Analysis of Topological Characteristics of Huge Online Social Networking Services

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Young-Ho Eom, Sue Moon, Hawoong Jeong

Abstract— **Social networking services are a fast-growing business in the Internet. However, it is unknown if online relationships and their growth patterns are the same as in real-life social networks. In this paper, we compare the structures of three online social networking services: Cyworld, MySpace, and orkut, each with more than 10 million users, respectively. We have access to complete data of Cyworld's *ilchon* (friend) relationships and analyze its degree distribution, clustering property, degree correlation, and evolution over time. We also use Cyworld data to evaluate the validity of snowball sampling method, which we use to crawl and obtain partial network topologies of MySpace and orkut. Cyworld, the oldest of the three, demonstrates a changing scaling behavior over time in degree distribution. The latest Cyworld data's degree distribution exhibits a multi-scaling behavior, while those of MySpace and orkut have simple scaling behaviors with different exponents. Very interestingly, each of the two exponents corresponds to the different segments in Cyworld's degree distribution. Certain online social networking services encourage online activities that cannot be easily copied in real life; we show that they deviate from close-knit online social networks which show a similar degree correlation pattern to real-life social networks.**

I. Introduction

The Internet has been a vessel to expand our social networks in many ways. Social networking services (SNSs) are one successful example of such a role. SNSs provide an online private space for individuals and tools for interacting with other people in the Internet. SNSs help people find others of a common interest, establish a forum for discussion, exchange photos and personal news, and many more. Cyworld, the largest SNS in South Korea, had already 10 million users 2 years ago, one fourth of the entire population of South Korea. MySpace and orkut, similar social networking services, have also more than 10 million users each. Recently, the number of MySpace users exceeded 130 million with a growing rate of over a hundred thousand people per day. It is reported that these SNSs "attract nearly half of all web users" [1]. The goal of these services is to help people establish an online presence and build social networks; and to eventually exploit the user base for commercial purposes. Thus the statistics and dynamics of these online social networks are of tremendous importance to social networking service providers and those interested in online commerce.

The notion of a network structure in social relations dates back about half a century. Yet, the focus of most sociological studies has been interactions in small groups, not structures of large and extensive networks. Difficulty in obtaining large data sets was one reason behind the lack of structural study. However, as reported in [2] recently, missing data may distort the statistics severely and it is imperative to use large data sets in network structure analysis.

It is only very recently that we have seen research results from large networks. Novel network structures from human societies and communication systems have been unveiled; just to name a few are the Internet and WWW [3] and the patents, Autonomous Systems (AS), and affiliation networks [4]. Even in the short history of the Internet, SNSs are a fairly new phenomenon and their network structures are not yet studied carefully. The social networks of SNSs are believed to reflect the real-life social relationships of people more accurately than any other online networks. Moreover, because of their size, they offer an unprecedented opportunity to study human social networks.

In this paper, we pose and answer the following questions:

What are the main characteristics of online social networks? Ever since the scale-free nature of the World-Wide Web network has been revealed, a large number of networks have been analyzed and found to have power-law scaling in degree distribution, large clustering coefficients, and small mean degrees of separation (so called the small-world phenomenon). The networks we are interested in this work are huge and those of this magnitude have not yet been analyzed.

How representive is a sample network? In most networks, the analysis of the entire network is infeasible and sampling is unavoidable. We evaluate the popular snowball sampling method using the complete Cyworld network topology.

How does a social network evolve? From the historical data of the Cyworld network, we reconstruct the past snapshots of the Cyworld network. The evolution path of a social network may exhibit patterns of major change. It would be of tremendous interest to latecomers, for it is literally a prediction of what they might turn into.

This paper is organized as follows. In Section 2, we review related works. In Section 3, we describe social network topologies we have gained access to and, in Section 4, the snowball sampling method and metrics of interest in network analysis. Then we begin our topology analysis with that of Cyworld. In Section 5, we analyze the friend relationship network of Cyworld, perform a historical analysis on the network's evolution, evaluate the snowball sampling method, and compare a special subset of Cyworld network (the network of testimonials) with the complete network, and . In Section VI, we analyze MySpace and orkut networks. We compare all three networks in Section VII, along with a discussion on the origin of power-law in online social networks and the resemblance to real social networks. We conclude with a discussion on future work in Section VIII.

II. Related Works

The structural propertes, such as degree distribution, of large-scale social networks have received much attention since the uncovering of the network of movie actors [3]. It is followed by the analysis on the network of scientific collaboration network [5] and the web of human sexual contacts [6]. However, a link in these networks is different from a normal *friend relationship* and the large-scale analysis on the such networks has remained uncharted.

Recently, the rapid growth of online social networking services made it possible to investigate the huge online social network directly. Since the rise of Cyworld, many SNSs including MySpace and orkut have grown. However, the analyses on these huge networks have been limited to cultural and business viewpoint [7].

Here, we introduce two relevant works on online social networks. First work is on an Internet dating community, called pussokram.com [8]. The dataset consists of about $30,000$ users and time series of all interactions. By network analysis, fat tails are found in all degree distributions from the networks made by several interaction layers: messages, guest book, and flirts. An interesting feature is super-heavy tails which go beyond the trend of small degree region. Another work investigates an online blog community, LiveJournal [9]. The number of users examined is $1,312,454$, about half of whom publicize their snail mail addresses. By examining this partial list of real addresses of bloggers, the work uncovers the connection between online friendship and geography. The network's degree distribution also shows a weak but significant super-heavy tail, which deviates from the trend of small degree region. The huge size of online communities makes the sampling an inevitable process in analyzing the networks. Recently, extensive simulations are performed for several network sampling methods [10] and the effect of missing data in social network analysis is studied [2].

III. Online SNSs

Social networking services (SNSs) provide users with an online presence that contains shareable personal information, such as a birthday, hobbies, preferences, photographs, writings, etc. Most SNSs offer features of convenience that help users form and maintain an online network with other users. One such feature is a "*friend*." A user invites another user to be one's friend. If the invited user accepts the invitation, a friend relationship is established between them. This friend feature is often used as a shortcut to others' front pages and, in some SNSs, a convenient tool to exchange messages and stay in touch.

