10-2008

Comparison of online social relations in volume vs interaction: A case study of Cyworld

Hyunwoo CHUN

Haewoon KWAK
Singapore Management University, hkwak@smu.edu.sg

Young-Ho EOM

Yong-Yeol AHN

Sue MOON

See next page for additional authors

Comparison of Online Social Relations in Terms of Volume vs. Interaction: A Case Study of Cyworld

Hyunwoo Chun
Dept. of Computer Science
KAIST, Daejeon, Korea
hyunwoo@an.kaist.ac.kr

Haewoon Kwak
Dept. of Computer Science
KAIST, Daejeon, Korea
haewoon@an.kaist.ac.kr

Young-Ho Eom
Dept. of Physics
KAIST, Daejeon, Korea
thinking22@gmail.com

Yong-Yeol Ahn*
Center for Complex Network Research
Boston, U.S.A.
yongyeol@gmail.com

Sue Moon
Dept. of Computer Science
KAIST, Daejeon, Korea
sbmoon@kaist.edu

Hawoong Jeong
Dept. of Physics
KAIST, Daejeon, Korea
hjeong@kaist.ac.kr

ABSTRACT

Online social networking services are among the most popular Internet services according to Alexa.com and have become a key feature in many Internet services. Users interact through various features of online social networking services: making friend relationships, sharing their photos, and writing comments. These friend relationships are expected to become a key to many other features in web services, such as recommendation engines, security measures, online search, and personalization issues. However, we have very limited knowledge on how much interaction actually takes place over friend relationships declared online. A friend relationship only marks the beginning of online interaction.

Does the interaction between users follow the declaration of friend relationship? Does a user interact evenly or lopsidedly with friends? We venture to answer these questions in this work. We construct a network from comments written in guestbooks. A node represents a user and a directed edge a comments from a user to another. We call this network an *activity network*. Previous work on activity networks include phone-call networks [34, 35] and MSN messenger networks [27]. To our best knowledge, this is the first attempt to compare the explicit friend relationship network and implicit activity network.

We have analyzed structural characteristics of the activity network and compared them with the friends network. Though the activity network is weighted and directed, its structure is similar to the friend relationship network. We report that the in-degree and out-degree distributions are close to each other and the social interaction through the guestbook is highly reciprocated. When we consider only those links in the activity network that are reciprocated, the degree correlation distribution exhibits much more pronounced assortativity than the friends network and places it close to known social networks. The k-core analysis gives yet another corroborating evidence that the friends network deviates from the known social network and has an unusually large number of highly connected cores.

We have delved into the weighted and directed nature of the activity network, and investigated the reciprocity, disparity, and network motifs. We also have observed that peer pressure to stay active online stops building up beyond a certain number of friends.

The activity network has shown topological characteristics similar to the friends network, but thanks to its directed and weighted nature, it has allowed us more in-depth analysis of user interaction.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Human Factors, Measurement

Keywords

Online social network, Cyworld, Friend relationship, Guestbook log, Degree distribution, Clustering coefficient, Degree correlation, K-core, Reciprocity, Disparity, Network motif

1. INTRODUCTION

Online social networking services are among the most popular Internet services according to Alexa.com and have become a key feature in many Internet services. Not only online social networking services (*e.g.*, Myspace and Facebook) but also other major web 2.0 services (*e.g.*, Flickr, Del.icio.us, and YouTube) offer social networking features on their sites. Through various features of online social networking services, users establish friend relationships, share photos, and write short messages. The friend relationship lays the foundation for other systems to build upon for recommendation engines, cooperation-based security, online search, and other personalization functions. Understanding friend relationships is the first step towards achieving them.

A friend relationship is an explicit and static declaration of a relationship. For actual interaction between users, the friend relationship may not be the best representation. The friend relationship only marks a beginning of online interaction. Activities, such as looking at friends' photos, reading their articles, and leaving comments on their guestbooks, follow once the friend relationship is established. Macroscopically, the number of users, the number of

*This work was conducted while Ahn was at KAIST.

IMC'08, October 20–22, 2008, Vouliagmeni, Greece.

daily visitors, and page views are the three most basic metrics to measure the status of online social networking services (OSNSs) [1]. These metrics compose an overall view of livelihood of an online social network service itself, but they do not provide any information about the livelihood of interaction between users.

In this paper, we shift the focus regarding analysis of online social networks from a friends network to an *activity network* for better understanding of online social networks. We construct an activity network from logs of actual interaction rather than from declared relationships. The two main sets of questions we raise in this work are:

- Does the friend relationship reflect underlying user interaction? Does a user interact only with one's friends or explore the social network more widely? If the social interaction does not follow the friends network closely or evenly, tracking user interaction should become a core design feature in any service site.

- How does information flow through the network? Do all users receive the same attention from their friends? How often do they interact? Is the interaction one-way or reciprocated?

We take a top-down approach and begin our analysis with network growth. The *activity network* built for the present work is directed and weighted. The direction represents the flow of interaction and the weight the amount of interaction. We first look at the numbers of users in the friend and activity networks and compare their growth over time. We then compare the topological characteristics–namely, the degree distribution, the clustering coefficient, and the degree correlation–of the two networks. We use reciprocity, disparity, and network motifs to investigate the activity network's unique characteristics in the form of a weighted and directed graph.

For our work, we use more than two years of guestbook logs from the largest online social networking site in Korea and build a graph from the comments recorded in these logs. The friends network is a complete set of friend relationships. Access to this data set allows us unique opportunities otherwise not possible, as the friend lists of some users are often kept private and data collected by crawling contains unavoidable bias.

Previous work on activity networks includes phone-call networks [34, 35] and MSN messenger networks [27]. Online social networks are unique in that they have this reference network of friends. To the best of our knowledge, this is the first attempt to compare the *explicit* friend relationship network and *implicit* activity network.

The remainder of this paper is structured as follows. In Section 2 we describe our guestbook logs and the features specific to our logs. In Section 3 we compare the topological characteristics between the friend and activity networks. We then delve further into the weighted and directed aspects of the activity network in Section 4 and other activity-related aspects In Section 5. In Section 6 we compile related work and Section 7 concludes with a discussion on future work.

2. ACTIVITY IN GUESTBOOK

In this section, we describe the social network data we use for this study. Cyworld, launched in 2001, is the largest online social network service in Korea. As of October 2007, the number of registered Cyworld users has surpassed 20 million, which is more than a third of the entire South Korea population[2].

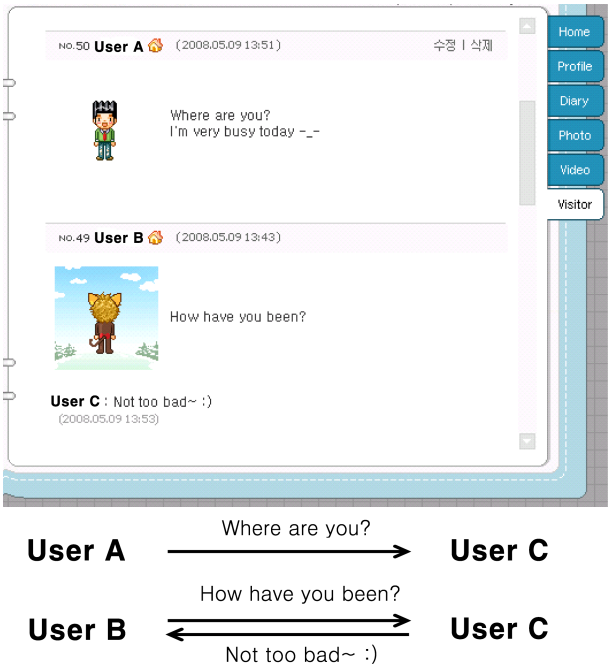


Figure 1: Screen capture of user C's guestbook

When a user joins Cyworld, one is given a homepage (called *minihompy*) that contains an avatar, a photo gallery, a public diary, a testimonial board, a guestbook, etc. A user can establish *friend* relationships with other users and share information only with those established relationships. Users browse through friends' photos and leave comments. They read others' public diaries and write testimonials for those established friends. Some of the features, such as writing a testimonial and viewing photos, are often limited to only those with established online friend relationships. The owner of the minihompy can choose the buttons or features on one's minihompy. Some features, such as the profile and the diary, are read-only, while access to other features are owner-configurable except for the guestbook. Once the owner includes the guestbook on the minihompy, then it is open to anyone to write, a friend or not. Even a person not registered as a Cyworld user can still visit and write a comment on a guestbook. The photo gallery and the bulletin board can be configured to be writable by visitors, but many users keep the default setting of write-by-owner-only. The guestbook is the most used feature in Cyworld where friends and visitors leave a note of greetings to the minihompy owner[3]. We include a screen capture of a typical interaction on a Cyworld guestbook in Figure 1. A comment writer's name and avatar are displayed along with the comment.

