Statistics for the structure and time-evolution of a large, complex, evolving, real-world network is presented. The data comes from the communication flow in an Internet community of around 25 000 users with user activity logged for about 500 days. The vastness and precise time resolution of an Internet community offers unique possibilities to monitor social network formation and dynamics. The data allows us to study the dynamic process during an extended period of time compared to previous studies. Time evolution of well-known quantities, such as clustering, mixing (degree-degree correlations), average geodesic length, degree, and reciprocity is studied. The effective sampling time is long enough for most of these measures to converge. In contrast to earlier analyses of scientific collaboration networks, mixing between vertices is found to be disassortative. Furthermore, both the evolutionary trajectories of the average geodesic length and of the clustering coefficients have minima.

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 (Buchanan, 2002). 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 (Skvoretz, 1990; Fararo and Sunshine, 1964), most sociological and anthropological studies of networks have focused on small-group interaction or cognitive networks. One reason for this being that data collection of large social networks, behavioral or cognitive, is cumbersome and often practically impossible to carry through. Therefore, although recent analysis (Newman, 2001b; Watts and Strogatz, 1998; Watts, 1999) 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, 2001b). 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 urging need to analyze other types of large empirical social networks. In this paper we report 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, Mielsch, 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, in press). 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 "off-line" 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 the next section we give a detailed description of the functions of the Internet community in focus. The third section contains statistical analyses and presentation of results that we summarize and discuss in the fourth and concluding section.

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 25 000 active users during the spring and summer 2002, the mean user age is 21 years, and approximately 70 percent of the users are women (therefore, and to simplify, we will use the female gender when referring to users in this paper). Our data consists of all the user activities on pussokram.com logged for 512 days from 13:39:25 on February 13, 2001 \((t = 0)\) to 13:28:19 on July 10, 2002. The smallest time-unit on the log is 1 second. 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.\(^1\)

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.

\(^1\) 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 don’t even have data that can be cross-examined with other databases (like computer IP-addresses) to detect users’ identity.
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 network generated by any of these contacts. A brief description of the four types of contacts follows:

- **Messages** are e-mail-like and 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 has received are retrievable for other users.
- **In Guest book** signing, each user has a guest book that every community member is free to write in.

Fig. 1. Screenshot of a typical user’s homepage at pussokram.com. “User A”, “User B”, etc. symbolizes user names. (The translation is due to the authors. Italics denote a description rather than a translation.)
• **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’re 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 cancelled by any of the friends.

### 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 that is revealed publicly at pussokram.com is that of the guest book, which means that there is very little opaque 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 user’s homepage.
- 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 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.

### 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, arcs goes from user A to user B if user B is in user A’s list of contacts (for details see Smith, 2002). In the scientific collaboration networks (Newman, 2001b) 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 (Shen-Orr, Milo, 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 can how the 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 go 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. This procedure is iterated and the quantities are averaged over 100 iterations (except the nioki.com data which is averaged over 30 iterations).

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. Some times we study properties of the undirected graph obtained by taking the reflexive closure of \( G \).

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. 2a, and the average degree, i.e. the average number of arcs of each vertex, \( M / N \), is displayed in Fig 2b. As can be seen, both the number of vertices and the average degree are increasing as a function of time, but at a decreasing growth rate. The average degree appears to converge to a constant, but for \( t < 100 \), it increases as a power law. 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. 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, 2002), as well the linear growth of scientific co-authorship networks extracted from article databases (Newman, 2001a, b; Barabási, Jeong, 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.

Reciprocity varies between networks

Various types of social relations differ in direction, intensity, and frequency (Granovetter, 1973). Messages between agents with different social status for example, tend to be unevenly distributed (Gould, 2002). In the present analysis, we can disseminate the reciprocity of communicative action by looking at the direction of the communication flow between any two users. For example, if user A send a friendship request to user B, we observe a link between user A and user 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 bi-directional dyads (a dyad is a pair of actors connected by a tie in one of two directions). More analytically we define \( R \) as:
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.

\[
R = \frac{2M}{M_2} - 1
\]  

(1)

---

The time evolution of the reciprocity can be seen in Fig. 3a. As is evident from the figure, reciprocity levels differ largely 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 zero. For the other two networks, the curves converge to values around 0.3, for the guestbook network and all contacts network, and 0.45 for the messages network (see Table 2). It’s 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 an actor feels less social pressure to respond to a communicative act over the Internet than in a face-to-face, or telephone encounter, for example.