As SNSs have become very popular, leading sites boast of an extensive user base of tens of millions of subscribers. For this work, we have collected four sets of online social network topology data from three popular SNSs. All of them have more than 10 million users. We have obtained the entire network topology of Cyworld directly from its provider, and sampled others through web crawling. We believe our work is the first to analyze social networks of such magnitude. In Table I, we summarize the four data sets described above. The metrics in the table are explained later in Section IV.

Below we describe each network in detail.

A. Cyworld

Cyworld is the largest and oldest online social networking service in South Korea. It began operation in September 2001, and its growth has been explosive ever since. Cyworld's 15 million registered users, as of November 2006, are an impressive number, considering the total population of 48 million in South Korea. As any SNS, Cyworld offers users to establish, maintain and dissolve a friend (called *ilchon*) relationship online.

From SK Communications, Inc. the provider of the Cyworld service, we received an anonymized snapshot of the Cyworld user networks taken in November 2005 and November 2006. The snapshot contains 191 (291) million friend relationships among 12 (15) million users, respectively. We have access to additional data to study the network evolution and describe it in Section V-B.

Cyworld offers a mechanism called *ilchon pyung* for friends to leave a *testimonial* on a user's front page[1] Friends can leave one testimonial each, though modifiable more than once, and it usually describes one's impression or a word of encouragement. Not all friends write a testimonial and thus it can be construed as a manifestation of a close relationship in real life. We have decided to include testimonials in our analysis for comparison with the complete Cyworld network. We have used snowball sampling and collected testimonials from about 100,000 users. Note that the testimonial network is a directional graph, where a link represents one user's testimonial on another user's front page, and the complete Cyworld friend network is an undirected graph.

B. MySpace

As of August 2007, MySpace is the largest social networking service in the world, with more than 190 million users. It began its service in July 2003, and the number of users grew explosively. According to Alexa.com[2], it is the world's 5th most popular website (4th among English websites).

A new user in MySpace by default gets a friend relationship with Tom Anderson, the cofounder of MySpace. In our dataset, we exclude links to him, since he has links to everyone. MySpace offers similar features with other social networking services, such as writing testimonials to friends on their front pages, checking upcoming birthdays, shortcuts to friends' front pages.

In this paper, we use two data sets of MySpace. We have obtained 100,000 user information and the link between them from the MySpace friend network by crawling the MySpace online web site from September to October, 2006. The crawler randomly selects a starting user site, and crawl the user's friends' pages, their friends' pages, and so on. We have left out

[1]Only 101 testimonials were allowed on front page originally, but the restriction was lifted.

[2]http://www.alexa.com/

Set #	I	II	III	IV	V
	Cyworld (2005)	Cyworld (2006)	Testimonial(Cy)	MySpace	orkut
sampling ratio p	100%	100%	0.77%	$\sim 0.08\%$	$\sim 0.30\%$
no. of nodes N	$12,048,186$	$15,149,764$	$92,257$	$100,000$	$100,000$
no. of edges L	$190,589,667$	$291,075,290$	$703,835$	$6,854,231$	$1,511,117$
mean degree $\langle k \rangle$	31.6	38.4	15.3	137.1	30.2
avg. clustering coefficient C	0.16	0.16	0.32	0.26	0.31
assortativity r	-0.005	-0.003	0.43	-0.20	0.03
estimated degree of separation ℓ	5.3	4.5	7.1	2.7	3.8

TABLE I

SUMMARY OF DATA SETS FROM ONLINE SOCIAL NETWORKING SERVICES

users who do not publicize their firends' list, and the amount of those users were about 23% out of all the nodes we have crawled.

Another data set is intended to see the degree distribution more accurately. Thanks to the regular form of an user's id, we generate many user ids randomly and visit their homepage to obtain the number of friends. We sampled $1,238,502$ nodes.

C. orkut

In September 2002, orkut began its trial service by a few Google employees, and became an official Google service in January 2004. Until recently, orkut accounts were given only to people invited by already existing users, which is different from Cyworld or MySpace. As it has permitted any user to create an account without invitation, it has expanded fast; the number of users reached 1 million at the end of July and surpassed 2 million by the end of September 2004[3]. Today, the number of orkut users exceeds 33 million. Once a user joins orkut, one can publish one's own profile, upload photos, and join communities of interest. Orkut also offers friend relationship. The maximum number of friends per user was limited to 1000, but this limit has also been lifted. Crawling in a similar way to MySpace, we have collected orkut friendship data on 100,000 users from June to September, 2006.

IV. ANALYSIS METHODOLOGY

In this section, we outline the sampling method employed to crawl and capture MySpace and orkut networks, and describe briefly the metrics of topological characteristics and their interpretations.

A. Snowball sampling

We have gained access to the entire topology of Cyworld through human contact, but were not successful with MySpace or orkut. Falling back on crawling for data collection, we are limited in the number of nodes we could crawl in a finite time frame.

There are several network sampling methods: node sampling, link sampling, and snowball sampling [10]. In node sampling, we randomly select a fraction of nodes. Links between selected nodes are included in the final sample network. Link sampling is similar. We randomly select a fraction of links and construct a sample network. In contrast, snowball sampling randomly selects one seed node and performs a breadth-first search (hence the name, snowball sampling), until the number of selected nodes reaches the desired sampling ratio. Only those links between selected nodes are included in the final sample network.

The snowball sampling method is the only feasible one to crawl the web for the following reasons. First, if the sampling fraction is not large enough, the sample network is likely to consist of many small and isolated clusters and be far from the original network in many aspects of interest. Second, node and link sampling methods are inefficient. In node sampling, if sampled nodes are not well connected with each other, most links crawled should be dumped, and it damages the efficiency of algorithm. Meanwhile, link sampling obtains information of only one neighbor at each visit to a web page, it's efficiency is far worse than the snowball sampling method. Furthermore, the expected mean degree of the sample network is always much smaller than of the original network. In short, we always underestimate the node degree, cannot estimate the degree of separation, nor find hubs (nodes with a very large number of neighbors), if we use node or link sampling. By these reasons, we use snowball sampling method to obtain a sample network from web. In addition, we use random degree sampling for MySpace network to obtain more accurate degree distribution of MySpace. In this method, we visit many random nodes and crawl the degree information of the nodes, without crawl the information that who is connected to whom.