Ahn *et al.* have analyzed Cyworld's topological characteristics of bi-directional friend relationships [2]. Once established, a friend

[1]Ranking sites, such as Alexa, Rankey, and Ranking, use these metrics in their web site rankings.

[2]Upon joining, a new user must have its personal identification number (equivalent of U.S.'s social security number) verified. Foreigners have special provisions for membership. All user accounts on Cyworld map to real users, unless a user make an illicit use of other people's personal identification numbers.

[3]We were offered logs of comments on the photo gallery and the bulletin board of the same period, but they were far smaller than guestbook logs.

relationship remains rigid regardless of the actual relationship [40]. It is an assertion that some relationship existed, currently active or not. In this work, we delve deeper into the web of social networking and study the user interaction captured in the guestbook. Unlike a friend relationship, which is bi-directional, a message on a guestbook represents a directional interaction between users. On a guestbook, people write greetings, recent updates, replies, and so on.

We have obtained the complete guestbook logs of Cyworld from June 2003 to October 2005[4]. This period is very important in the development of Cyworld, as the number of users grew exponentially from 2 million to 16 million and the friend relationship network began to show a sign of densification [2]. In this work we investigate whether the growth in actual user interaction, a key aspect of social networking services, has kept up with the growth in sheer size. Our guestbook log consists of three-tuples: the writer, the guestbook owner, and the time of the guestbook comment. All user identifiers have been anonymized. As of October 2005, the number of Cyworld subscribers reached $16,146,817$. Among those 16 million users, 74.6% or $12,048,186$ users have formed friend relationships with others, and 64.8% or $10,476,604$ users have written or received a comment on a guestbook at least once during the period of our guestbook logs. Compared to $381,602,530$ friend relationships, the number of the writer and guestbook owner pairs is larger: $537,970,431$. Table 1 summarizes our dataset. The numbers in the parentheses exclude messages written by the owner of the guestbook on one's own guestbook. We explain more about this type of messages in Section 2.2.

Table 1: Summary of Cyworld Guestbook Logs

Period	2003.06∼2005.10
# of 3-Tuples	8,423,218,770
# of unique writer-owner pairs	537,970,431
# of guestbook users	17,788,870
Mean # of msg per writer	637 (397)
Mean # of msg received per owner	484 (297)

2.1 Growth in Guestbook Activity

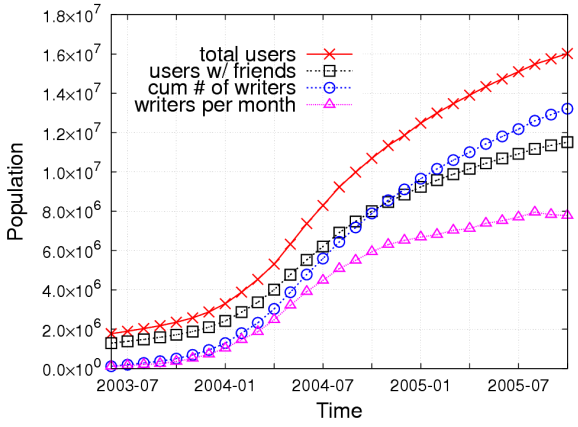


Figure 2: Cyworld growth in numbers

[4]The period of guestbook logs does not match that of the friends network in [2]. We could not retrieve the friends network from the same period as the guestbook logs.

As the number of Cyworld subscribers grew almost ten times between 2003 and 2005, its guestbook had also seen explosive growth in activity. We plot the number of Cyworld subscribers and the relevant statistics in Figure 2[5]. The top graph represents the total number of subscribers. The next two graphs crisscross each other in about October 2004. The graph marked with a square represents the cumulative number of guestbook writers, and that marked with a circle the number of users with friend relationships. The former could be larger than the latter, because the guestbook is open to anyone. Even if a person has not established a friend relationship with the guestbook owner or is not even a registered user of Cyworld, one can still write on a guestbook. The bottom graph represents the number of guestbook writers in that month. The number of guestbook users was very small at the beginning, but caught up with the total number of Cyworld users fast. It surpassed the number of users with friend relationships, attesting that it is the most used feature.

However, the monthly statistics of guestbook users started to abate in growth. Here we observe a hint of slow-down in Cyworld growth. The slow-down tendency is also observed in the growth rate of the number of guestbooks and messages per writer. Figure 3 shows the total number of guestbook comments and the number of user pairs against the number of guestbooks users.

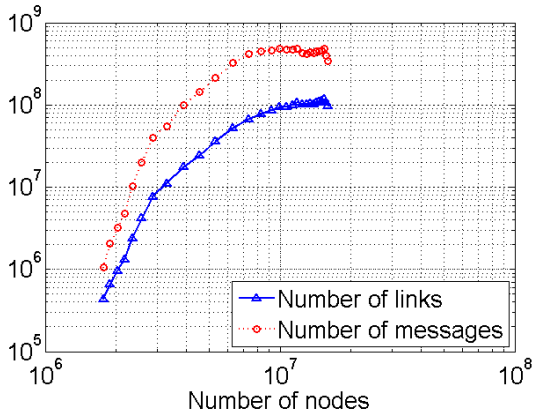


Figure 3: Growth in the numbers of guestbooks and messages versus the number of users

The total number of guestbooks that users have written does not increase very fast after the number of users exceeds 10 million. No social network can sustain an explosive growth forever, and its growth rate must slow down at some point. As Cyworld is limited to Korean-speaking populace of 70 million to 100 million[6], the slow-down in growth in around July 2004 or at about 8 million is markedly interesting. We do not have data from other social networking services, but take a mental note that at about 10% of the target market size the growth slows down. Similar slow-down in growth has been observed in bulletin board systems (BBS) of a university as well [17]. In this network, BBS users are connected through message posting like leaving comments in the guestbook. The number of users in BBSs grow exponentially, but the growth rate eventually drops below an exponential rate. The total number of links and the total weight of BBS networks also grow exponen-

[5]The total number of friends and the number of users with friends in this figure are a courtesy of SK Communications, Inc.

[6]Cyworld has opened service in China, Japan, Taiwan, and USA. Each service runs independently and the user based is not shared.

tially at the beginning, and then their growth rates slow down similarly [15]. The starting point of inevitable slow-down in growth is of interest to online social networking service (OSNS) providers, as it marks a transition from a fast-growing phase to a steady growth. Our observation is just one exemplary data point, and we leave the correlation between the point of transition and the expected population of the service for future work.

2.2 Self-Posting in Guestbook

When a friend writes a message on a guestbook, the owner of the guestbook often replies in one's own guestbook, instead of visiting the friend's guestbook and writing there. This activity is captured in our guestbook log as a 3-tuple that has the same writer and owner. We call this tuple a self-post. Self-posts take up about a third or 38.9% in all posts, and they are evenly distributed over time. Also 81.8% of users who have written at least once have written a self-post. For half of the users, a third of messages they wrote are self-posts. Self-posts make up a non-negligible portion and we should determine how to interpret self-posts before analyzing user activities of guestbook logs.

A self-post serves either of the two purposes: a message for viewing by all others (a notice) or a reply specifically for a preceding message. We cannot distinguish a notice from a reply in the guestbook log, as they both appear as 3-tuples with the same writer and owner. As Cyworld offers two other features, the bulletin board and the public diary, that both serve a similar purpose for notices and announcements, we assume most self-posts are replies for this work.

