solna, Sweden

Abstract

We present statistics for the structure and time evolution of a network constructed from user activity in an Internet community. The vastness and precise time resolution of an Internet community offers unique possibilities to monitor social network formation and dynamics. Time evolution of well-known quantities, such as clustering, mixing (degree–degree correlations), average geodesic length, degree, and reciprocity is studied. In contrast to earlier analyses of scientific collaboration networks, mixing by degree between vertices is found to be disassortative. Furthermore, both the evolutionary trajectories of the average geodesic length and of the clustering coefficients are found to have minima.

Keywords: Internet community; Time evolution; Communication; Complex networks

1. Introduction

With the growing interest in social network analysis from the physics community, a new research area is emerging in the intersection between statistical physics and sociology (Albert and Barabási, 2002; Dorogovtsev and Mendes, 2002a,b; Newman, 2003). Sociologists have been interested in network analysis for at least half a century, and with mathematicians and statisticians they have developed a set of tools to analyze positions, structures, and processes of social networks (Wasserman and Faust, 1994; Butts, 2001). Although there are exceptions (Fararo and Sunshine, 1964; Skvoretz, 1990), most sociological and anthropological studies of networks have focused on small-group interaction

* Corresponding author. Tel.: +46-90-7867760; fax: +46-90-7866673.
E-mail address: holme@tp.umu.se (P. Holme).

0378-8733/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2004.01.007

or cognitive networks. In one respect this is quite natural as most groups and formal organizations are of small size. Also, a pragmatic reason for this is that data collection of large social networks, behavioral or cognitive, is cumbersome and often practically impossible to carry through. Therefore, although recent analyses (Watts and Strogatz, 1998; Watts, 1999; Newman, 2001a,b) have brought new attention to comparative analysis of large-scale social networks, the statistical physics method, emphasizing the limit of large system sizes (Albert and Barabási, 2002), has been of limited utility. However, the extended use of database technology provide new possibilities for constructing real world networks for the analysis of, e.g. movie–actor networks (Watts and Strogatz, 1998) and co-authorship in science (Newman, 2001a,b). Surely, these networks reflect social interaction, but they are also heavily constrained by the logic of a particular industry or a particular professional activity. Thus, to allow for exploration of the possible universal properties of social networks in general, there is still an urgent need to analyze other types of large empirical social networks. In this paper we report on an investigation of a large social network, aiming to give a phenomenological description that will hopefully shed some new light on the processes forming the structure of social networks. To put results in context, we try to compare our findings to other studies whenever possible, and to contrast parameters to what would be expected from a random network with similar characteristics.

To construct network data and large graphs based on more spontaneous patterns of human interaction than, e.g. co-authorship and co-actorship, one can consider data from e-mail exchange (Ebel et al., 2002) or user activity in Internet communities (Rothaermel and Sugiyama, 2001; Smith, 2002). The present work belongs to the latter category, with a strong focus on the dynamics of the network. In contrast to previous studies of Internet communities (Smith, 2002), we use down-to-the-second timing of the communication to investigate time evolution and obtain steady-state estimates of well-known measures of graph structure. We use data from a Swedish Internet community called pussokram.com (roughly "kiss‘n’hug" in English) that is primarily targeted at adolescents and young adults. The community provides an arena for flirting, dating, and other romantic communication; as well as communication for non-romantic friendship.

Studies suggest that online interaction is driven by the same needs as face-to-face interaction, and should not be regarded as a separate arena but as an integrated part of modern social life (Wellman and Haythornthwaite, 2002). Thus, communicative actions taken by members of the community can be expected to share many features with the web of human acquaintances and romances in the social offline world. Indeed, for many people in contemporary western societies, interaction on the Internet is as real as any other interaction (Wellman, 2001). Internet communities are interesting by and for themselves, but this suggests that the formation and dynamics of social networks in an Internet community can share the same generic properties as all social acquaintance networks, and that the study of Internet communities can provide important information for enhancing our understanding of social networks in general.

The paper is divided into four sections. In Section 2, we give a detailed description of the functions of the Internet community in focus. Section 3 contains statistical analyses and presentation of results that we summarize and discuss in Section 4.

2. The Internet community pussokram.com

pussokram.com is a Swedish Internet community primarily intended for romantic communication and targeted at adolescents and young adults. The community had around 30 000 active users during the spring and summer 2002, the mean user age is 21 years, and approximately 70% of the users are women (therefore, and to simplify, we will use the female gender when referring to users in this paper). Both age and sex are self-reported. It is possible to have multiple accounts on the community. A crude check on the number of accounts linked to every unique e-mail address indicates that this is not very common (more than 99.7% of the membership accounts are associated with a unique e-mail address and no e-mail address are associated with more than five accounts).[1] Our data consist of all the user activities on pussokram.com logged for 512 days from 13:39:25 h on 13 February 2001 ($t = 0$) to 13:28:19 h on 10 July 2002. The smallest time-unit on the log is 1 s. We analyze the activity of all users registered at time $t = 0$, as well as the activity of any new users during this time span.[2] Time $t = 0$ defines the start up day for this particular community. However, prior to $t = 0$ there was a mail server for sending anonymous love messages on the Internet. Registered users of this service had their accounts automatically transferred to pussokram.com. We only study activity on the community, nevertheless this recruitment might induce higher initial growth of active users.

pussokram.com has a pronounced romantic profile, where

- users are encouraged to send messages to others that they are secretly in love with;
- the provider answers questions related to love and sex posed by the users under the pseudonym Dr. Love;
- the design of the HTML pages makes use of a romantic iconography well known to the targeted users (with Valentine's hearts, deep red colors, etc. see Fig. 1). Nevertheless, a quick glance through some of the public guest books reveals that many of the contacts taken are also non-romantic.

2.1. Types of contacts in pussokram.com

There are four major modes of communication at pussokram.com. We study each of the networks generated by these four types of contacts separately and we also study the union of these networks generated by any of these contacts. A brief description of the four types of contacts follows:

- The *Messages* are in effect intra-community e-mails. These are private in the sense that no one in the community, except the sender and receiver, can access them. Not even information on how many messages other users have received are retrievable for other users.

[1] Of course it is possible to use an unique e-mail address for every unique e-mail account but since this information is not revealed its hard to see way on would go through the extra effort so doing.

[2] Personal integrity is of course an issue here. For the analysis, we study the anonymized data to prevent any intrusion of privacy, and we do not have access to specific message contents. Like everyone else, we can read the guest books, but still we cannot link an user (and her guest book) to the vertices of the network. Thus, we cannot identify any specific individual person in the data. We do not even have data that can be cross-examined with other databases (like computer IP-addresses) to detect users identity.