### 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, and it means that vertices of high degree preferably attaches to each 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.

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

\[
r = \frac{\langle k_{to} k_{from} \rangle - \langle k_{to} \rangle \langle k_{from} \rangle}{\sqrt{\langle k_{to}^2 \rangle} - \langle k_{to} \rangle} \left[ \sqrt{\langle k_{from}^2 \rangle} - \langle k_{from} \rangle \right]^{-1}
\]  

(2)

In equation 2, \( \langle \ldots \rangle \) 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 display a positive assortative mixing coefficient. Maybe 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). In many social settings the fitness function is not as clearly defined as in e.g., scientific collaboration. Therefore, and despite recent discussions on maximizing so-called social capital (Lin, 2001), in many social situations it is not obvious with whom it is optimal to interact. This might be especially clear in online interaction. Another issue is the skewness of the degree distribution. Intuitively, a large spread in the degree distribution will reinforce the effect of other structural biases, but whether or not a skewed degree distribution can induce finite degree-degree correlations are harder to say (since there is no known way of make an unbiased sampling the set of graphs of a given degree distribution). However, we note that some network models with skewed degree distributions produce networks of zero or positive assortative mixing (Newman, 2002).

Table 1. Assortative mixing coefficients, $r$, for the five pussokram.com networks, and nioki.com and arXiv.org networks. Statistics for corresponding randomized networks is underlined. Differences between the various mixing coefficients are discussed in the text. Double hyphens indicate missing data.

<table>
<thead>
<tr>
<th>network</th>
<th>$N$</th>
<th>$r$</th>
<th>$r_{dir}$</th>
<th>$r_{in,in}$</th>
<th>$r_{out,in}$</th>
<th>$r_{in,out}$</th>
<th>$r_{out,out}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>all contacts</td>
<td>24 349</td>
<td>-0.062</td>
<td>-0.078</td>
<td>-0.082</td>
<td>-0.052</td>
<td>-0.092</td>
<td>-0.057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.039(2)</td>
<td>-0.0303(3)</td>
<td>-0.0253(5)</td>
<td>-0.0172(5)</td>
<td>-0.0460(3)</td>
<td>-0.029(3)</td>
</tr>
<tr>
<td>messages</td>
<td>21 545</td>
<td>-0.055</td>
<td>--</td>
<td>-0.054</td>
<td>-0.056</td>
<td>-0.076</td>
<td>-0.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.05286(3)</td>
<td>--</td>
<td>-0.0143(4)</td>
<td>-0.0120(7)</td>
<td>-0.05833(8)</td>
<td>-0.0574(2)</td>
</tr>
<tr>
<td>guest book</td>
<td>20 691</td>
<td>-0.073</td>
<td>-0.085</td>
<td>-0.097</td>
<td>-0.043</td>
<td>-0.089</td>
<td>-0.053</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0387(4)</td>
<td>-0.0384(6)</td>
<td>-0.024(1)</td>
<td>-0.0155(8)</td>
<td>-0.0419(6)</td>
<td>-0.0262(5)</td>
</tr>
<tr>
<td>friends</td>
<td>14 278</td>
<td>-0.042</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0255(6)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>flirts</td>
<td>8 186</td>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.086</td>
<td>-0.022</td>
<td>-0.012</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.12014(4)</td>
<td>-0.1016(4)</td>
<td>0.008(4)</td>
<td>-0.003(3)</td>
<td>-0.1018(4)</td>
<td>0.014(4)</td>
</tr>
<tr>
<td>nioki.com</td>
<td>50 259</td>
<td>-0.13</td>
<td>-0.10</td>
<td>-0.088</td>
<td>-0.084</td>
<td>-0.10</td>
<td>-0.095</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0339(3)</td>
<td>-0.0145(3)</td>
<td>-0.0184(3)</td>
<td>-0.0141(3)</td>
<td>-0.0207(3)</td>
<td>-0.0161(3)</td>
</tr>
<tr>
<td>arXiv.org</td>
<td>52 909</td>
<td>0.36</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Next we turn to the comparison of the values of real and randomized networks in Table 1. We emphasize that a finite number of iterations of the rewiring procedure of (Shen-Orr, Milo, et al., 2002) (as described above) is not an unbiased sampling of all graphs of a given degree-sequence;
rather that it removes much of the effects of other structures (than the degree-sequence) in the graph. We also stress that the it is hard to estimate the error in the measurement on the actual pussokram.com network data (if the community was reconstructed and the quantities remeasured the outcome would differ). That much said we can still draw some conclusion from the Table 1: Most of the pussokram.com and nioki.com networks are quite close to the values for corresponding randomized networks. However, there seems to be a small but significant deviation towards smaller negative $r$ values for these networks. Nevertheless, we believe that the negative assortative mixing coefficient is likely to stem from the degree distribution. The scientific co-authorship network is in complete contrast to the Internet community networks—here there is a strong assortative mixing in the real network, but the randomized networks show (just as the Internet community networks) a slight negative value. As mentioned above, this can most probably be explained by the strategy and customs of scientific collaborations.