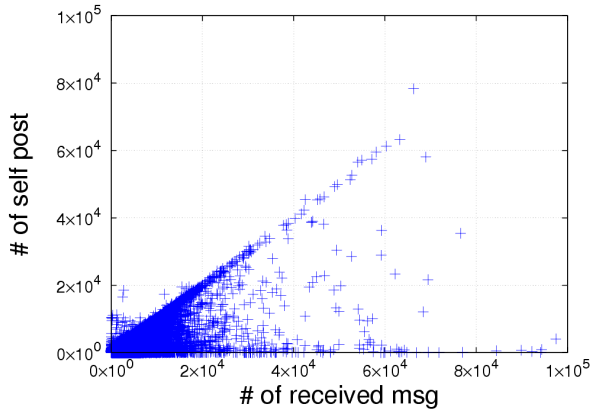


Figure 4: Self-posts vs messages received

In Figure 4 we plot the number of self-posts against the number of messages received per user. There is a strong positive correlation between the two numbers; the Pearson correlation coefficient between the two numbers is 0.8201. Most points lie below the line, $y = x$, and about 95.1% of users' self-posts are smaller than their received messages. We see a small number of points above $y = x$ in the left bottom corner. Our guestbook logs only include comments between registered users of Cyworld and do not contain comments by non-Cyworld users. Non-Cyworld users can browse minihompies and write on guestbooks, as long as owners of the guestbooks allow it. Non-Cyworld users do not have a user id, and their activity is not logged in our data. This explains those points above $y = x$ in the left bottom corner.

From above, we conclude that the self-posts represent reciprocal activity, but face a dilemma because we cannot disambiguate the actual recipient. For example, a minihomy owner has received messages from users i and j, and writes one self-post a few days later. Is the self-post meant for user i, user j, or both? We cannot tell from the data we have. However, self-posts are an important aspect of user activity, and we cannot drop them completely in our analysis. In the rest of the paper, we make it explicit whether we include self-posts in the analysis or not.

2.3 Activity Network

Graph representation of a social network is an apt abstraction of their connected nature and allows us to tap into the rich repository of graph and complex network theories. In this section we describe how we represent the user interaction on the guestbook as a graph and define metrics of interaction.

In a network of nodes and edges without directions, a node degree refers to the total number of edges. For the guestbook activity, we construct a network with weighted and directed edges. We map a user to a node and a message to a directed edge from a writer to a reader (we refer to a user and a node interchangeably). An edge from node i to node j denotes that user i has written a message on user j's guestbook. The weight, w_{ij}, of a directed edge from i to j is the number of messages user i has written to user j. A node in a directed network has two degrees: an in-degree and an out-degree. We often refer to an out-degree in a directed network as the degree, and specify in-degrees when necessary.

In a weighted network, a node strength represents the sum of all weights of outgoing edges. The strength of node i with out-degree k is defined as: $s_i = \Sigma_{j=1}^{k} w_{ij}$.

We call a weighted and directed network constructed from the guestbook log the *activity network*. Note that the nodes of the activity network is not a proper subset of that of the friends network, for users without friend relationships can still write onto one's guestbook.

Self-posts map to a reflexive edge pointing back at the originating node itself, and the weight is the number of self-posts. It is reasonable to include self-posts in the strength, as self-posts are meant for other users.

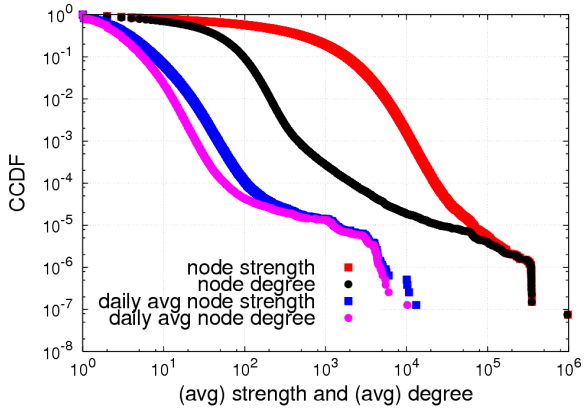


Figure 5: CCDF of strength and degrees of the activity network

In Figure 5 we plot four complementary cumulative distribution functions (CCDFs) of the strengths and out-degrees of the activity network; two of them are daily averages of the strength and the out-degree.

As our guestbook data is from the period of explosive growth, a large number of users have joined and the time of membership

initiation should be taken into consideration. Not all users write on a guestbook as soon as they join Cyworld and a gap exists between the times of a membership initiation and the first guestbook activity. Hence, daily averages of node strength and degree are calculated by taking the cumulative node strength and degree by October 2005 and dividing them by the number of days since the first time a user has written a message during our dataset period.

The plots in Figure 5 all have conspicuous straight drops near the end. As the weight represents the total number of comments on a guestbook, those drops mark the physical limit of a Cyworld users' activity. The strength and daily average strength drop at around 980,000 and 4,000, respectively. If a user took 10 seconds for each comment, then the daily average strength of $4,000$ means that the user spent more than 10 hours a day! This is a remarkable feat unless we suspect them of using an automated script. We see a need to investigate those users who seem to have hit the physical limit in more detail.

All Cyworld users have to register with one's national identification number (like the social security number in the US). Foreigners have a similar provision but with a different identification. At this point we have no basis to consider rampant illicit use of identity. However, occasional appearances of spammers are possible, and they must have used an automated script. Even though automated scripts are banned in Cyworld, a limited number of users have taken the liberty to use them. The site has kept up with catching those abusers, but the list is never complete. Guestbook comments entered by automated scripts are most likely to be close in time, and the number of daily comments to surpass a reasonable number expected from manual input. However, we are aware of legitimate users whose use of automated scripts is condoned by the site.

The goal of this work is to conduct a macroscopic comparison of friend relationships and underlying social interaction. As we have seen in [2], cyber-only relationships are common and they often lead the growth of the service site. Thus we include all comments, suspected of automated scripts or not, in this study and leave the separation of scripted comments for future work.

3. STRUCTURE OF ACTIVITY NETWORK

In this section, we compare the topological characteristics of the friend relationship and activity networks to analyze the basic structural characteristics of the network, and to observe how the activity network differs from the friend relationship network in structural aspects. We analyze three fundamental metrics: degree distribution, the clustering coefficient distribution, the degree correlation. We include the plots of the friend relationship network of Cyworld as a reference (as of November 2005) [2].

3.1 Degree

First, we plot the degree distribution of the activity network and the friend relationship network. The degree distribution of the friends network has revealed that the Cyworld friends network has two different scaling regions in a power-law distribution. The second region in the power-law distribution attests to existence of users with very high node degrees, higher than predicted by a power-law distribution. Ahn *et al.* have attributed this to cyber-only relationships. It would be interesting to see if a similar pattern emerges in the activity network.

If we take into account the direction of the edges in the network, the degree of a person can be defined in several ways since there are both incoming and outgoing edges. We plot three distinct types of degree distributions: out-degree, in-degree, and bi-directional degree. One's out-degree is the number of guestbooks one has written on, the in-degree is the number of users who have visited and written on one's guestbook, and the bi-directional degree represents the number of people who have reciprocally interacted through the guestbook. Note that each kind of degree represents a distinctive perspective; one's out-degree is one's own activity, while the in-degree is the collection of all other people's activity on the person. The bi-directional degree is the number of people who reciprocally interact with the person and filters out one-sided interaction.

In Figure 6, the in-degree distribution and the out-degree distribution look similar. It is possible that it is just a coincidence, *i.e.* the one's own activity distribution (out-degree) coincides with the popularity distribution (in-degree). However, there is a much more persuasive explanation that this similarity is the manifestation of strong underlying reciprocity of the communication. The bi-directional degree distribution also supports this explanation. It still shows two scaling regions, the second of which stretches over $10,000$. There are apparently a fair number of people who interact reciprocally with more than hundreds or thousands of people. This finding supports that the activity is highly reciprocal. We will delve into the reciprocity in the Section 4.1.

On the other hand, the in-degree and out-degree distributions have small but notable differences. The out-degree distribution exhibits sharp cut-off between 10^5 and 10^6, while the in-degree distribution shows a smooth tail. It is consistent with our previous observation that the node strength has an upper bound due to physical limits, while the in-degree has none.