The main estimation error of the snowball sampling lies in the likelihood of oversampled hubs [10], for they have many links and are easily picked up in the first few rounds of the breadth-first search. In order to evaluate the deviating impact of snowball sampling on the metrics of interest, we take full advantage of the complete Cyworld network and compare various metrics between partial and complete networks.

It is known that the power-law nature in the degree distribution is well conserved under snowball sampling [10] since the snowball sampling method easily picks up hubs. This property reduces the degree exponent and produces a heavier tail, but it is difficult to get a power-law degree distribution from a network without the power-law decaying degree distribution.

B. Metrics of interest

We begin the analysis of online social network topologies by looking at their degree distributions. Networks of a power-law degree distribution, $P(k) \sim k^{-\gamma}$, where k is the node degree

[3]http://en.wikipedia.org/wiki/orkut

and $\gamma \leq 3$, attest to the existence of a relatively small number of nodes with a very large number of links. These networks also have distinguishing properties, such as vanishing epidemic threshold, ultra-small worldness, and robustness under random errors [11], [12], [13], [14]. The degree distribution is often plotted as a complementary cumulative probability function (CCDF), $\wp(k) \equiv \int_k^\infty P(k')dk' \sim k^{-\alpha} \sim k^{-(\gamma-1)}$. As a power-law distribution shows up as a straight line in a log-log plot, the exponent of a power-law distribution is a representative characteristic, distinguishing one from others.

Recently, the method of maximum likelihood was suggested as an un-biased and accurate estimator of power-law exponent [15], [16]. The approximate expression for the power-law exponent in the discrete case is given by the following expression:

$$\gamma \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{min} - \frac{1}{2}} \right]^{-1}, \quad (1)$$

where k_{min} is the point where the power-law tail starts, and $k_i, i = 1, ...n$ are the degree of the nodes such that $k_i \geq k_{min}$.

Next, we examine the clustering coefficient. The clustering coefficient of a node is the ratio of the number of existing links over the number of possible links between its neighbors. Given a network $G = (V, E)$, a clustering coefficient, C_i, of node $i \in V$ is:

$$C_i = 2|\{(v, w)|(i, v), (i, w), (v, w) \in E\}|/k_i(k_i - 1) \quad (2)$$

where k_i is the degree of node i. It can be interpreted as the probability that any two randomly chosen nodes that share a common neighbor have a link between them. For any node in a tightly-connected mesh network, the clustering coefficient is 1. The clustering coefficient of a node represents how well connected its neighbors are. The clustering coefficient of a network is the mean clustering coefficient of all nodes. Often it is insightful to examine not only the mean clustering coefficient, but its distribution. We denote the mean clustering coefficient of degree k as $C(k)$ and analyze its distribution. Unless stated otherwise, the clustering coefficient in the rest of the paper refers to the mean clustering coefficient of a network.

The degree correlation, k_{nn}, is a mapping between a node degree k and the mean degree of nearest neighbors of those nodes of degree k. Its distribution is often characterized by the assortativity (r), which is defined as the Pearson correlation coefficient of the degrees of either nodes which is connected by a link [17]. It is expressed as follows:

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{(\langle k_i^2 \rangle - \langle k_i \rangle^2)(\langle k_j^2 \rangle - \langle k_j \rangle^2)}}, \quad (3)$$

where k_i and k_j are degrees of the nodes located at either end of a link and the $\langle \cdot \rangle$ notation represents the average over all links.

If a network's assorativity is negative, a hub tends to be connected to non-hubs, and *vice versa*. When $r > 0$, we call the network to have an assortative mixing pattern, and when $r < 0$, disassortative mixing. Most social networks exhibit an assortative mixing pattern, whereas other networks show a disassortative mixing pattern [17], [18]. The assortative mixing pattern is considered as a unique characteristic of social networks and its origin was suggested as rich community structures of human relationships [19].

As introduced in Stanley Milgram's experiment of mail forwarding [20], the degree of separation (ℓ) is the mean distance between any two nodes of the network. Accurate calculation of the degree of separation or the average path length, as we call it in this paper, requires the knowledge of the entire topology and the time complexity of $O(NL)$, where L is the number of links and N is the number of nodes. In huge networks like Cyworld, MySpace, and orkut, the calculation is infeasible. Only approximation is possible. From a snowball sample network, we measure the number of nodes at each round of breadth-first search. By extrapolating this number sequence, we predict how many steps are needed to cover the entire network, and obtain an estimate of the average path length by the following formula [21].

$$\frac{\log(N/n_1)}{\log(n_2/n_1)} + 1, \quad (4)$$

where N is the total number of nodes and n_1 and n_2 are the average numbers of first and second neighbors respectively.

Palmer *et al.* propose an approximation for the effective diameter of a massive graph [22]. The effective diameter is the 90th-percentile of the path length distribution, and is a better metric than the maximum diameter in estimating the network size, as the maximum diameter can be an outlier from a small number of nodes forming a chain. Palmer's approach is useful when the complete network topology is known, but too large to load on memory and calculate the exact neighborhood function. As our Cyworld topology can be loaded onto our server's 4 GB memory and complete network topologies of MySpace and orkut are not available, we do not use their approximation. For Cyworld data, we estimate the average path length and the effective diameter from sample networks and use (4) to calculate average path lengths for MySpace and orkut networks.

V. Analysis of Cyworld

We begin our network analysis with the Cyworld data set. As it is the most extensive data set we have, this section includes basic analysis of topology-related metrics, evaluation of the snowball sampling method, historical analysis of the network evolution, and online networks' similarity to real-life social networks. In Section V-A, we first calculate the degree distribution, clustering coefficient distribution, and degree correlation distribution of the complete Cyworld network. Then we study how the Cyworld network has evolved over time in Section V-B. In Sction V-C, we evaluate the snowball sampling method by comparing sample networks to the complete Cyworld network. In Section V-D, we analyze the network of testimonials.