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 $r_{\text{in\_out}}$ to be the same as e.g. $r_{\text{in\_in}}$, as a large $r_{\text{in\_out}}$ means that actors that are active in the community (have a high $k_{\text{out}}$) tend to associate with those who are successful in promoting themselves in the community (have a high $k_{\text{in}}$), 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 $r_{\text{dir}}$ (the time development of the other assortative mixing coefficients of Table 1 is qualitatively similar). We see that $r_{\text{dir}}$ 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 $t = 300$ 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.)

Cumulative degree distributions are highly skewed

The degree distribution has received much attention in comparative analyses of complex network since the work of Barabási and Albert (Barabási and Albert, 1999). A skewed degree distribution is commonly regarded as a cumulative effect in the attachment of new arcs to the network (Barabási and Albert, 1999; Simon, 1955), and it offers a way to classify different types of networks (Amaral, Scala, 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-freeness. 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, Edling, 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 $-\gamma$ then the cumulative distribution will have the exponent $-\alpha = -\gamma + 1$. All pussokram.com networks are highly skewed, but none of them perfectly fits a power-law form. We would prefer to classify them as stretched exponential distributions. 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 conclude that the dynamics shaping the degree-distribution is to a large extent the same for the two communities.

---

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

---

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

$$l_{GC} = \frac{1}{|A_{GC}|} \sum_{(u,v) \in A_{GC}} d(u,v),$$  \hspace{1cm} (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),

$$l^{-1} = \frac{1}{M} \sum_{(u,v) \in M} \frac{1}{d(u,v)},$$  \hspace{1cm} (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 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 data set, 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, Jeong, 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).

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

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 person B and person C, then person B and person C are likely to be acquainted. This is commonly measured with Watts and Strogatz’ clustering coefficient (Watts and Strogatz, 1998; Newman, Watts, et al., 2002), which gives the fraction of triangles out of the connected 3-paths of the graph (a quantity that was defined for bi-directional
graphs, but is trivially generalized to directed graphs, for which we use subscript “dir”). If we let \( p(n) \) denote the number of paths, and \( c(n) \) denote the number of circuits, of length \( n \), then we can express the clustering coefficient, \( C \), as:

\[
C = \frac{c(3)}{p(3)}. \tag{5}
\]