3.2 Clustering

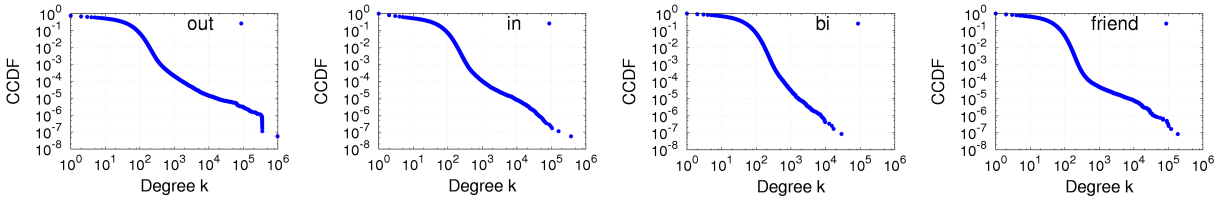
In a network with only unweighted and undirected edges, the clustering coefficient of a person, which represents how closely one's friends are connected, is defined by the ratio of the actual number of connections over all possible connections between one's friends. The clustering coefficient of a network is defined by the average of individual clustering coefficients. Barrat *et al.* has proposed a slightly modified definition for a weighted network [7]. The definition of a weighted clustering coefficient for node i is:

$$c_i^w = \frac{1}{s_i(k_i-1)} \sum_{j,h} \frac{(w_{ij}+w_{ih})}{2} a_{ij}a_{ih}a_{jh} \qquad (1)$$

where $k_i = \Sigma_j a_{ij}$ and $a_{ij} = 1$ if there exists an edge from node i to node j; $a_{ij} = 0$, otherwise. In the case of the friends network, $(w_{ij} + w_{ih})/2 = 1$ and $s_i = k_i$.

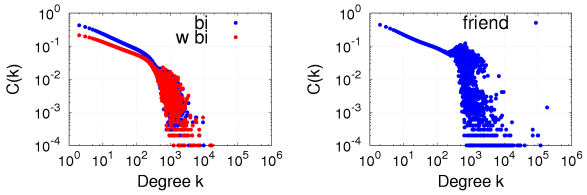
We calculate the clustering coefficient of the activity network with only bi-directional edges. We compare the clustering coefficient for the following two cases: first, without the weight of edges (that is, as if the network is unweighted), and second, with the bi-directional edges.

The clustering coefficient distribution, $C(k)$, is a mapping of the mean clustering coefficient of all nodes with degree k to k. The plots of $C(k)$ is in Figure 7. The mean clustering coefficient of all nodes, C, is 0.1665. We denote the mean weighted clustering coefficients of all nodes and of nodes with degree k as C^w and $C^w(k)$, respectively. Figure 7 includes $C(k)$ of the bi-directional network and $C^w(k)$ of the weighted bi-directional network. If $C^w(k) > C(k)$, then the edges with larger weights are more likely connected. If $C^w(k) < C(k)$, then network topology owes more to the lightly weighted edges than to those with large weights [7]. In Figure 7(a) we see that $C^w(k) < C(k)$ for the most part. The bi-directional weighted activity network has C^w of 0.0965, which is smaller than $C = 0.1665$. To translate this to our context with the activity network, much of the user interaction (here, weights) is carried over edges not belonging to triangularly clustered connections (a.k.a. completely mutual triads). We revisit this issue of microscopic clustering in Section 4.3.



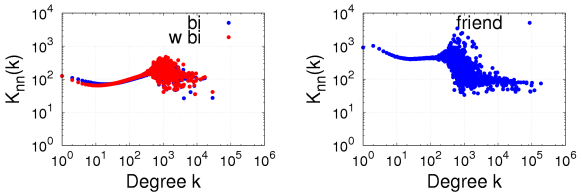
(a) out-degree (b) in-degree (c) bi-directional (d) friend

Figure 6: Degree distributions



(a) bi- and weighted (b) friend

Figure 7: Clustering coefficients



(a) bi-and weighted (b) friend

Figure 8: Degree correlations

3.3 Degree correlation

The degree correlation shows the individual's propensity to connect to nodes with similar degrees. The term *assortative mixing* stands for the situation that a person is likely to be connected to other people with similar degrees, and the term *disassortative mixing* stands for the opposite. It is known that most real-world networks exhibit the disassortative mixing, while the human social networks exhibit the assortative mixing. The mixing pattern can be quantified by the *assortativity*, which is defined as follows [31,32]:

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}, \quad (2)$$

where M is the number of links and j_i and k_i are the degrees of two nodes connected to edge i.

Besides the assortativity, a plot of degree correlation discloses more details. The degree correlation is the ratio of the degree of a user over the mean degree of one's neighbors, and we denote it as $k_{nn,i}$ for node i. Normally, the quantity $k_{nn}(k)$ is calculated by averaging the degree correlations of all the nodes of degree k. The degree correlation is shown in the Figure 8.

The Cyworld friends network exhibits the assortativity value of -0.13, but a close inspection on the degree correlation has shown complex behaviors [2]. We report similarly complex behaviors in degree correlation from the activity network. The assortativity of the activity network is 0.0089. The bi-directional activity network still has a glut of points spread out beyond the degree of 500 or above, but the clear assortative mixing pattern between $k = 30$ to 500 shows up. The slight disassortative mixing or a negative trend of $k_{nn}(k)$ for $k < 30$ is consistently observed in both Figures 8(a) and (b). It is due to highly asymmetric connection between users of very small degrees and those of very large degrees. Even when only a small fraction of people with degree 1 is connected to a user with degree $100,000$, they more or less act as outliers in the calculation and have a big impact on the average, $k_{nn}(1)$. One plausible explanation for such an asymmetric relation is fans of a celebrity.

It is also possible to modify the definition of the degree correlation to take into account the edge weight [7]. The weighted degree correlation of node i is:

$$k^w_{nn,i} = \frac{1}{s_i}\sum_{j=1} N a_{ij} w_{ij} k_j \quad (3)$$

The weighted degree correlation is plotted along the degree correlation in Figure 8(a). It dips slightly under the plot of the bi-directional activity network for $k < 30$, but lines up almost right on top of it for $k \leq 30$. We conclude that the the interaction pattern (weight) are not affected much by the degree of friends.

3.4 K-core

The assortativity captures the "birds of a feather" phenomenon, but says little about grouping amongst users. We use *k-cores* to see how strongly connected people of similar degrees are. A k-core is a subgraph in which all nodes have at least k neighbor nodes. It identifies strongly connected components of similar degrees in a network. Note that a k-core is different from a set of nodes whose degree is k in that all the nodes in a k-core are connected.

A network partitions into multiple k-cores, if $k \geq 1$. We plot the number of nodes in all the k-cores of two networks: the bi-directional activity network and the friends network in Figure 9. As in the experiment by Leskovec *et al.* [27], we observe a sharp decrease in the number of nodes at certain values of k. In each network, the number of nodes included in k-cores decreases steadily until $k = 34$ (bi-directional) and $k = 38$ (friend) and drops quickly over an order of magnitude in just one step. The bi-directional activity and friends networks exhibit a very similar behavior until the transition points. However, after the transition, the number of nodes in k-cores in the friends network decreases much slower than in the bi-directional activity network. This result indicates that there are excess edges that are not active but play an important role in holding the core people together. Grouping in a social network offers much insight into the operational structure of the society. We leave more detailed analysis of OSNS grouping for future work.

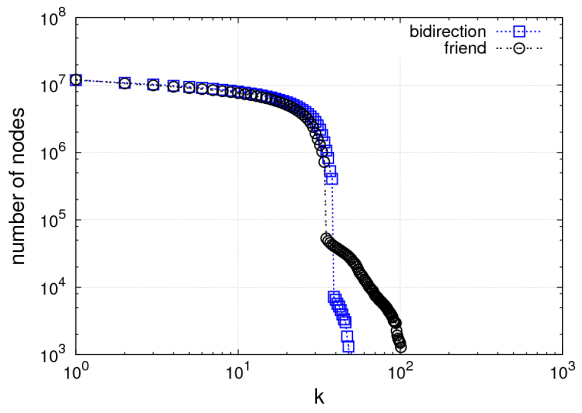


Figure 9: K-core

3.5 Summary

In this section we have analyzed the fundamental structural metrics of the activity network and compare them with the friends network. Though the activity network is weighted and directed, its structure is similar to the friend relationship network. We report that the in-degree and out-degree distributions are close to each other and the social interaction through the guestbook is highly reciprocated. When we consider only those links in the activity network that are reciprocated, the degree correlation distribution exhibits much more pronounced assortativity than the friends network. This sets the activity network apart from the friends network and places it close to known social networks. The k-core analysis gives yet another corroborating evidence that the friends network deviates from the known social network and has an unusually large number of highly connected cores.