A. Complete friend relationship network

The first metric we investigate is the degree distribution. Figure 1(a) plots the complementary cumulative distribution

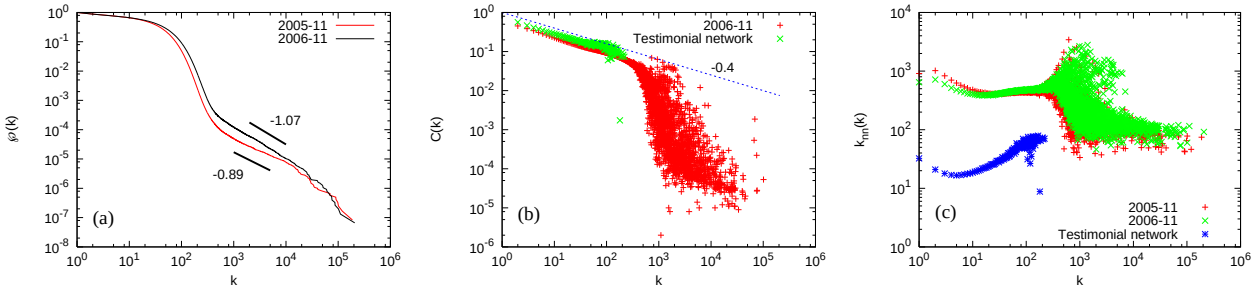

Fig. 1. Topological characteristics of the complete Cyworld network

function (CCDF) of the complete Cyworld friend relationship network, Set I and Set II. Most networks in nature and human societies have a power-law degree distribution with a single exponent, γ, between 2 and 3, attesting to the existence of hubs or people with a very large number of friends. However, the Cyworld network shows two different scaling regions as in Figure 1(a). The crossover takes place between $k = 100$ and $k = 1000$ and divides the CCDF into two regions: a rapid, exponentially decaying region and a heavy tailed ($\gamma \sim 2$) region. This behavior has not been reported previously about any SNS topologies. (Note that the exponent of a CCDF is smaller than the exponent of a probability distribution function itself by one.)The multi-scaling behavior observed in Cyworld suggests that Cyworld consists of two different types of networks, *i.e.*, two types of users. Interestingly, as shown in Figure 2, the degree distribution is remarkably well fitted by the simple functional form with the combination of power-law and exponential function.

$$P(k) = ae^{-\frac{k}{k_c}} + bk^{-\gamma}. \quad (5)$$

To find the constants, we first calculate $\gamma = 2.07$ by the method of maximum likelihood. Then, we use least square fit algorithm with three remaining parameters, which turn out to be $a = 0.018$, $k_c = 46.3$, and $b = 0.21$.

Further analysis of distributions of the clustering coefficient and the degree correlation also support the existence of heterogeneous types of users. We revisit this issue in Section V-C against the analysis of sample networks and discuss further in comparison with other social networks in Section VII.

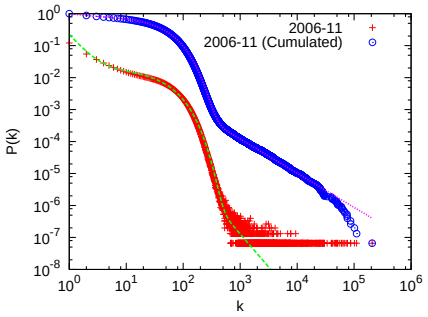

Fig. 2. Fitting of Cyworld degree distribution

Next, we examine the clustering coefficient. Figure 1(b) displays the entire distribution of the clustering coefficient, $C(k)$, versus degree k. (Figures 1(b) and 1(c) also include graphs from the testimonial networks that we defer discussion on until Section V-D.) The average clustering coefficient of the Cyworld network is 0.16, which is smaller than those of other network data sets to be discussed in later sections. It implies that Cyworld friend relationships are more loosely connected than other networks. However, a closer examination of the clustering coefficient distribution reveals that it actually has two different scaling regions, as in the degree distribution. Up until $k = 500$, the graph of the clustering coefficient distribution has a power-law distribution with an exponent of 0.4. Then beyond $k = 500$, the graph suddenly drops and disperses. (The small number of nodes in that region also contributes to the dispersion.) That is, neighbors of nodes with degree larger than 500 are very loosely connected and different from those of nodes with smaller degrees.

We plot the distribution of degree correlation, k_{nn}, of the complete Cyworld network in Figure 1(c). It exhibits not a simple pattern, but a more complex and intractable one, just as the degree distribution of Figure 1(a) or the clustering coefficient distribution of Figure 1(b) do. If a network has an assortative mixing pattern, the degree correlation distribution exhibits an increasing scaling behavior; if disassorative, then a decreasing one. The overall trend of Figure 1(c) is decreasing, and the Cyworld network's assortativity is a negative value: -0.13. As human social networks are known to exhibit an assortative mixing pattern from hubs attracting hubs, lack of such a pattern and the complex, heterogeneous structure in the degree correlation distribution imply mixing of different types of users.

Far-fetched, we could claim that the top left corner of Figure 1(c) shows negative correlation and the remaining upper cluster displays (although very slight) positive correlation, which is the main characteristic of social networks [19]. The degree correlation of large hubs seems to be neutral or slightly disassortative. We revisit this issue later with sample network analysis in Section V-C.

Leskovec *et al.* reports that the average node degree of a wide range of real graphs increases over time and thus the graphs densify [4]. They also report that the effective diameters of those graphs shrink over time.

We estimate the average path length between nodes from sample networks. We randomly choose a seed node, run a breadth-first search of the network, and obtain a distribution of path lengths between the seed and all other nodes. We repeat

this network sampling with 100, 2000, and 3000 seeds, and obtain Figure 3. The inset in the figure plots the cumulative distribution function. We can tell from the figures that the average path length between 90% of nodes is less than 6, even though the maximum distance may reach 18, more than double the value for 90%. In order to confirm that the distribution of the average path length eventually converge to that of the complete network, we have looked at the incremental change as the number of sample networks increased. We observe that the root mean square of the difference between the empirical probability density distributions of sample networks with h and $(h+1)$ seeds steadily decreases fitting to the $y = 1/x$ line, confirming that our estimation would converge to the true average path length.

B. Historical analysis

As of March 2006, the population of Korea reached 48 million, and there were 24 million Internet users of ages over 15, the 6th largest in the world [23]. Internet users can be classified as "weekly" users, if they use the Internet at least once a week, or "daily" users, if they do every day. According to a market research company, Korean Click, the number of weekly users in 2000 was about 15 million and that of daily users was 10 million. When we extrapolate the numbers of daily and weekly users from 2000 to March 2006 along the increase in the overall Internet users, we obtain the graphs labeled as weekly and daily in Figure 4.