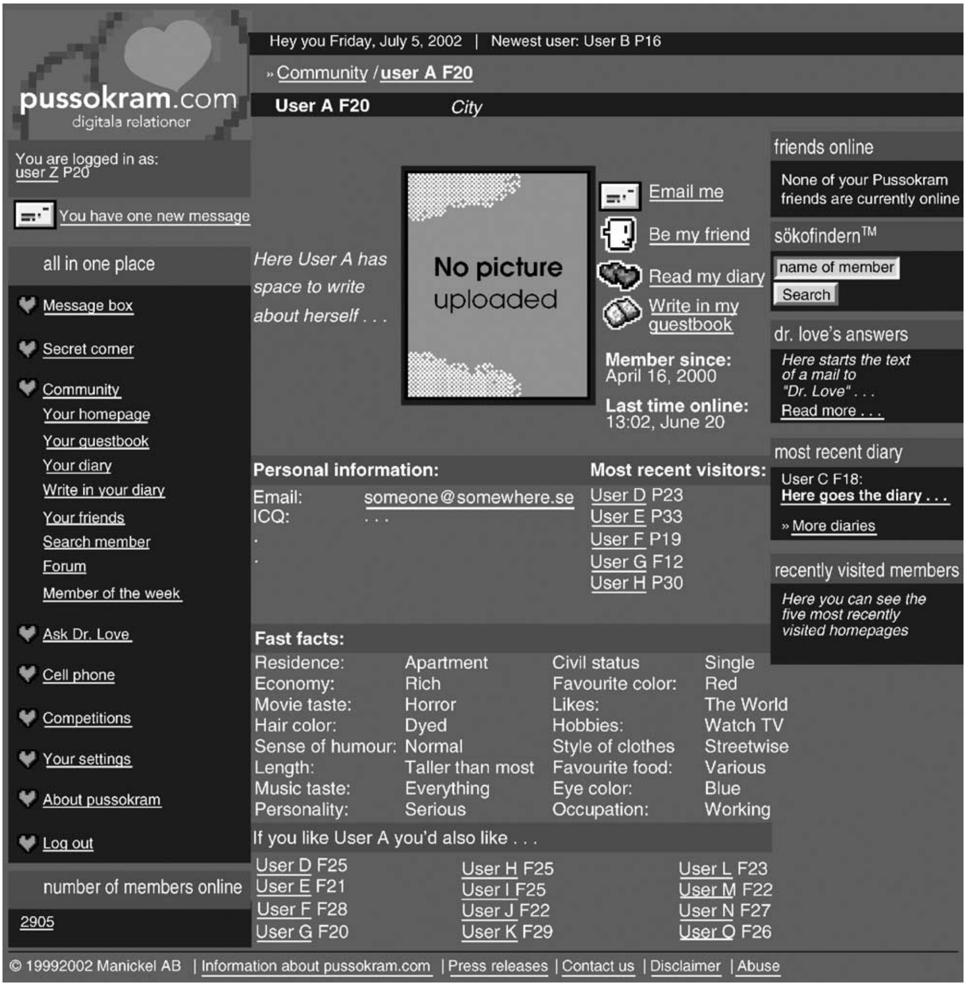

Fig. 1. Screenshot of a typical user homepage at pussokram.com. "User A", "User B", etc. symbolize user names. (The translation is due to the authors. Italics denote a description rather than a translation.)

- In *Guest book* signing, each user has a guest book that every community member is free to write in.
- *Flirt* or "friendship request": user A can ask user B to be her friend. If user B accepts user A's request then they can both easily see if the other is online whenever they are logged onto pussokram.com. Information on the friends of a specific user is private to the user only.
- *Friendship*: a friendship relation is established after acceptance of a friendship request, as described above. The friendship network is thus bi-directional. A friendship can be canceled by any of the friends.

2.2. Ways to receive attention and search users

Unless engaged in peer-to-peer contact of some sort, users at pussokram.com are relatively anonymous towards each other. There is reason to believe that knowledge about the prior interactive behavior of other individuals structures the present interactive behavior of a given individual (the so-called imitation factor). The only information about a user's interaction history available to other users. But there are several ways for an user to draw attention to herself (i.e. to direct other users to her community homepage), and for users to find information about others. Here we summarize various ways that can be used to receive attention, search for other users, and promote oneself at pussokram.com. The following information is displayed when a logged on user browse the pussokram.com website:

- the username of the most recently registered community member;
- the name of the most recently edited diary (each user has space open for others to read, intended as a diary);
- the names of the most recent users to browse a specific user's homepage;
- the names of similar users are displayed on a specific users homepage. Similarity is assesses through self-reported background variables;
- a long interview with the "user of the week" (although updated more seldom than weekly). This is an epithet that users can apply for
- photographs of 10–20 users are displayed at the login-page.

A user can search out other users with a search engine (the "sökofinder"—in English "search'n'finder"—in Fig. 1) that handles the following criteria: sub-string of the username, gender, age, place of residence, online status, and if a user has provided a photograph of herself. Presumably, these are the characteristics that drive user activity, but because it is hard to assess their validity, and because we are only interested in structural properties, we do not conduct any analysis on them.

2.3. Comparisons with other empirical and statistical networks

For comparison we also use networks by instant messaging at the French Internet community nioki.com and scientific collaboration (or, rather, co-authorship) networks. nioki.com and pussokram.com are rather similar, both in terms of content and design, but compared to pussokram.com, nioki.com is even more youth oriented and not as focused on romantic relations as pussokram.com. Besides the possibility of searching for user names, nioki.com has two search procedures *recherche l'amitié* (search for friendship) and *recherche l'amour* (search for love), where one can fill out questionnaires to find other users that match ones preferences. In the nioki.com network, an arc connects user A to B if user B is in A's list of contacts (for details see Smith (2002)). In the scientific collaboration networks (Newman, 2001a,b) the vertices are scientists who have uploaded manuscripts to the Los Alamos preprint repository arXiv.org, arcs are added between scientists who have co-authored a paper. In contrast to the pussokram.com and nioki.com networks, ties in the scientific collaboration network is bi-directional. Note, that the pussokram.com networks are dynamic, while we only have access to snapshot data of nioki.com and scientific collaboration networks. For this reason we can only make comparisons between the static properties of these networks.