4. ACTIVITY NETWORK AS A WEIGHTED DIRECTED GRAPH

In Section 3 we have investigated the topological differences between the unequivocal friend relationship and social interaction behind it. In this section we focus solely on the activity network and anatomize the flow of interaction through reciprocity, disparity, and motifs.

Reciprocal interaction is considered as one mechanism to help the evolution of cooperation in society [5, 20, 33] and characterizes the strength of ties [21]. Although gender, religion, age, or culture differences induce a gap in the needs of reciprocal intimacy among people [4], reciprocal interaction "pervades every relation of primitive life" [39] and in all social systems [37]. In this section we look into the reciprocity embedded in the activity network.

As previously stated, we do not include self-posts in our analysis because of difficulties in disambiguating the true recipient of a self-post. Due to the massive nature of the data and the even spread of self-posts across all users, we expect self-posts would not alter the outcome of the analysis much. We leave the self-post disambiguation for future work.

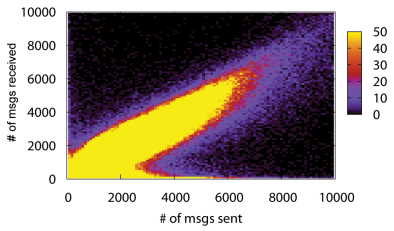
Depending on the type of expected reciprocity, reciprocal interaction can be classified into two models: the *actor-reactor* model and the *actor-receiver* model [22]. In the actor-reactor model, the actor and the reactor compare their relative ranks in reciprocal interaction. That is, if user i considers user j to be most important, user j should also consider user i with a rank of matching importance among other users. We could raise questions, such as "Who are the top three friends that a user interacts with most?" and "Does that ranking match that of the friend's?" The significant drawback of this model is that users are burdened with heavy cognition load, keeping track of the reactor's interaction with other users. Hemelrijk concludes this type of reciprocity is not representative of the utility we experience and get out of daily social interaction [22].

Under the actor-receiver model, "Hey, I visited your guestbook last week and left you a note. When will you get back to me?" is more likely than "Why don't I get your most attention?" A user checks only what one gives and receives. If user i gives value k to user j, then user i only checks the value returned from user j. There lies an inherent assumption in the actor-receiver model that the cost of interaction is comparable. If it is not, then the difference should be factored into the evaluation of reciprocity. In Cyworld, all users are presented with the same interface to a guestbook but for the skin. In this work, we investigate the reciprocity under the actor-receiver model and assume all posts to have an equal cost.

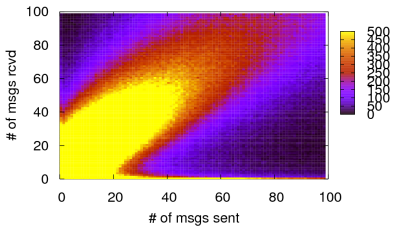
4.1 Reciprocity

As the first step towards looking at reciprocity, we compare the numbers of messages a user received on one's own guestbook and written on others' guestbooks. For clarity we refer to the former as messages received and the latter as messages sent in the rest of the paper. Figure 10(a) plots the number of messages received versus the number of messages sent per user. We have found a scatter plot of the data not easy to read, and used 100 by 100 grids instead. The color of each grid represents the number of data points in the grid region. Data points of value greater than 50 are colored the same as 50. Out of 17.8 million guestbook users, about 10 million or 58% of them are located in the grid bounded by (0,0) and (100, 100). That single data point is two orders of magnitude larger than any other point and makes it hard to plot the rest of the data points and elicit a pattern. We zoom into that single grid and plot it in Figure 10(b). The grid size is 1 by 1 and all points greater than 500 are colored the same as 500. Though in different scales, Figures 10(a) and (b) display similar patterns. Most users are centered along $y = x$. Figure 10(c) plots the median number of messages received versus the number of messages sent per user in log scale and confirms the trend. From the figures, we see three types of people: type (I) with comparable numbers of messages sent and received, type (II) with far more messages sent than received, and type (III) with more messages received than sent. The first and largest type of people appears along the $y = x$ line. People of type (I) are clearly rewarded for the messages sent with reciprocated messages. The other two types present opposite characteristics of two user groups. Type (II) writes far more messages than receives. We conjecture a good part of these users are likely to be spammers or very passionate fanboys. The total number of data points in the grids along the x-axis is $374,443$. The grid at the origin is not included. Type (III) represents those who write only a few replies, but they receive many messages. The total number of data points in the grids along the y-axis is $667,652$, not counting the grid at the origin. They are likely to be very popular people, such as celebrities, but we could not verify due to user anonymization and lack of supplementary data on user profiles.

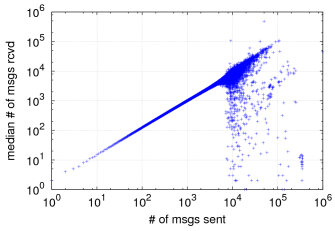
Now in order to investigate pair-wise reciprocity, we plot the number of messages exchanged between a pair of users. Figure 11(a) plots the number of messages received versus the number of messages sent between all pairs of users. The figure is symmetric along the line $y = x$, as we plot both data points, (w_{ij}, w_{ji}) and (w_{ji}, w_{ij}) per pair. The graph shows a somewhat predictable trend of symmetric reciprocity between pairs of users. We confirm this



(a) 100 x 100 grid plot of messages received vs messages sent

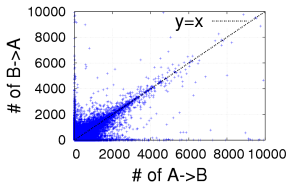


(b) Zoom into (0, 0) and (100, 100) of (a)

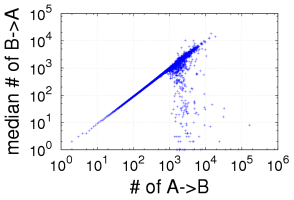


(c) Median number of messages received vs. messages sent

Figure 10: Comparison of messages received and sent per user



(a) no. of received messages vs. no. of written messages between a pair of users



(b) median no. of received messages vs. no.of written message in log scale

Figure 11: Comparison of message received vs written between a pair of users

symmetric reciprocity by plotting the median number of received messages versus the number of written message in Figure 11(b). Note that the x-axis in Figure 11(b) is log scale, while that of Figure 11(a) is in linear scale and covers only up to $10,000$. Beyond $x = 10,000$, a very small number of points exist in Figure 11; almost one data point for each value of x, if any. Figure 11(b) fits well onto the line, $y = x$, up to about $x = 1,000$. Beyond $x = 1,000$, it is hard to interpret the data points, as there are a limited number of data points.

In the Cyworld activity network, the ratio of the number of messages received against that of messages sent pair-wise is fixed to 1. In his pioneering work of operational definition for reciprocity, Hemelrijk has distinguishes three types of reciprocity: 'relative', 'absolute', and 'qualitative' [22]. According to his classification, the reciprocity in the activity network is *close to* 'absolute'; Hemelrijk projects that 'absolute' reciprocity is expected when there are no differences in individuals' capacities. His interpretation still holds in our problem setting as the development of technology gives a nearly equal power to all users in electronic communication. Cyworld facilitates users to surf to friends' homepages by simply clicking a button, and has almost no barrier for any users.

To quantify the reciprocity in the activity network, we use *link reciprocity* defined in [43]. When user i wrote p messages to user j and user j wrote q messages to user i, we measure the correlation between p and q, which map to w_{ij} and w_{ji} in our case. We use Garlaschelli and Loffredo's method to calculate the reciprocity coefficient ρ of the entire network based on link reciprocity [16].

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \overline{a})(a_{ji} - \overline{a})}{\sum_{i \neq j}(a_{ij} - \overline{a})^2} \tag{4}$$

where $\overline{a} = \sum_{i \neq j} a_{ij}/N(N-1)$ and N is the number of nodes. The reciprocity coefficient tells whether the number of mutual links in the network is more or less than that of a random network. If the value ρ is bigger than 0, the network is reciprocal; otherwise, anti-reciprocal. The value ρ for the our activity network turns out to be 0.7775. The reciprocity coefficient in (4) does not take weights into consideration. We substitute a_{ij} in (4) with w_{ij} and recalculate the reciprocity coefficient to obtain 0.7650. Compared to 0.5165 of the World Wide Web network [16], 0.231 of Email Networks [16], 0.28 of Slashdot [19], 0.58 of Twitter [24], and 0.32 of wikipedia [45], the quantitative link reciprocity of the Cyworld activity network is second only to World Trade Web with exceptionally high 0.952 [16]. Our analysis on reciprocity demonstrates that Cyworld users interact in a strongly reciprocal way through the guestbook, and it is common characteristics of social systems.