From Table I, we know that the number of Cyworld users reached 12 million in November 2005. In addition to Set I and Set II, we have acquired several other data sets to investigate the historical growth of Cyworld. We have monthly statistics on the total numbers of Cyworld users from the very beginning of the Cyworld service in 1999, as well as the numbers of users with friend relationships. The former includes all users with or without any friend relationship. We have acquired two additional snapshots of existing friend relationships from April and September of 2005, which we call Snapshots I and II, respectively. Snapshots I and II, as well as Set I and Set II, include friend relationships. In other words, those users who have not established any friend relationship are not included. We also have acquired partial data (about 42%) on dates of establishments and dissolutions of friend relationships from SK Communications, Inc., which we call Dated Partial Set (DPS). Due to lack of knowledge about the biases in DPS, we assume that it is a random sample. We reconstruct partial friend relationship networks at every six month interval starting in December of 2002 from Dated Partial Set.

Figure 4 plots all the historical data we have about Cyworld and the Internet users. The curvy graph just below the extrapolated line of daily Internet users represents the all-inclusive Cyworld users and the graph below it represents only those with friend relationships. The latter overlaps with four data points obtained from Snapshots I, II, Set I, and Set II. They do not coincide exactly, for data was collected on different dates and marginal differences exist between data sets.

The total numbers of all-inclusive Cyworld users and those with friend relationships follow each other closely up until June 2004 and start to deviate. In June 2006, the gap between the two widens to over 4 million. The dramatic increase in the number of users without friend relationships could be contributed to several factors. One of the most likely factors is additional services offered by SK Communications, Inc. that require a Cyworld membership. People who want to use the service, but are not interested in SNS only join Cyworld and do not engage in building a social network. Whether a Cyworld user has a friend or not, the number of Cyworld users almost reaches the extrapolated number of daily Internet users in Korea in December 2005, about a third of the entire Korean population and more than 60% of Internet users. Considering the language barrier, it is unlikely that the user demography of Cyworld expands outside Korea much, as Brazilians represent a very active and significant portion in orkut or there are many Spanish-speaking users in MySpace. Thus we may conclude that Cyworld has almost reached saturation in terms of the number of users.

The lowest graph in Figure 4 represents partial networks reconstructed from Dated Partial Set based on the date information of relationship establishment and dissolution. Partial friend relationship networks obtained from DPSs clearly do not sample the complete network equally every six months. However, we note that the sampling ratios in terms of the number of nodes are always higher than 5%.

In order to study the historical growth of Cyworld in depth, we use 3 reconstructed partial networks from DPS, each in December of 2002, 2003, 2004, and Set1 from 2005. We refer to those DPS sets as DPS I, II, III, respectively. These networks are not snowball sampled, but are good representations of the complete networks, due to the high sampling ratios of

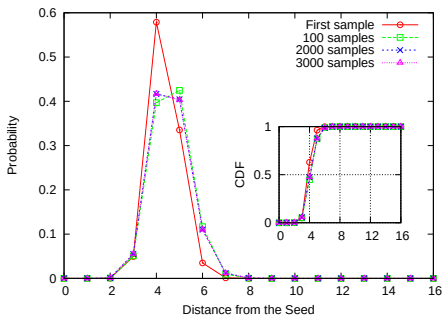

Fig. 3. Probability distribution function of distances between nodes

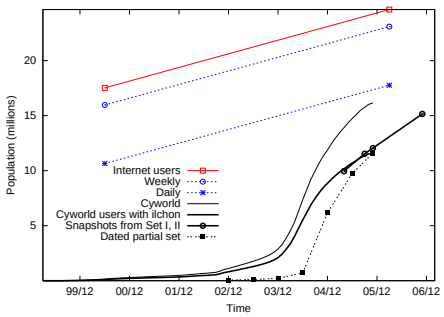

Fig. 4. Comparison of the national population, Internet users, and Cyworld users

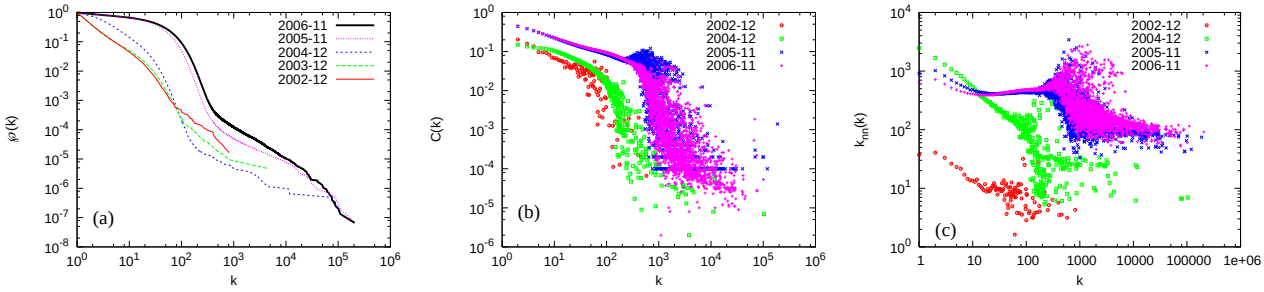

Fig. 5. Evolution of the topological characteristics of Cyworld.

5% or higher. Using these four sets, we study how the Cyworld network has evolved over time in terms of the degree distribution, the clustering coefficient distribution, and the degree correlation distribution.

We first plot the CCDF of the degree distributions in Figure 5(a), and study the evolution of the degree distribution. The graph of DPS I is the shortest, reaching only up to $k = 1000$. It is close to a straight line, and we cannot see if it has a multi-scaling behavior. The graph of DPS II extends longer than that of DPS I and the tail part for $k > 100$ starts to exhibit a different slope. However, it is only with DPSs III and Set I that the multi-scaling behavior is clearly pronounced. From Figure 5, we can infer that user heterogeneity has started to materialize around December 2004.