In addition, following (Anderson et al., 1999; Pattison et al., 2000; Shen-Orr et al., 2002), we compare some observed quantities to the corresponding average values from randomized networks with the same degree-sequence as the original. By this approach, we examine how aspects of structures other than the degree-sequence, influences the quantities. Every known real social network deviates from the average randomized network in a larger or lesser extent, depending on the social forces structuring the interaction. For example, with regards to the present case, we believe that an Internet community network will be closer to the average randomized network than several other types of social networks, because time and space constraints are much less pressing than in, e.g. a kinship network. These randomized networks are generated by sequentially going through all directed arcs A–B, and for every such arc randomly select another arc, C–D, and then rewire so that A–D forms one arc, and C–B forms another. The choice of C–D is done with uniform randomness among all arcs that would not introduce a loop or a multiple arc. We use this algorithm to generate $\sim$3000 networks and the quantities are averaged over these networks. This procedure is inspired by Roberts (2000). However, it differs from Roberts in the sense that we use sweeps over all arcs (where each arc is rewired at least once) as the unit of iterations of the algorithm.[3]

3. Statistical analysis

The pussokram.com network consists of all registered users and the communication flow between these users as described above. Communication is conceived of as directed links between users. This is translated into a graph of vertices (users) and arcs (ties). Vertices are added to the network the first time a registered user is active, i.e. the first time the user sends or receives a message, signs a guest book, or sends or accepts a friendship request as described above. Each of these interactions defines a unique network, and by adding an arc for any activity one gets a total network of online activities. We thus study five networks, and for each of them the vertex set is empty at $t = 0$. We represent the network as a directed graph, $G = (V, A)$, where V is the vertex set and A is the set of arcs, or ordered pairs of vertices. $N = |V|$ denotes the order (number of vertices) of G, and $M = |A|$ represents the number of arcs. Sometimes we study properties of the undirected graph obtained by taking the reflexive closure of G.[4]

3.1. Decreasing growth rate of network size and convergence of average degree

For each network, the number of vertices of each network, N, as a function of time during the sampling is displayed in Fig. 2(a), and the average degree, i.e. the average number of arcs per vertex, M/N, is displayed in Fig. 2(b). As can be seen, both the number of vertices and the average degree are increasing as a function of time, but with at a decreasing growth rate. The average degree appears to converge to a constant, but for $t < 100$, it

[3] To be precise our algorithm run as follows: we go sequentially through the arc set A (see Section 3). For every arc (v, w) we construct a set A' of arcs such that if a member (v', w') of A' is to be rewired with (v, w)—i.e. so that (v, w) and (v', w') are replaced by (v, w') and (v', w)—then no loops or multiple arcs are formed. Then we choose one of A's arcs with uniform randomness and rewire that arc with (v, w).

[4] That is, the graph obtained if for every $(u, v) \in A$ and $(v, u) \notin A$ then (v, u) is added to A.

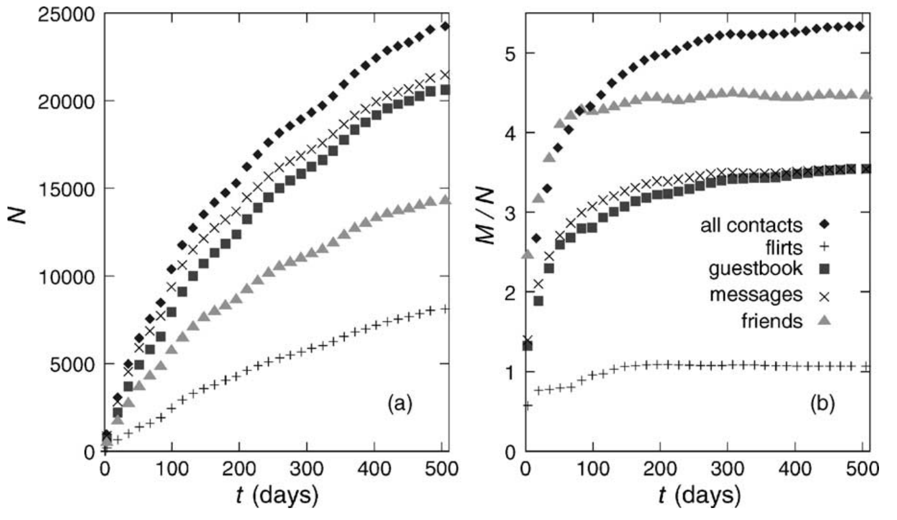

Fig. 2. Time evolution of the number of vertices (a) and average degree (b) as a function of time.

increases as a power function. The more rapid growth rate in the beginning of the period is explained by the fact that old users log on for the first time during our sampling period (see discussion in Section 2). The decreasing growth, and apparent approach to equilibrium, stand in contrast to the accelerated growth of the Internet and the World Wide Web (Dorogovtsev and Mendes, 2002a,b), as well as the linear growth of scientific co-authorship networks extracted from article databases (Newman, 2001a,b; Barabási et al., 2002). However, in social networks, the average degree cannot be increasing without bounds, and this goes for scientific collaboration networks too. We believe the difference stems from a wider effective sampling time frame—due to the much more rapid dynamics of an Internet community (compared to scientific collaborations) we are, relatively speaking, able to follow the process for a much longer period. In the sense that G is a steadily growing dynamic network, we deal with a non-equilibrium representation of the social situation. When we speak of the network "reaching equilibrium", we refer to when all quantities that are bounded as a function of N (such as the average degree) are reaching their constant limits.

3.2. Reciprocity varies between networks

Various types of social relations differ in direction, intensity, and frequency (Granovetter, 1973). Messages between agents with different social status, e.g. tend to be unevenly distributed (Gould, 2002). In the present analysis, we can investigate the reciprocity of communicative action by looking at the direction of the communication flow between any two users. For example, if user A sends a friendship request to user B, we observe a link between users A and B, and note an arc between the two vertices. But it makes quite a difference whether user B accepts the invitation or not, i.e. whether we note one or two arcs between the vertices. We define reciprocity R, as the fraction of mutual dyads, i.e. the ratio between the number of vertex-pairs $\{v, w\}$ occur in two arcs ((v, w) and (w, v)) and vertex-pairs that

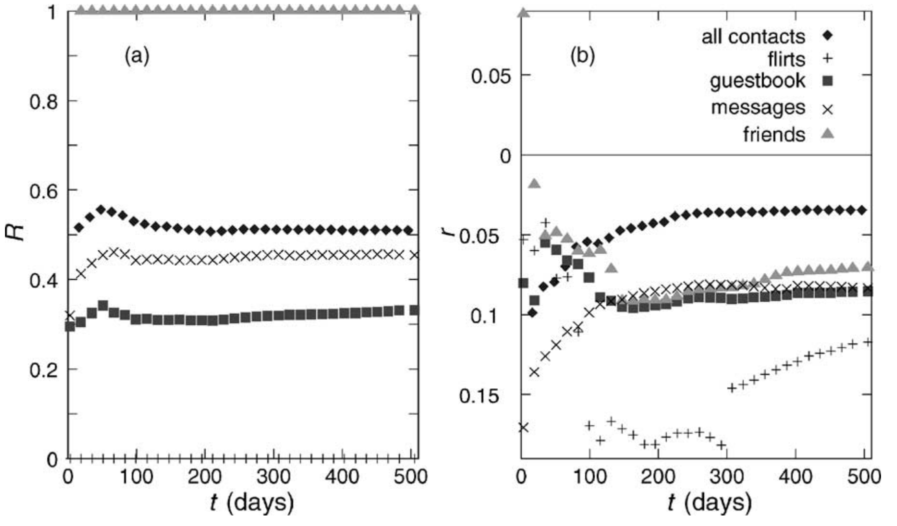

Fig. 3. (a) Reciprocity R and (b) assortative mixing coefficient $r_{\rm dir}$ as functions of time.

occur in at least one arc. More analytically:

$$R = \frac{2M}{M_2} - 1 \tag{1}$$

where M_2 is the number of arcs in the reflexive closure of G. R lies strictly in the interval $[0, 1]$; if (u, v) is an arc then $R = 0$ implies that (v, u) is not an arc and $R = 1$ implies that (v, u) is an arc.