4.2 Disparity

The median numbers of messages received and sent represents how active a user is, but do not tell if the user interacts evenly with all friends or not. The intuition is that a user is more likely to interact evenly if the number of friends is small. Disparity of a node is a metric that shows the spread of activity of a user over all the friends [3, 11]. $Y(k, i)$, as defined below, is a sum of squares of the number of messages sent over the total number of messages received for node i with out-degree k and in-degree k_{in}.

$$Y(k, i) = \sum_{j=1}^{k} \left\{ \frac{w_{ij}}{\sum_{l=1}^{k_{in}} w_{li}} \right\}^2 \tag{5}$$

$Y(k)$ represents $Y(k, i)$ averaged over all nodes with the same node degree k. When the weights of out-degree edges of a node is comparable to those of in-degree edges, then $kY(k) \sim 1$. If the majority of the activity is carried by a single in-degree or out-degree edge, then $kY(k) \sim k$. We plot $kY(k)$ against k in Figure 12. From the figure, we observe two distinct regions of different disparities. Between degrees of 1 to about 500, $kY(k)$ does follow k, mapping to $kY(k) \sim k$, but deviates beyond $k > 500$. Actually, once k reaches 1000, $kY(k)$ falls to 1, thus showing $kY(k) \sim 1$. This observation is actually counterintuitive. The distribution of disparity in Figure 12 tells us that users with a smaller number of correspondents tend to interact more with a subset of correspondents, while users with a very large number or more than 1000 correspondents actually spread their activity evenly across all of the correspondents.

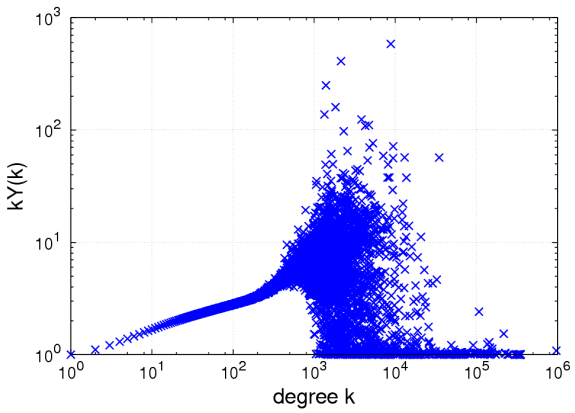


Figure 12: Disparity

4.3 Network Motifs

In Sections 4.1 and 4.2 we have conducted a macroscopic analysis of the activity network and seen that interaction is highly reciprocal, but not evenly spread amongst friends. In this section we delve deeper to the interaction patterns among users and their evolution.

The basis of group interaction begins with three people. There are 13 possible patterns of directional interaction among three people, and they are called *network motifs*. Milo *et al.* have proposed a network-motif-based categorization of networks [29]. The main idea is to calculate the proportions of 13 network motifs in the network of interest, and compare them against random graphs. We can group networks of similar nature based on the prevalence or scarcity of a certain subset of motifs relative to the random graphs.

The Z-score of a motif represents its proportion in a network. It is calculated as follows:

$$Z_i = \frac{N_{real,i} - m(N_{random,i})}{\sigma_{random,i}}, \quad (6)$$

where $N_{real,i}$ is the number of motif i in the network of our interest, and $m(N_{random,i})$ and $\sigma_{random,i}$ are the mean and the standard deviation of motif i in random graphs, respectively. The Z-score as in (6) is not normalized and grows with the network size. We normalize Z scores by

$$Z_i/(\sum {Z_i}^2)^{0.5} \quad (7)$$

as in [28] and remove the bias. The motif definitions and Z-scores we use here do not take the edge weight into consideration. Thus motifs in this section capture the interaction patterns, not the frequency or intensity of the interaction.

We conduct our motif analysis using FANMOD, a network motif detection tool developed by Wernicke and Rasche citewernicke06. In contrast to previous work that focuses on static snapshots of networks [1, 28, 29, 36] and infer evolutionary paths, we take snapshots from different periods and compare the evolution in motif composition. Instead of building a single activity network from the two-year-long guestbook logs, we build the activity network by the month. More specifically, we choose five representative months, July 2003, January, April, June 2004, and January 2005, and count the motifs in them. FANMOD takes as input the in-degree and out-degree distributions of the target network and the number of random networks. We use 100 random networks for the first 4 activity networks, but only 20 random networks for the last and largest activity network from Januanry 2005. Motif analysis is computationally costly and the computation of motifs from June 2004 took more than two weeks on our 64-bit 3.4 GHz dual-core Intel Xeon dual-processor server with 8 GB main memory. We had to curb the number of random networks for the January 2005 network.

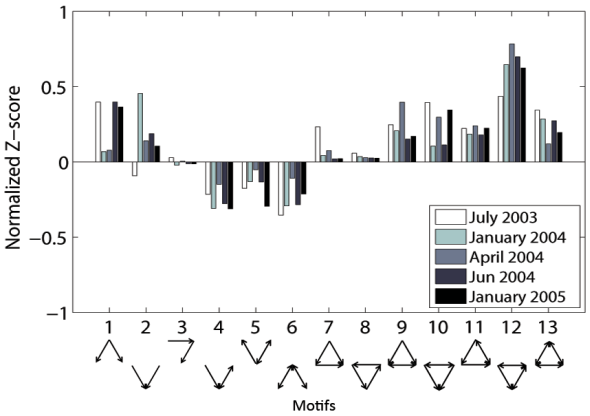


Figure 13: Normalized Z-scores of five activity networks

Figure 13 shows the result of motif analysis. We observe that transitive motifs (motifs 9, 10, 12, and 13 in Figure 13) are abundant, while intransitive motifs (motifs 4, 5, 6 in Figure 13) are rarely constructed in Cyworld. These findings comply with results from small-scale social networks in [28]. The normalized Z-scores for motifs 1 and 2 in Figure 13 deviate from what we expect in social networks. Both motifs are intransitive relations: motif 1 is a broadcasting type. A user writes to others, but they do not know each other nor respond back to the writer. Motif 2 presents the opposite relations. Two users do not know each other, but write to the same user; the recipient does not respond.

People instantly think of spammers for motif 1 and celebrities for motif 2. As we have not excluded comments suspected to be from automated scripts in the construction of our activity network, the prevalence of motif 1 is easily understood. Celebrities physically cannot communicate with a large number of their fans, and there is no need for their fans to know each other. Thus much contribution to motif 2.

Another point to note about the motif distribution in the activity network is self-posts. We remind you that self-posts take up 38.9% of all posts. If disambiguated, self-posts add a directed edge to a motif, if the edge is not already present. In general, self-posts will decrease the proportion of intransitive motifs, and raise the Z-scores of transitive motifs, once disambiguated.

The monthly window for the activity incurs boundary effects for interaction at the beginning and the end of the month. Most users' posts are spaced at one day or shorter apart as we see in Figure 15 of Section 5.2 and we conclude the boundary effects are negligible.

In summary the network motif analysis of the activity network demonstrates that the online interaction through the Cyworld guestbook feature follows other social networks [29] closely, but massive one-way communication suspected to be from spammers and celebrities distort the Z-scores for motifs 1 and 2. In spite of the explosive growth during the period, the Z-score distributions of the network motifs have consistently shown the most proximity to those from previously analyzed social networks.

5. OTHER ACTIVITY-RELATED ASPECTS

Here we investigate two other activity-related aspects, namely, capacity cap and time intervals. We examine whether the number of messages a user writes increases with the number of friends the user has online. The number cannot increase indefinitely and is bound to fall off at some point. The fall-off point speaks for the innate upper bound on human capacity of guestbook-like online social interaction. The next topic we look at is the time intervals between messages sent. The online social network services are a rather new phenomenon and not much is known about human behaviors on those services. The time interval analysis should provide a macroscopic understanding on the frequency of service usage.