| Table 2. | Statistics for the fully-grown networks of pussokram.com, nioki.com and arXiv.org networks provided for comparison. Statistics for corresponding randomized networks is underlined. Double hyphens indicate missing data. |
|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| network | all contacts | messages | guest book | friends | flirts | nioki.com | arXiv.org |
| N       | 24 349       | 21 545   | 20 691     | 14 278  | 8 186   | 50 259     | 52 909     |
| M       | 101 313      | 76 257   | 73 346     | 31 871  | 8 744   | 405 742    | 490 600    |
| R       | 0.34         | 0.45     | 0.33       | 1       | 0       | 0.69       | 1          |
| \( l_{GC} \) | 4.4     | 4.3      | 4.6        | 5.1     | 5.7     | 4.1        | 6.1        |
| \( \Gamma^3 \) | 0.12   | 0.10     | 0.084      | 0.18    | 4.0 \times 10^{-4} | 0.209      | 0.121      |
| \( C \) | 0.023         | 0.0012   | 0.014      | 0.020   | 0.00057 | 0.0065     | 0.45       |
| \( 0.00887(5) \) | \( 0.00262(2) \) | \( 0.00879(8) \) | \( 0.00470(9) \) | \( 0.000289(9) \) | \( 0.00806(2) \) | \( 0.00203(1) \) |
| \( C_{dir} \) | 0.015        | 0.0052   | 0.014      | --      | 0       | 0.0076     | --         |
| \( 0.00887(5) \) | \( 0.00348(3) \) | \( 0.00541(5) \) | --      | 0       | 0.00767(2) | --         |
| \( D \) | 0.024         | 0.0061   | 0.021      | 0.020   | 0.088   | 0.013      | 0.35       |
| \( 0.01015(5) \) | \( 0.00262(2) \) | \( 0.01008(8) \) | \( 0.00539(8) \) | \( 0.0177(4) \) | \( 0.00812(2) \) | \( 0.00209(6) \) |
| \( D_{dir} \) | 0.012        | 0.0081   | 0.015      | --      | 0       | 0.016      | --         |
| \( 0.00548(3) \) | \( 0.00342(3) \) | \( 0.00536(5) \) | --      | 0       | 0.00766(2) | --         |

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

\[
D = \frac{c(4)}{p(4)}. \tag{6}
\]

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 the for a paper with \( n_{auth} \geq 3 \) authors represents a fully connected subgraph of \( G \) (contributing with \( n_{auth} (n_{auth} - 1) (n_{auth} - 2) / 3 \) triangles). However, we would like to stress that the values themselves is 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 dependent on only the local network structure the density of short circuits is an intrinsic quan-

---

\(^3\) Presumably, a same-sex relationship is not a typical type of romantic relationship among Swedish heterosexual adolescents. Therefore we expect few triangles. 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.
tity; and, as seen for the clustering coefficient (Barabási, Jeong, et al., 2002), these quantities approaches 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. 5, 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; and 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, Jeong, et al., 2002; Barabási and Albert, 1999; Barabási, Albert, et al., 1999) 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; Holme and Kim, 2002) is likely not the mechanism responsible for the finite clustering (remember that in network models such as the Erdös-Rényi (Erdös and Rényi, 1959) and Barabási-Albert (Barabási and Albert, 1999; Barabási, Albert, et al., 1999;
Barabási, Jeong, et al., 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.  

**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 condition 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 about the Internet community pussokram.com, we see that:

- The average degree converges over time. Previous studies do suggest that there is an upper limit to the number of contacts that a person can have (Marsden, 1987), but it is interesting that we find this socio-cognitive limitation despite the fact that time and space is of less important here.
- 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 disassortative mixing. The main explanation for these $r$-values are probably the skewed degree sequence. We also find that mixing coefficients as a function of time converge rapidly. The disassortative mixing in the Internet community networks is in striking contrast to the strong assortative mixing seen in scientific collaboration networks.
- The cumulative degree distributions are highly skewed, lying somewhere in between previous mappings of acquaintance networks (Amaral, Scala, et al., 2000) and partnership networks (Liljeros, Edling, et al., 2001).
- 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 $l-1$ shows 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 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 twofold—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

---

4 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.
in that the density of 4-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 4-circuits.

We conclude by once again establishing that 
an Internet communities 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 and Stefan Praszalowicz at nioki.com for granting us access to data, Niklas Angemyr for helping out with data manipulation, Mark Newman for assortative-mixing comments, and Reginald Smith for sharing contacts. PH is partially supported by the Swedish Natural Research Council (F 5102-659/2001). CRE is supported by the Bank of Sweden Centenary Foundation. FL is supported by the National Institute of Public Health.

References

“e-print arXiv:” refers (yet unpublished) to manuscripts uploaded to the database arXiv.org.