The time evolution of the reciprocity can be seen in Fig. 3a. As is evident from the figure, reciprocity levels differ little between the different networks. By definition, the friendship network has reciprocity of 1. And by the same token, the flirt network has a reciprocity equal to 0. For the other two networks, the curves converge to values around 0.4 for the guest book and messages networks, and 0.5 for the all contacts network (see Table 1). It is hard to judge whether these are high or low values of reciprocity. They are however compatible with data for the French Internet community nioki.com. We normally assume acquaintance networks to have a high degree of reciprocity, but one reason to expect a lower value for online interaction is that, e.g. an actor feels less social pressure to respond to a communicative act over the Internet than in a face-to-face, or telephone encounter.

3.3. *Disassortative mixing coefficients of the pussokram.com networks*

Together with the degree distribution, the degree–degree correlation is considered to govern much of the network's robustness towards disturbances as well as the information flow. In other contexts the discussion is usually phrased in terms of resilience against epidemics and attack. A positive degree–degree correlation is also referred to as assortative mixing by degree, and it means that vertices of high degree preferably attaches to each

Table 1
Assortative mixing coefficients, r, for five pussokram.com networks, and for nioki.com and arXiv.org networks

Network	N	r	r_{dir}	$r_{\mathrm{in\,in}}$	$r_{\mathrm{in\,out}}$	$r_{\mathrm{out\,in}}$	$r_{\mathrm{out\,out}}$
All contacts	29341	−0.048* [−0.043]	−0.059* [−0.041]	−0.063* [−0.028]	−0.046* [−0.021]	−0.071* [−0.049]	−0.050* [−0.035]
Messages	21545	−0.055* [−0.053]	−0.083* [−0.061]	−0.054* [−0.013]	−0.056* [−0.011]	−0.076* [−0.058]	−0.087* [−0.057]
Guest book	20691	−0.073* [−0.049]	−0.085* [−0.038]	−0.097* [−0.024]	−0.043* [−0.015]	−0.088* [−0.042]	−0.053* [−0.026]
Friends	14278	−0.042* [−0.031]	–	–	–	–	–
Flirts	8186	−0.12* [−0.12]	−0.12* [−0.10]	−0.006 [0.016]	−0.022 [−0.002]	−0.12* [−0.10]	−0.042* [−0.013]
nioki.com	50259	−0.13 [−0.034]	−0.10 [−0.014]	−0.088 [−0.018]	−0.084 [−0.014]	−0.10 [−0.020]	−0.095 [−0.016]
arXiv.org	52909	0.36 [−0.034]	–	–	–	–	–

Statistics for corresponding randomized networks are within square brackets. Differences between the various mixing coefficients are discussed in the text.
* $P \leq 0.01$ (nioki.com and arXiv.org data are not tested for significance).

other, and vice versa. For example, assortative mixing makes the networks more vulnerable to outbreaks of diseases, and more robust against strategic attack (Newman, 2002), because if people with many contacts are connected to other people with many contacts, the epidemic threshold will be lowered. Disassortative mixing, on the other hand, gives rise to larger epidemics (Morris and Kretzschmar, 1995).

We measure assortative mixing by calculating Pearson's correlation coefficient r for the degrees at either side of an edge as suggested by Newman (2002):

$$r = \frac{\langle k_{\mathrm{to}} k_{\mathrm{from}} \rangle - \langle k_{\mathrm{to}} \rangle \langle k_{\mathrm{from}} \rangle}{\sqrt{\langle k_{\mathrm{to}}^2 \rangle - \langle k_{\mathrm{to}} \rangle^2} \sqrt{\langle k_{\mathrm{from}}^2 \rangle - \langle k_{\mathrm{from}} \rangle^2}} \tag{2}$$

In Eq. (2), $\langle \cdot \rangle$ denotes the average over arcs, k_{from} is some (in, out, or total) degree of the vertex that the arc starts from, and k_{to} is some degree of the vertex that the arc leads to. We look at r for total degree of both bi-directional (where the reflexive closure has been taken if the network is not bi-directional by definition) and directed graphs r_{dir}. Furthermore, we measure the four combinations of in- and out-degree correlations, e.g. the out–in correlation coefficient indicates whether users that have many contacts (high out-degree) prefers to communicate with those users that themselves receive communication from many users (high in-degree).

The values for pussokram.com and other networks are displayed in Table 1. Interestingly enough all the pussokram.com networks, as well as the nioki.com network display a significant disassortative mixing for all types of degree–degree correlations. This is in contrast to what have been measured for (scientific, actor, and business) collaboration networks (Newman, 2002). To set these results in perspective we also measure r for a scientific collaboration network, which clearly displays a positive assortative mixing coefficient. May be an assortative mixing is significant only to interaction in competitive areas, such as professional collaborations (where only already big names are likely to be successful in collaborating with other big names). This result relates to research on exchange networks that claim that negative mixing is optimal when actors are substitutable, as for example, in friendship and dating network (Cook et al., 1983). In contrasts, professional collaboration is positive because both knowledge and already established channels for co-operation screen off potential alternative collaborators. Another issue is the skewness of the degree distribution. Intuitively, a large spread in the degree distribution will increase the likelihood of observing negative mixing. And as can be seen from the randomized networks in Table 1, given the degree distribution we would expect a negative mixing coefficient. However, the observed coefficients are consistently, and significantly, higher than expected. This strongly suggests that negative mixing arise from this particular form of social interaction in which alters are substitutable (Cook et al., 1983). Note though, that some network models, analyzing completely different forms of interaction, with skewed degree distributions produce networks of zero or positive assortative mixing (Newman, 2002; Park and Newman, 2003).

The six different assortative mixing coefficients of Table 1 are all of the same sign and roughly of the same magnitude. This is interesting since it suggests that the r-values is a result of other structures (presumably the degree-sequence) rather than from the behavior of individuals. There are no a priori reasons for $r_{\mathrm{in\,out}}$ to be the same as, e.g. $r_{\mathrm{in\,in}}$, as a large $r_{\mathrm{in\,out}}$ means that actors that are active in the community (have a high k_{out}) tend to associate with those who are successful in promoting themselves in the community (have a high k_{in}), while a large $r_{\text{in in}}$ means that the latter category has a preference towards each other.