5.1 Capacity Cap

We first ask the following question: "Are people socially more active, if they have many friends?" We would like to know if one's number of friends plays an encouraging role, as the more friends have joined the same online social networking service, the more peer pressure one might receive.

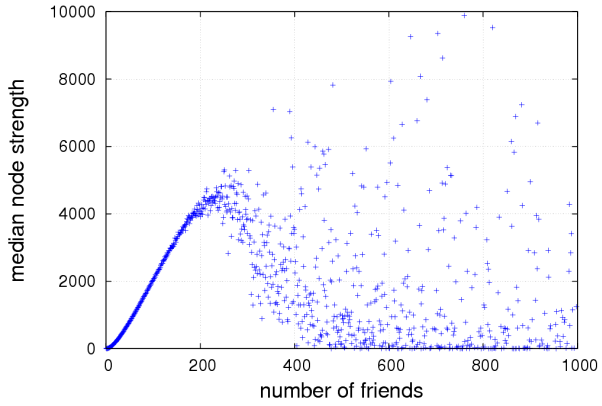


Figure 14: Median node strength vs number of friends

In Figure 14 we plot the median node strength against the number of friends per user. We see that the number of friends does influence the node strength up to users with about 200 friends. That is, people with up to 200 friends respond to peer pressure to stay active online. Then the node strength starts to decrease, even when the number of friends keeps increasing. The Pearson correlation coefficient of the overall graph in Figure 14 is 0.2071. We split the users into two separate groups, those with 200 or fewer friends and with more than 200, and compute the correlation coefficients. For the first group, the Pearson correlation coefficient is 0.6235, that is, strongly positive; for the other group, only 0.00913. Intuitively, the more friends one has, the more active one should be socially. However, beyond 200 or so friends, one must reach a limit in one's socializing capacity. We have plotted the number of messages without self-posts against the node degree, and have observed the same cut-off at 200. In order to reflect the sparsity of data points beyond the degree of 200, we have used different bin sizes for $x > 200$. Still, we see a clear cut-off at $x = 200$.

This striking behavior is in agreement with the previous work that reports a fall-out from a single scaling behavior in the node degree distribution [2]. We have conjectured the emergence of online-only relationships for the multi-scaling behavior in the degree distribution and referred to Dunbar's number of 150 for a plausible explanation [13]. Dunbar bases his work on the analysis of the role of languages and the corresponding development in human brain in the evolutionary path. Dunbar's law extrapolates a limit on the number of manageable relationships by the species based on its neocortex size and "the limit imposed by neocortical processing capacity is simply on the number of individuals with whom a stable inter-personal relationship can be maintained." [12].

As a recent news article points out [8], the technology-assisted social network size is fundamentally intriguing as it challenges our innate capacity for social grooming. Figure 14 tells one of the first evidences that people respond to comments and manage a social network size up to 200 online. Whether they manage off-line contact with these 200 friends or not is beyond the scope of this work. The young generation are more at ease and faster in adopting new technologies. Correlation between the group size and the age might reveal the generational gap quantitatively. Also the microscopic analysis of intra- and inter-group dynamics would provide baseline facts about online socializing behaviors. We leave these questions for future work.

5.2 Time Interval

Daily visitors to the site probably write and visit more guestbooks than infrequent visitors. The time interval between visits tells much about the inherent underlying cognitive behavior as well as patterns of induced traffic for network engineering purposes. In this section we analyze the time intervals between consecutive messages written by one user.

The call arrival process in the telephone network has long been known to follow the Poisson distribution. With the advent of the Internet and new killer applications, the arrival process of new traffic type requires close examination for any change in the underlying stochastic nature. Arrival patterns of HTTP requests at a web server are critical to monitoring and management of the system performance and quality-of-service. Crovella and Bestavros have published one of the first work on world-wide web traffic [9]. They have shown that the distribution of HTTP request arrivals and transmission times follow heavy-tail. Much work on traffic modeling followed, mostly from traffic log mining.

Figure 15 shows time intervals between two consecutive guestbook messages by users. In contrast to earlier sections, we include self-posts, because they are relevant user activity and we need not identify the recipients. We can divide the time interval distribution into three regions: $x < 36$ min, 36 min $< x < 1$ day, and $x > 1$ day. All the three regions follow power-law, while the third region has an undulating pattern with daily peaks.

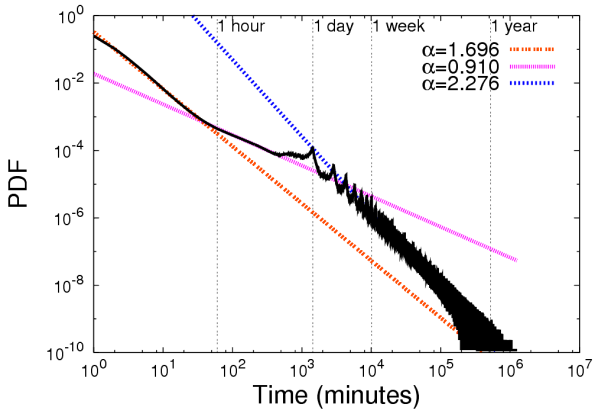


Figure 15: Time intervals between guestbook comments

The first region ($x < 36$ min) maps to comments written at a very short time interval of 36 minutes or shorter. The first data point on the y-axis is from when $x = 1$ min. Writing a comment within a minute or less after the previous comment is somewhat not humanly, although writing quick, short messages, such as "Okay!" or "Hey!", is not unheard of. Golder *et al.* have used 5 seconds as the threshold of inter-message time for automatically generated messages and manual ones [18]. They report 43% of the collected Facebook messages to be spam. The time granularity in our guestbook logs is in minutes, and we cannot verify their choice of the threshold. However, we can extrapolate from Figure 15 that the amount of spam classified with their threshold value is likely to be far smaller. In the Microsoft Messenger network analysis, Leskovec and Horvitz have presented the power-law exponent of 1.5 of time interval distributions between consecutive conversations started by one user [27]. Barabasi points out that human beings receive tasks and execute them in two separate processes, and has built a model for the task execution intervals for variable-length tasks with the power-law exponent of 1.5 [6].

The power-law exponent from our data is slightly larger than from the other two. Though the Cyworld guestbook has no upper limit on the number of characters per entry, most entries are terse enough to be viewed without scrolling down. We believe the finite nature of the guestbook entries contribute to the slightly larger power-law exponent. However, we do not have data about the length of messages, and cannot confirm our proposition.

The power-law exponent for the second region (36 minutes $<$ $x < 1$ day) comes out to be 0.910. If the first region maps to intra-session intervals, this region is representative of inter-session intervals in a day. In [6] Barabasi begins with a simple model for human task execution and verifies his model with email data. The time intervals between consecutive emails have the power-law distribution with an exponent of $\alpha = 1$. As we expect people to write more emails a day than logging in multiple times to Cyworld a day, the gap between Barabasi's exponent of 1 and ours of 0.910 from (36 minutes $< x <$ 1 day) is acceptable.

The time intervals beyond a day maps to login intervals. The undulating pattern with daily peaks is the same as in literature [18, 27]. Temporal synchronization across days is possibly due to users logging in during regular breaks from school and work and in the evening hours.

In summary, the distribution of inter-message times has three distinctive regions: for the short time range, the inter-message time is inhumanly short, explaining the presence of spam and their estimated portion. Yet the distribution agrees with other known task distributions (*e.g.* MSN messengers). For inter-message times longer than 36 min and shorter than 1 day, the behavior closely matches that of email transmissions, but at a lower rate. Beyond 1 day, we detect daily synchronization of activity, but otherwise a fast decaying distribution.

6. RELATED WORK

Social network analysis has been mostly an area for sociologists and anthropologists [43]. As electronically compiled social network data has enabled researchers to gain access to large-scale statistics of networks, it has opened up new possibilities to researchers in other fields, such as physics and computer science. Before the emergence of large-scale online social network servicing portals, a variety of other online social networks have been analyzed. Email networks are one of the most studied such networks [14, 25, 41, 42]. Valverde and Solé study the social network of open source communities [41, 42]. The massive data of mobile communication has recently been analyzed [34, 35]. Using mobile phone records of millions of people, they examine the communication pattern of people. They argue that the stability of the communication network largely depends on the weak ties in the network.