In Figure 5(b) we plot the distributions of clustering coefficients from December 2002, 2004, November 2005, and November 2006. The distribution from December 2003 is very similar to that of 2004 and is omitted in order to avoid cluttering the plot. Up until December 2004, the distribution of clustering coefficient, $C(k)$, is not power-law and looks similar from one year to another. From 2004 to 2005, the distribution of clustering coefficient changes significantly, and shows different regions of scaling behavior in November 2005.

The degree correlation from December 2002, plotted in the lower left corner of Figure 5(c), exhibits an disassortative mixing pattern. The disassortative mixing pattern continues to manifest up to December 2004 for $k \leq 50$. In November 2005, as we have observed already, the disassortative mixing pattern for $k \leq 100$ is no longer clear, and the degree correlation for $k \geq 100$ is spread out. In November 2006, the mixing pattern of the region between 10 and 200 goes assortative. The manifestation of different types of users is shown in all plots of metrics.

As a last metric of historical analysis, we examine the evolution of the average path length and the effective diameter. In Figure 6, we plot the log of the total number of nodes, the average path length, and the effective diameter as described in Sections IV-B and V-A. We use DPSs I to IV, as well as reconstructed networks from the original DPS for June of 2003, 2004, and 2005. Previously, a complex network is known to have the average path length scale logarithmically to the number of nodes, while a power-law network with $2 \leq \gamma \leq 3$ scales to $log(log(N))$ [13]. However, recent study by Leskovec *et al.* has reported that a wide range of graphs exhibit a densification trend and their diameters shrink over

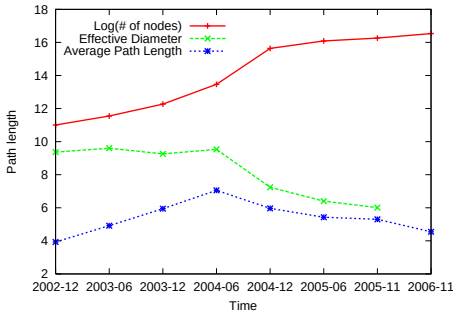

Fig. 6. Average path length vs time

time [4].

In case of Cyworld, the average path length increases up to June 2004, and then starts to decrease, while the effective diameter vascillates until June 2004 and decreases afterwards. Due to space limitation, we only report that the number of nodes plotted against number of edges plotted in log-log scale exhibits a power-law distribution and thus the Cyworld network concurs to the same densification trend.

We plot the effective diameter in Figure 6 and observe that it increases for the first year and a half of Cyworld service, and then starts to decrease. This junction coincides with the point in time when the number of users without friend relationships increases and the multi-scaling behavior starts to manifest in the degree distribution. We conjecture that the Cyworld has reached a saturation point as an active friend relationship network, and then graph densification has taken over, resulting in a decrease of both the average path length and the effective diameter.

C. Evaluation of the snowball sampling method

As discussed in Section IV-A, snowball sampling is a better technique than node or link sampling methods, especially for huge population, if its tendency to overrepresent hubs can be deliberated. In this section, we simulate a snowball sampling crawl on the complete Cyworld network, and evaluate the estimation error in snowball sampling.

Here we randomly select 10 different seed nodes with $k > 100$ and get 10 sample networks starting from each seed. We set the sampling ratio to 0.33%, resulting in the final $40,000$ nodes in each sample network. Starting from seed nodes with smaller degrees does not make a difference, for within a few rounds of breadth-first search, a hub of $k > 100$

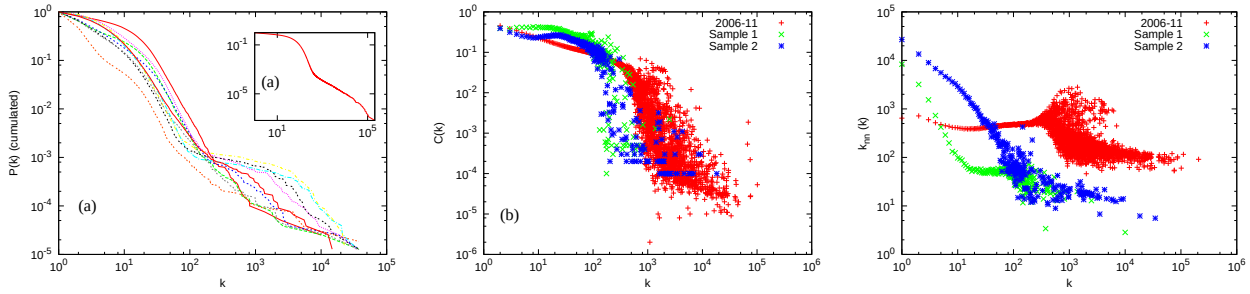

Fig. 7. Topological characteristics of sampled network

is reached, almost certainly as expected from the power-law degree distribution.

Figure 7(a) plots the CCDF of the degree distributions of all 10 sample networks of Cyworld. For a reference, we include the cumulative degree distribution of the complete Cyworld network as an inset. The cumulative degree distributions of sample networks have similar multi-scaling behaviors to that of the complete network: there are two regions of slow and rapid decaying. The exponents of the heavy tailed region range from 1.2 to 1.8. As observed in [10], the exponents of sample networks are supposed to be smaller than that of the complete network. We can summarize sample networks have varying smaller degree exponents than the complete network, but clearly the same multi-region scaling behavior.

Out of the above 10 sample networks, we choose two sample networks randomly and plot only the two in the rest of the analysis for convenience. We verified that including other sample networks did not change the qualitative conclusion in our analysis. Figure 7(b) depicts the distributions of clustering coefficient, $C(k)$, of the two sample networks and the complete network. Recall that the complete network had a very clean scaling behavior up to $k = 500$ with an exponent of 0.4. The limited scaling behavior in the complete network vanishes; moreover, the two sample networks exhibit no scaling behavior in their clustering coefficient distributions.

The average clustering coefficient of sample networks are larger than 0.16 of the complete network. The mean clustering coefficient of all 10 sample networks is 0.29. However, the converging behavior of the clustering coefficient (whether it increases or not) under the snowball sampling is not deterministic in general [10] and we cannot predict whether the original clustering coefficient of orkut and MySpace is larger than those from sample networks.

We plot the degree correlation distributions of the two sample and complete networks of Cyworld in Figure 7(c). The sample networks exhibit a more definite disassortative mixing pattern in their degree correlation distribution. The distributions from the two sample networks exhibit a clear decreasing pattern for $k < 100$ and then disperse.