Fig. 3b shows the time development of the assortative mixing coefficient r_{dir} (the time development of the other assortative mixing coefficients of Table 1 is qualitatively similar). We see that r_{dir} converges more quickly than the average degree. This is not surprising since the correlation coefficient is a function of the way ties are formed rather than the size or average degree of the network. An interesting detail of Fig. 3b is the jump at $t \approx 300$ days in the flirt (friendship request) network. This is due to the formation of a tie between two of the most connected actors (the fact that the flirt network is by far the sparsest strengthens this effect).

3.4. Cumulative degree distributions are highly skewed

The degree distribution has received much attention in comparative analyses of complex networks since the work of Barabási and Albert (1999). A skewed degree distribution is commonly regarded as a cumulative effect in the attachment of new arcs to the network (Simon, 1955; Barabási and Albert, 1999), and it offers a way to classify different types of networks (Amaral et al., 2000). Indeed it has been demonstrated that many apparently dissimilar types of networks share the same highly skewed degree distributions of a (truncated) power-law form (Albert and Barabási, 2002), indicating an emerging scale-free structure. Such degree distributions are generated through a growth process in which new arcs are drawn between already existing vertices and new vertices only. However, a process that reasonably describes the activity of an Internet community would allow also for new arcs to be drawn between two already existing vertices. Such a mixed process however, would result in a stretched exponential distribution, and not a power-law, and thus a stretched exponential distribution is what we would expect to observe. Another process that can be responsible for cutting the tails of power-law degree distributions in real-world networks is a limited capacity of the actors.

Following (Liljeros et al., 2001) we measure the cumulative degree distribution of all the pussokram.com networks (see Fig. 4) . If the degree distribution follows a power-law with exponent γ then the cumulative distribution will have the exponent $\alpha = \gamma + 1$. All pussokram.com networks are highly skewed, but none of them fits a power-law form across the whole range observed. However, it is interesting to note that there are no clear signs of the (inevitable) high-degree truncation in any of the graphs (Fig. 4). A previous study of the French nioki.com has reported a power-law fit of the cumulative degree distribution (Smith, 2002). Our result might appear to set the pussokram.com community apart from the nioki.com community, but a closer inspection of our graphs and (Smith, 2002) reveals a striking similarity in the functional form of the distribution. We therefore conclude that the dynamics shaping the degree distribution is to a large extent the same for the two communities.

3.5. Evolution of average geodesic length

As a general measure of how closely connected a graph is, the average geodesic (shortest path) length is one of the most studied network quantities. There is no unique natural

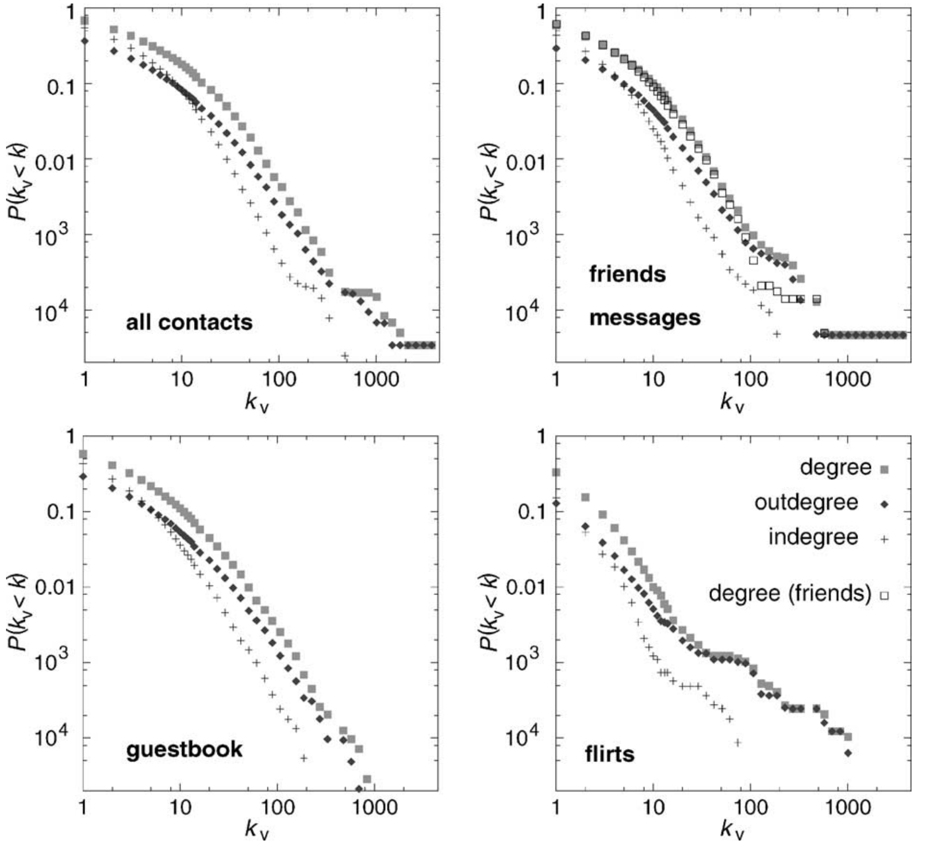

Fig. 4. Cumulative degree distribution for the networks at the largest times for (a) all contacts, (b) friendship confirmations and messages, (c) guest book and (d) flirts.

definition of average geodesic length in an arbitrary directed graph—the problem is the contribution from disconnected pairs of vertices. One choice is to measure the geodesic distance averaged over pairs of vertices in the giant component:

$$l_{\mathrm{GC}} = \frac{1}{|A_{\mathrm{GC}}|} \sum_{(u,v)\in A_{\mathrm{GC}}} d(u, v) \tag{3}$$

where $d(u, v)$ is the distance between u and v, and A_{GC} is the arc-set of the giant component. Another option is to average the inverse geodesic length (Latora and Marchiori, 2001):

$$l^{-1} = \frac{1}{M} \sum_{(u,v)\in A} \frac{1}{d(u, v)} \tag{4}$$

where $1/d(u, v)$ is defined as zero when no path exists from u to v. In the present paper, we focus on l^{-1}, and l_{GC} for the reflexive closure of G. If the two measures agree, we can infer

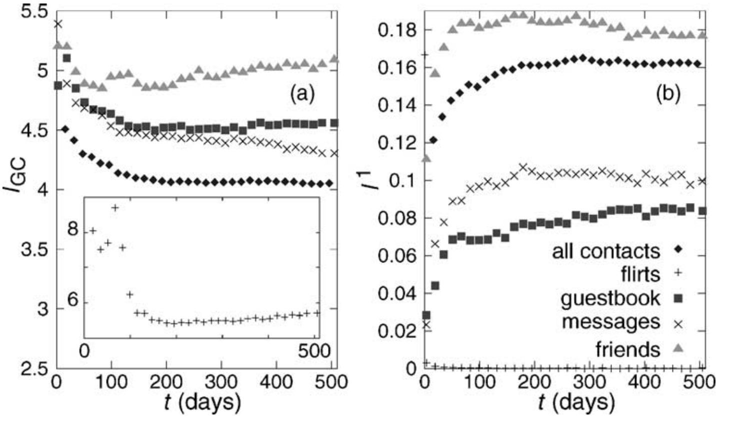

Fig. 5. Time evolution of the average geodesic length within (a) the giant component of the reflexive closure and (b) the average inverse degree.

that there is no additional effect influencing the shortest paths in a substantial way, other than the bi-directional structure of the largest connected subgraph.