Holme *et al.* analyze an online dating community in detail [23]. In their work, the time evolution of activity shows the saturation of degree and power-law activity pattern. Mislove *et al.* investigate not only online SNS, but also other web services that have social networking features [30]. Java *et al.* conducted a research on microblogs [24]. In their work, the in-degree and out-degree distribution of the network, and the activity pattern of users were analyzed. Leskovec *et al.* analyzed the largest social network (instant messaging network) ever published [27]. They analyzed various aspects such as the number of buddies, the duration of conversations, the time interval between each conversation, geographic usage patterns.

Adamic *et al.* conduct the motif analysis of Yahoo! Answers [1]. They report that the feed-forward loop [36] (motif 7 in Figure 13) is abundant in 'Programming' topic. This abundance of the feed-forward loop is argued as a unique characteristic of knowledge communities because experts help those who have lower-level expertise. Also, Milo *et al.* shows that transitive interactions and transitive triads are more abundant than intransitive ones [28].

The friendship in online is different from that in offline. Cummings *et al.* compares the quality of means of social interactions [10]. They show that online relationships are not considered valuable as much as offline ones. It is concurrent to our knowledge that the online relationship seems to be more shallow and easy to be formed [2].

7. CONCLUSIONS

In this work we have investigated the guestbook logs of Cyworld, the biggest OSNS in Korea. We construct a weighted and directed activity network based on the comments users wrote in other guestbooks. The Cyworld activity network witnessed explosive growth in 2004, and our data spans from July 2003 to October 2005, covering the peak growth period. Two limitations of the present data set come from self-posts and likely spams. The activity network contains self-loops that map to self-posts. Self-posts are likely to be answers to previous posts, but we could not disambiguate self-posts according to the intended receiver. Although banned, automated scripts have been in use in Cyworld and our guestbook logs contain comments generated by such scripts. We suspect most automatic-script-generated comments are spam, but have no means to validate this.

We have analyzed the structural characteristics of the activity network and compared them with the friends network. Although the activity network is weighted and directed, its structure is similar to the friend relationship network. The in-degree and out-degree distributions are close to each other and social interaction through the guestbook is highly reciprocated. When we consider only those links in the activity network that are reciprocated, the degree correlation distribution exhibits much more pronounced assortativity than the friends network and places it close to known social networks. The k-core analysis gives further corroborating evidence that the friends network deviates from the known social network and has an unusually large number of highly connected cores.

Thanks to its weighted and directed nature, the activity network lends insight into the actual dynamics of interaction between users of an OSNS. In our analysis, we have seen that a user is reciprocated for his or her activity in terms of the total number of messages, and unequivocal reciprocity holds between most pairs of users. However, the analysis of disparity shows that users who have roughly two hundred friends or less tend to communicate unevenly,

while those users with very large numbers of friends interact more evenly. The network motif analysis of the activity network demonstrates that online interaction through the Cyworld guestbook feature follows trends of other social networks [29] closely. However, massive one-way communication suspected to be from spammers and celebrities distorts the Z-scores for motifs 1 and 2. In spite of explosive growth during the period of interest, the Z-score distributions of the network motifs consistently show high proximity to those from previously analyzed social networks.

We report an interesting observation that the activity measured by the node strength increases linearly with the number of friends, but starts to fall off beyond 200. This number is larger than what we anticipated through Dunbar's number, and requires further investigation. We have found that the distribution of inter-message time intervals has three different regions, but for a comprehensive understanding of this tri-modal behavior has yet to be reached. This is yet another point for more work.

Online social networks are spreading very rapidly and affect many corners of our daily web experience. The present work offers insight into the macroscopic behavior of an online social network. We plan to extend this work to include an analysis of group dynamics and microscopic analyses from the perspecitves of individual users. We hope this work offers insight into information flow in cyber space.

Acknowledgements

H. Chun, H. Kwak, and S. Moon were supported by the IT R&D program of MKE/IITA [A1100-0801-2758, "CASFI: High-Precision Measurement and Analysis Research"]. Y.-H. Eom, Y.-Y. Ahn, and H. Jeong were supported by Acceleration Research CNRC of MOST/KOSEF (R17-2007-073-01001-0). We thank Jeongsu Hong and Jaehyun Lim of SK Communications, Inc. for providing the Cyworld data. We would like to express sincere gratitude to our shepherd Thomas Karagiannis and anonymous reviewers. Their feedback was absolutely vital in improving the presentation of this work.

8. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *WWW '08*, pages 665–674. ACM, 2008.

[2] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM Press.

[3] E. Almaas, B. Kovacs, T. Vicsek, Z. N. Oltvai, and A.-L. Barabasi. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427:839–843, Feb 2003. Letters to Nature.

[4] R. Aukett, J. Ritchie, and K. Mill. Gender differences in friendship patterns. *Sex Roles*, 19, 1988.

[5] R. Axelrod and W. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[6] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[7] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of National Academy of Sciences*, 101(11):3747–3752, March 16 2004.

[8] C. Bialik. Sorry, you may have gone over your limit of network friends. *The Wall street journal*, November 16 2007.

[9] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE /ACM Transactions on Networking*, 5(6):835–846, 1997.

[10] J. N. Cummings, B. Butler, and R. Kraut. The quality of online social relationships. *Commun. ACM*, 45(7):103–108, 2002.

[11] B. Derrida and H. Flyvbjerg. Statistical properties of randomly broken objects and of multivalley structures in disordered systems. *Journal of Physics A: Mathematical and General*, 20:5273–5288, 1987.

[12] R. Dunbar. Co-evolution of neocortex size, group size, and language in humans. *Behavioral and brain scineces*, 16(4):681–735, 1993.

[13] R. Dunbar. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, 1998.

[14] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66:035103, 2002.

[15] Y.-H. Eom, C. Jeon, H. Jeong, and B. Kahng. Evolution of weighted scale-free networks in empirical data. accepted in Phys. Rev. E.

[16] D. Garlaschelli and M. I. Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93:268701, 2004.

[17] K.-I. Goh, Y.-H. Eom, H. Jeong, B. Kahng, and D. Kim. Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions. *Phys. Rev. E*, 73:066123, 2006.

[18] S. A. Golder, D. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *3rd International Conference on Communities and Technologies (CT2007). East Lansing, MI.*, June 2007.

[19] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceedings of the 17th international conference on World Wide Web*, New York, NY, USA, April 2008. ACM.

[20] A. W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25:161–178, april 1960.

[21] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[22] C. K. Hemelrijk. Models of, and tests for, reciprocity, unidirectionality and other social interaction patterns at a group level. *Animal Behavior*, 39:1013–1029, 1990.

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

[24] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[25] G. Kossinets and D. Watts. Emprical Analysis of an Evolving Social Network. *Science*, 311(88), 2006.

[26] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of ACM KDD*, 2006.

[27] J. Leskovec and E. Horvitz. Planetary-scale views on an instant-messaging network. http://arxiv.org/abs/0803.0939, Mar 2008.

[28] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, 2004.

[29] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[30] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *ACM Internet Measurement Conference*, October 2007.

[31] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct 2002.

[32] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126, Feb 2003.

[33] M. A. Nowak. Five Rules for the Evolution of Cooperation. *Science*, 314(5805):1560–1563, 2006.

[34] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9:179, 2007.

[35] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proc. Nat. Acad. Sci.*, 104(18):7332, 2007.

[36] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[37] G. Simmel and K. H. Wolff. *The Sociology of Georg Simmel*. Free Press, 1950.

[38] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *IMC '06*, pages 27–40, New York, NY, USA, 2006. ACM.

[39] R. Thurnwald. *Economics in Primitive Communities*. Oxford University Press, 1932.

[40] S. T. Tong, B. V. D. Heide, L. Langwell, and J. B. Walther. Too much of a good thing? the relationship between number of friends and interpersonal impressions on Facebook. *Journal of Computer-Mediated Communication*, 13:531–549, 2008.

[41] S. Valverde and R. V. Solé. Evolving social weighted networks: Nonlocal dynamics of open source communities. arXiv:physics/0602005v1.

[42] S. Valverde and R. V. Solé. Self-organization versus hierarchy in open-source social networks. *Phys. Rev. E*, 76:046118, 2007.

[43] S. Wasserman and K. Faust. *Social network analysis*. Cambridge University Press, Cambridge, 1994.

[44] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22:1152–1153, 2006.

[45] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks, Jul 2006.