In our preliminary work, we have evaluated how close topological characteristics of snowball sampled networks are to the complete network as we vary the sampling ratios [24]. From our numerical analysis, we suggest a practical guideline on the sampling ratio for accurate estimation of the topological metrics, excluding the clustering coefficient, where the explicit sampling ratio for accurate estimation is charted for the other metrics; 0.25% or larger for degree distribution, 0.2% or larger for degree correlation, and 0.9% or larger for assortativity. In the case of the clustering coefficient, even with a sampling ratio of 2%, it is inconclusive if the clustering coefficient of the sample network has converged close to that of the complete network.

In summary, we observe the same multi-scaling behavior in degree distribution from sample networks, while the exponents of the distributions are different from that of the complete network. Other metrics, such as the clustering coefficient distribution, its mixing pattern, and the degree correlation, are not easy to estimate from samples. Though not rigorously validated, we demonstrate that only with a sampling ratio above a certain threshold, the scaling behavior of the node degree distribution is captured correctly in sample networks.

D. Testimonial network

As described in Section III-A, a testimonial represents a directional relationship of intimacy and interest in Cyworld. Not all online friends leave testimonials on their friends' front pages. We view testimonials as an online manifestation of close off-line relationship, and expect it to have close resemblance to real-life social networks.

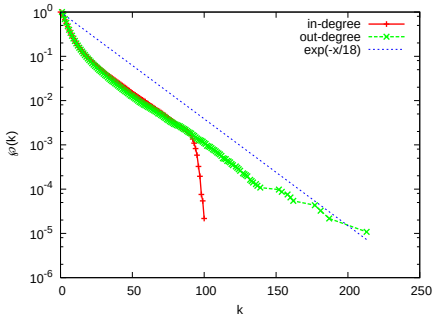

Fig. 8. Cumulative distribution of in-degree and out-degrees of Cyworld's testimonial network

We have collected a sample network of 100,000 nodes for our testimonial network analysis. In contrast to the complete Cyworld network, both the in-degree and out-degree distributions of the sampled network show an exponential distribution. There is a sharp cut-off in in-degree distribution around $k \sim 100$ which is due to the fact that the testimonial space is limited in the front page. However, the few data points

near $k \sim 200$ in the out-degree distribution indicates that some users leave extraordinarily many testimonials on others' front pages.

To evaluate other metrics, we need an undirected network. We consider each link in this testimonial network as bidirectional in the following analysis. The clustering coefficient of Cyworld testimonial network is 0.32. As discussed before, the clustering coefficient from a sample network is hard to evaluate and we only report the value of 0.4. However the distribution of clustering coefficients as shown in Figure 1(b) has a clear scaling behavior, and the degree correlation in Figure 1(c) demonstrates an assortative mixing, a sign of human social networking. We also estimate the average path length and it is about 7, larger than those of the friend relationship network shown in Figure 6.

In summary, the testimonial network of Cyworld is a subset of the complete Cyworld friend network. However, its topological characteristics deviate from the complete network: it has exponential degree distributions, clear positive degree correlation and rather a large average path length. The testimonial network represents more tight and close-knit relationships than the simple friends network. Accordingly, the testimonial network shows a positive degree correlation, which is a unique property of the social networks.

VI. Analysis of orkut and MySpace

Figure 9(a) shows the degree distributions of orkut and MySpace. The degree distribution of orkut exhibits clear power-law with degree exponent $\gamma \sim 3.7$. There is a cutoff around $k \sim 10^3$, which is due to orkut's old policy that the number of friends cannot exceed 1000. Note that the real degree exponent will be a little larger than 3.7 since the snowball sampling underestimate the degree exponent. The degree distribution of MySpace is not from the sample by the snowball sampling, but from the sample by the randome degree sampling, which visit a random person's homepage and sample the degree of the person. The size of the sample is $1,238,502$ and it covers more than ten times larger number of nodes than the snowball sample. Note that we use this data only for the degree distribution. The degree distribution of MySpace also shows power-law with degree exponent $\gamma \sim 2.1$, which agrees well with the exponent of heavy tail of Cyworld. Note that the random degree sampling is unbiased in contrast to the snowball sampling. We should also note that as in the sampling analysis of Cyworld, the sampling fraction of MySpace is not conclusive enough, thus, multi-scaling behavior might not be observed, even if it exists. Figure 9(b) shows the distributions of clustering coefficient. $C(k)$ of orkut shows a tendency to decrease, but it is not evident. Our former analysis on the samples of Cyworld suggests that the decay in the real $C(k)$ of orkut is more rapid than observed. The similar tendency is observed from the MySpace.

The decreasing tendency of the degree correlation of orkut in the small and large degree region shown in Figure 9(c) may come from the effect shown in Cyworld sampling or the invitation system of orkut. Until recently, users could register to orkut only when he was invited by a user in orkut. So, new users were likely to be connected to heavy users. However, it now permits users to create accounts without an invitation. Orkut shares the property of the positive assortative mixing with the testimonial network of Cyworld, which is considered as a close-knit community. In contrast, the MySpace network's assortativity is clearly negative, $r \sim -0.2$. The disassortative mixing tells us that MySpace largely deviates from the traditional social networks.

Since we only have data from a single seed, the estimation of degrees of separation for orkut and MySpace is much rougher than the case of Cyworld. In order to reduce the error, we let the number of first neighbors (n_1) as the mean degree we measured, and we tune the number of second neighbors (n_2) as to preserve the ratio of the number of first and second neighbors by assuming that the ratio (n_2/n_1) is not so sensitive to the n_1. Note that the value n_2/n_1 cannot fluctuate much due to the averaging effect, when the degree correlation is not strong. The calculated degrees of separation of orkut is 3.8, which is a little larger than that of Cyworld. The degrees of separation of MySpace is calculated in the same way with orkut and the value is 2.7, which is the smallest value among our data sets, while the size of myspace is the largest.

VII. Discussion

A. Comparison of Cyworld, MySpace, and orkut

We plot the three degree distributions of Cyworld, MySpace, and orkut together in Figure 9(a) for ease of comparison. In Figure 9(a), to compare the exponents more easily, the degree distribution graph of MySpace is rescaled and shifted to align with the heavy-tailed region of Cyworld while the range of both axes are identical to others. As seen previously, Cyworld has a multi-scaling behavior, while MySpace and orkut exhibit simple power-law.