As time evolves there are two conflicting mechanisms governing the average geodesic length. The increasing number of vertices works for an increase of l, whereas the increasing average degree makes l shorter. For the pussokram.com data the latter effect dominates, during the time span of our dataset, to give a monotonously decreasing l_{GC} (monotonously increasing l^{-1}) as shown in Fig. 5. The same situation has been reported for scientific collaboration networks (Barabási et al., 2002). Assuming the community outlives its members, l will eventually start to increase (when the number of inactive users slows down the accelerated growth sufficiently).

3.6. Density of short circuits

Acquaintance networks are expected to have a high degree of transitivity (Wasserman and Faust, 1994), or in other words, a high density of triangles, since if person A knows B and C, then persons B and C are likely to be acquainted. We apply a commonly used measure that gives the fraction of triangles out of the connected three paths of the graph (a quantity that was defined for undirected graphs, but is trivially generalized to directed graphs, for which we use subscript "dir"). If we let $p(n)$ denote the number of representations of paths[5] and $c(n)$ denote the number of representations of circuits, of length n, then we can express the clustering coefficient,[6] C, as

$$C = \frac{c(3)}{p(3)} \tag{5}$$

[5] A representation of a path of length three is a triplet (u, v, w) such that (u, v) and (v, w) are arcs. In an undirected network a path have two representations and a triangle has six representations.

[6] This quantity is sometimes called transitivity, sometimes clustering coefficient. Note however that is not identical to the Watts and Strogatz (1998) of clustering coefficient (where they average a local transitivity measure over the vertex set).

One can expect that social networks with many heterosexual romantic relationships, such as the pussokram.com networks, to have rather few triangles.[7] To get a better picture of the density of short circuits we also measure the density of circuits of length 4:

$$D = \frac{c(4)}{p(4)} \tag{6}$$

The n-behavior of $c(n)/p(n)$ varies from network to network, and could possibly be an informative quantity in itself. A very high C will in most cases probably imply a high D (for $R = 1$ network, two triangles with one arc in common will contribute to $c(4)$), but the reverse is less certain.

Values for C_{dir} and D_{dir} and their undirected counterparts are shown in Table 2. We note that, with a few exceptions, the values for the real networks are significantly larger than the randomized; the difference, however, is far less dramatic than for the scientific collaboration network. This is contrast between the Internet community networks and the arXiv.org data is easily explained from the fact that a paper with $n_{\mathrm{auth}} \geq 3$ authors represents a fully connected subgraph of G (contributing with $n_{\mathrm{auth}}(n_{\mathrm{auth}}-1)(n_{\mathrm{auth}}-2)/3$ triangles). However, we would like to stress that the values themselves are not very informative, compared to their time dependence.

The time development of C and D for different networks is shown in Fig. 6. As a quantity depend on only the local network structure the density of short circuits is an intrinsic quantity; and, as seen for the clustering coefficient (Barabási et al., 2002), these quantities approach their equilibrium values from above. Interestingly, just as for the assortative mixing coefficient, the relaxation towards equilibrium is faster for C and D than for the average degree M/N, i.e. the density of short cycles is rather independent of the average degree.

As can be seen in Fig. 6, most C and D curves have extremes in the middle of the time range (the density of short circuits are at their minima). The reason for this comes from a conflict between counteracting mechanisms of different time-scales. There are three natural time-scales in the system: the average time between new registrations; the average time between new contacts for an individual user; the average life-span of a user in the community. The latter time-scale should be responsible for the long-term behavior such as the increase towards equilibrium of M/N. And as shorter circuits are more likely in a dense network, it is natural that C and D increase in the large t-limit. The decrease for early times is a finite size effect that can be seen in evolving network models with constant average degree such as the Barabási–Albert model (Barabási and Albert, 1999; Barabási et al., 1999, 2002) and extensions (Holme and Kim, 2002), where the C and D curves converge from above.

Another interesting aspect is that the values of C and D, although finite in the large t-limit, is much smaller than in the actor- and scientific-collaboration networks. In an Internet community the way by which people introduce strangers among their acquaintances to each other (Newman, 2001a,b; Holme and Kim, 2002) is likely not the mechanism responsible for

[7] Presumably, homosexual relationships are not the common type of romantic relationship among Swedish adolescents. Therefore, we expect few triangles. As a corollary, in a community populated largely by homosexual individuals, the number of triangles would be much higher. Regrettably we cannot test this hypothesis with available data.

Table 2
Statistics for the fully grown networks of pussokram.com, nioki.com and arXiv.org networks provided for comparison

	Network						
	All contacts	Messages	Guest book	Friends	Flirts	nioki.com	arXiv.org
N	29341	20691	21545	14278	8186	50259	52909
M	174662	76257	73346	31871[a]	8744	405742	490600[a]
R	0.51	0.40	0.38	1	0	0.69	1
l_{GC}	4.4	4.3	4.6	5.1	5.7	4.1	6.1
l^{-1}	0.12	0.10	0.084	0.18	4.0×10^{-4}	0.209	0.121
C	0.006 [0.006]	0.001* [0.002]	0.014* [0.007]	0.020* [0.0044]	0 [0.001]	0.0065 [0.0081]	0.45 [0.0020]
C_{dir}	0.012* [0.007]	0.005* [0.003]	0.014* [0.005]	– [0]	0* [0.0077]	0.0076	–
D	0.017 [0.009]	0.006* [0.004]	0.022* [0.008]	0.020* [0.004]	0.212* [0.004]	0.013 [0.0081]	0.35 [0.0021]
D_{dir}	0.016* [0.007]	0.008* [0.003]	0.015* [0.005]	– [0]	0 [0.0077]	0.016	–

Statistics for corresponding randomized networks are within square brackets.

[a] The ‘friends’ and ‘arXiv.org’ datasets are undirected, M denotes the number of undirected edges (which is half the number of M in a directed representation of the graph). nioki.com and arXiv.org data are not tested for significance.

* $P \leq 0.01$.

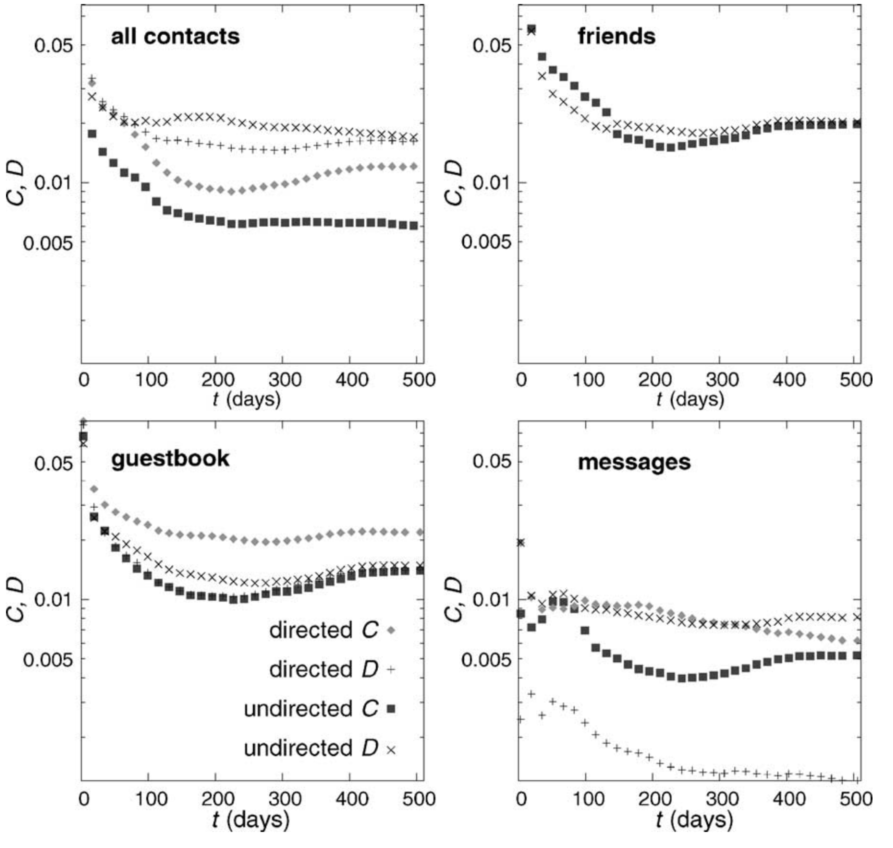

Fig. 6. Density of short circuits for the different networks (flirt network omitted as it contains very few three- and four-circuits).

the finite clustering (remember that in network models such as the Erdös and Rényi (1959) and Barabási–Albert (Barabási and Albert, 1999; Barabási et al., 1999, 2002) models the clustering goes to zero as the network grows). Instead a finite density of short circuits can be explained by the tendency formulated in the proverbial like-attracts-like, where the similarity is defined by signaled social, psychological, and physiological traits.[8]

To further convince ourselves that the sampling time is large enough we also use rewiring to examine the time evolution of two structural measures (the assortative mixing coefficient and the clustering coefficient for the undirected all-contacts network). As seen in Fig. 7, the rewired quantities converge in the same time-scale as r and C, which reconfirm that the sampling time frame is sufficient. We note that for $k > 200$ days the assortative mixing

8 Another possible explanation for the convergence of C and D to finite values is that short circuits are introduced from the offline world outside the community. Reading users' guest books, however, gives the impression that the vast majority of community-dyads were strangers offline. We believe that this effect is negligible, but we are unfortunately unable to go beyond speculation on this point.

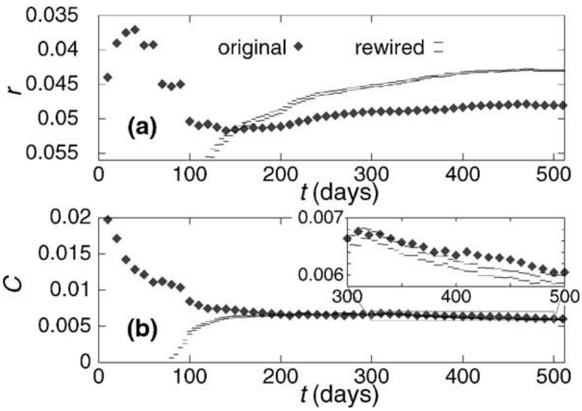

Fig. 7. Time evolution of original and rewired quantities: (a) data for the assortative mixing coefficient r for the undirected all-contacts network; (b) clustering coefficient for the same data. The rewired data is obtained from 100 updating sweeps over all links, and indicated by the upper and lower hinges (border values between the first and second quartiles, and third and fourth quartiles, respectively).

coefficient is significantly lower than the rewired reference curve. For the same time interval the rewired clustering coefficient closely overlap the measured C-value; for $t > 200$ days the actual value overlap the mid-quartiles of the rewired data during around 30% of the 512 days. For the initial 'non-equilibrium' part ($t < 100$ days) of the time evolution the curves of the rewired and real networks diverges. In this region the network is rather sparse (see Fig. 7) which explains the low C-values for the rewired C-curve. The high early values of C seems contradictory to the apparent absence of tendency towards triangle formation during latter times. This means that the contact patterns of the early network is not the same as later on. As it turns out, in the early community, a group of actors contact each other rather frequently (rather more like 'chatting' than romantic contact making), whereas another group makes a few contacts before quitting the community. We interpret this such that it requires a minimal number, or "critical mass" (cf. Schelling, 1978) of people for the community to function. Before the critical mass is reached, the users either have the community as a chat room (a usage with a presumably smaller critical mass) or leave it.

4. Summary and conclusions

We have investigated networks of communication between the users of the Internet community pussokram.com. The four different means of contact at pussokram.com defines five different networks in our study (one for each separately and one for all taken together). Apart from recent studies of scientific collaboration networks and movie actor networks, there are very few such phenomenological descriptions of large social networks, and thus there is limited knowledge that our findings can be related to.

It is obvious that the fact that the interaction under study takes place on the Internet creates special conditions for communication. We believe that the interaction online is exposed to less structural forces than what is typically the case in most other social settings. For example, simultaneous interaction is not a prerequisite for communication in an Internet community, i.e. time as a structural force is therefore of less importance than in most other settings. Neither does geographical space constraint communication. And in addition, that social signifiers are less visible (compared to, e.g. face-to-face interaction), and the relative ease with which you can conceal your identity and transform your appearance in online interaction, are factors reducing the structure forming forces at work in 'offline' social activity. It is therefore interesting to note, that despite these caveats, the networks under study here are much more structured than what would be expected in a random network.

To summarize our findings of the Internet community pussokram.com, we see that:

- The average degree converges over time, but surprisingly we observe no cut-off in the degree distribution. Previous studies do suggest that there is an upper limit to the mean number of contacts (Marsden, 1987), and on average we find this socio-cognitive limitation despite the fact that time and space is of less important here. The reason we see continued growth in the cumulative degree distribution might be that it's relatively costless to have a high turnover on ones contacts in an online community. Contacts are established without much investment, and can also be dropped without much sanctioning.
- Reciprocity is rather low, and presumably lower can be expected in a regular acquaintance network. Reciprocity levels quickly converge to a steady-state.
- Most assortative mixing coefficients have small negative values, suggesting a pattern of dissasortative mixing. This can partly be explained by the conventional effect from the skewed degree-sequence (Newman, 2002). The observed effect is significantly larger than can be expected solely from the degree distribution. An explanation for these higher r-values is the particular nature of the dating interaction (Cook et al., 1983). We also find that mixing coefficients as a function of time converge rapidly. The dissasortative mixing in the Internet community networks is in striking contrast to the strong assortative mixing seen in scientific collaboration networks, and the nice correspondence with previous work in sociology indicates that Internet communities indeed strongly resembles offline social communities.
- The cumulative degree distributions are highly skewed, being a mixture of previous mappings of acquaintance networks (Amaral et al., 2000)—for few contacts—and partnership networks (Liljeros et al., 2001)—for many contacts.
- The geodesic length initially increases as new vertices are added to the network. But as the network settles the increase is limited by the growing average degree. Both l_{GC} and l^{-1} show consistently that the average geodesic length is decreasing during the whole sample period (a situation that can only exist for a non-equilibrium network).
- Clustering—the density of triangles—converges over time to non-zero values (as opposed to completely random networks). Still, values are probably on a much lower level than would be expected in offline acquaintance networks. The explanation for these low values is two-fold—the lack of introduction as a mechanism for tie-formation, and the romantic profile of pussokram.com promoting romantic contacts. The latter aspect is also manifested in that the density of four-circuits is larger than the density of triangles for the pussokram.com networks. Once again, the Internet community networks are different

from the scientific collaboration network where clustering is larger than the density of four-circuits.

An Internet community such as pussokram.com defines a structured social network that share more of the structuring forces with general acquaintance networks than networks of professional collaborations do. We believe that the precise timing resolution and fast dynamics (giving a wide effective sampling time-frame) will make Internet communities an invaluable object for future social networks studies of the largest scale.

Acknowledgements

We thank Christian Wollter and Michael Lokner at pussokram.com, Stefan Praszalowicz at nioki.com, and Niklas Angemyr and Reginald Smith for granting and helping us getting access to data. We thank Mark Newman for comments on assortative-mixing, and the editor and anonymous reviewer for helpful comments. PH is partially supported by the Swedish Research Council through Contract No. 2002-4135. CRE is supported by the Bank of Sweden Tercentenary Foundation. FL is supported by the National Institute of Public Health.

References

Albert, R., Barabási, A.L., 2002. Statistical mechanics of complex networks. Review of Modern Physics 74, 47–97.

Amaral, L.A.N., Scala, A., et al., 2000. Classes of small-world networks. Proceedings of the National Academy of Sciences of the United States of America 97 (21), 11149–11152.

Anderson, B.S., Butts, C., et al., 1999. The interaction of size and density with graph-level indices. Social Networks 21 (3), 239–267.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Barabási, A.L., Albert, R., et al., 1999. Mean-field theory for scale-free random networks. Physica A 272 (1/2), 173–187.

Barabási, A.L., Jeong, H., et al., 2002. Evolution of the social network of scientific collaborations. Physica A 299, 559–564.

Butts, C.T., 2001. The complexity of social networks: theoretical and empirical findings. Social Networks 23 (1), 31–71.

Cook, K.S., Emerson, R.M., et al., 1983. The distribution of power in exchange networks: theory and experimental results. American Journal of Sociology 89 (2), 275–305.

Dorogovtsev, S.N., Mendes, J.F.F., 2002. Accelerated growth of networks. In: Bornholdt, S., Schuster, H.G. (Eds.), Handbook of Graphs and Networks: From the Genome to the Internet. Wiley–VCH, Berlin.

Dorogovtsev, S.N., Mendes, J.F.F., 2002b. Evolution of networks. Advances in Physics 51 (4), 1079–1187.

Ebel, H., Mielsch, L.I., et al., 2002. Scale-free topology of e-mail networks. Physical Review E 66 (Art. No. 035103).

Erdös, P., Rényi, A., 1959. On random graphs. Publicationes Matematicae Debrecen 6, 290–297.

Fararo, T.J., Sunshine, M.H., 1964. A Study of a Biased Friendship Net. Youth Development Center, Syracuse University, Syracuse.

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

Granovetter, M., 1973. Strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.

Holme, P., Kim, B.J., 2002. Growing scale-free networks with tunable clustering, Physical Review E 65 (2) (Art. No. 026107).

Latora, V., Marchiori, M., 2001. Efficient behavior of small-world networks, Physical Review Letters 87 (19) (Art. No. 198701).

Liljeros, F., Edling, C.R., et al., 2001. The web of human sexual contacts. Nature 411 (6840), 907–908.

Marsden, P.V., 1987. Core discussion networks of Americans. American Sociological Review 52 (1), 122–131.

Morris, M., Kretzschmar, M., 1995. Concurrent partnerships and transmission dynamics in networks. Social Networks 17 (3–4), 299–318.

Newman, M.E.J., 2001. Clustering and preferential attachment in growing networks. Physical Review E 64 (2) (Art. No. 025102).

Newman, M.E.J., 2001. Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E 64 (1) (Art. No. 016131).

Newman, M.E.J., 2002. Assortative mixing in networks. Physical Review Letters 89 (Art. No. 208701).

Newman, M.E.J., 2003. The structure and function of complex networks. SIAM Review 45 (2), 167–256.

Park, J., Newman, M.E.J., 2003. Origin of degree correlations in the Internet and other networks. Physical Review E 68 (2) (Art. No. 026112).

Pattison, P., Wasserman, S., et al., 2000. Statistical evaluation of algebraic constraints for social networks. Journal of Mathematical Psychology 44, 536–568.

Roberts, J.M., 2000. Simple methods for simulating sociomatrices with given marginal totals. Social Networks 22 (3), 273–283.

Rothaermel, F.T., 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.

Schelling, T.C., 1978. Micromotives and Macrobehavior. Norton, New York.

Shen-Orr, S.S., Milo, R., et al., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genetics 31 (1), 64–68.

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

Skvoretz, J., 1990. Biased net theory: approximations, simulations, and observations. Social Networks 12 (3), 217–238.

Smith, R., 2002. Instant messaging as a scale-free network, unpublished. cond-mat/0206378.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.

Watts, D.J., 1999. Networks, dynamics, and the small-world phenomenon. American Journal of Sociology 105 (2), 493–527.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of ‘small-world’ networks. Nature 393 (6684), 440–442.

Wellman, B., 2001. Computer networks as social networks. Science 293 (5537), 2031–2034.

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