From Figure 9(a), we observe an interesting relation between the three degree distributions. The rapid decaying behavior of orkut matches the rapid exponential decaying region of Cyworld, while the exponent of MySpace matches that of the heavy-tailed region of Cyworld. Since the linear region in the degree distribution of orkut span only a small region and the exponent is large, we cannot conclude whether the full network of orkut will have the power-law degree distribution or not. As shown in Section V-C, sample networks tend to have smaller exponents than the complete network and we circumspect that the real exponent of orkut be larger than 3.7 and it may be hard to determine whether it is a power-law or other distributions like exponential or log-normal. Meanwhile, The exponent of MySpace is $\gamma = 2.1$ and it perfectly matches that of Cyworld's heavy-tailed region, $\gamma = 2.1$. These facts gives us a hint about the peculiarity of Cyworld. The analysis of Cyworld reveals that there seems to be two heterogeneous networks mixed in Cyworld and the correspondence to orkut and MySpace may tell us that Cyworld's heterogeneity comes from the mixing of two types of users: one, with an orkut-like structure and the other with a MySpace-like structure.

Orkut is considered a relatively closed community, for a new user can register only by invitation. Relations in MySpace might be considered loose, as anyone can sign up without

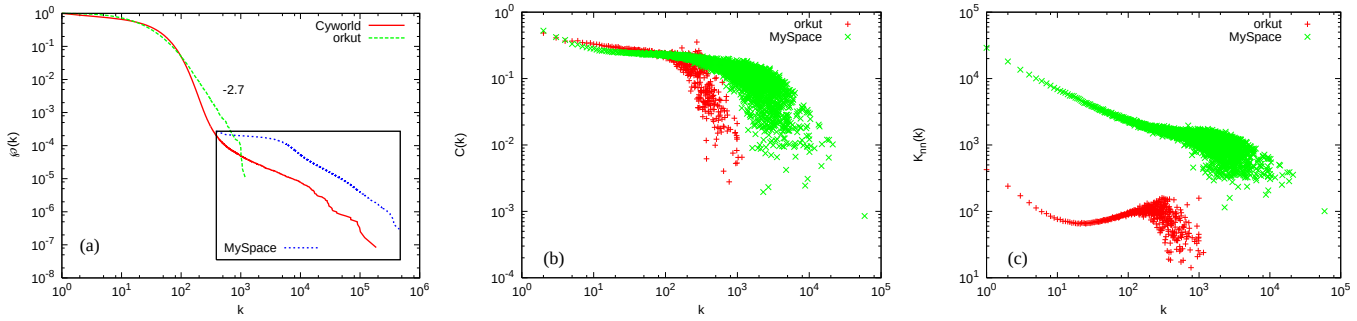

Fig. 9. Topological characteristics of three social networks: Cyworld friends network, orkut, and MySpace.

invitation and many are attracted to popularity, of which phenomenon stemming from early days as a popular site for the music community. The main webpage publishes "cool new people" and connecting to a new online friend is not a serious invitation-only affair. Cyworld has been a very private space in South Korea. By default, the list of friends is not made public, and neither are the photos and writings. On the other hand, Cyworld has also many features that encourage the creation of hubs. It picks two persons every day and puts their photos and link to their front pages at the main page. It is similar to the "cool new people" in MySpace. Moreover, Cyworld also has the club service that provides a homepage for various online communities. The organizer of a club tends to acquire many friends through the club activity.

Previous work has hinted at a multi-scaling behavior for online social networks. Holme et al. analyze the multitudes of layers of an Internet dating community: messages, guest books, and flirts. In all layers, the degree distribution is power-law, with the tail slightly heavier than normal [8]. Liben-Nowell et al. shows the degree distribution of a network of more than one million bloggers and it also has a short, but heavy tail [9]. Neither network has a reference network for comparison and its scaling behavior was not explicable. Our analysis of Cyworld is a confirmation that those previous revelations are not an isolated incident, but stem from a firm underlying structure in online social networks.

B. Dunbar's Law and Characteristic Degree of Cyworld

In Section V-A, we show that the degree distribution of Cyworld consists of exponential function and power-law function. The power-law function dominates in the tail region, while the exponential function dominates in small degree region. The exponential function has a characteristic scale, which turns out to be about 46. If we assume that the exponential decay is natural degree distribution for off-line like social network, it tells us that there is a certain limit in the number of friends. In sociology and anthropology, the theoretical limit in the number of social relationship is known as Dunbar's number and it is 150 [25]. Intriguingly, the characteristic number of friends, 46, is one third of Dunbar's number and the ratio coincides with the ratio between the number of Cyworld users and the South Korean population.

C. Origins of power-law behavior in online social networks

The origin of power-law has been extensively studied and there are many mechanisms that produce power-law distributions. Which of those best explains the origin of the power-law in online social networks? The best known mechanism to generate power-law distributions is preferential attachment. Not only the well-known Barabási-Albert model, but also many other mechanisms implicitly use preferential attachment. The *transitive linking* model [26], which is based on continuously completing triangles with only an edge missing, is one such example. Another noticeable viewpoint is fitness-based approaches. In any fitness-based approach, each node has its own fitness value and they are linked by the function of their fitness values [27], [28], [29]. In the case of the online social networks, both the preferential attachment and the fitness-based approach may contribute. More attractive and active persons are likely to have many online friends. Moreover, as one has more friends, it gets easier to have more friends through the transitive linking (a common friend of two persons introduces them to each other).

A real-life social relation is harder to maintain than the online counterpart. When online, you do not move to a new place nor spend much time to make new friends. It is a lot easier to leave a message to online friends than to meet a friend in real life. And as told about Cyworld users' behavior, most online relations are not severed, even when not active. Thus, the mechanisms that form the power-law degree distribution may not be severely blocked by the cognitive limit.

D. Online networks like real social networks

We conjecture that Cyworld's testimonial network is very similar to off-line social networks. First, strangers cannot write a testimonial , but only confirmed friends can. Even out of confirmed friends, not all are tempted or care enough to write a testimonial on a friend's front page for the rest of friends to view. Thus the testimonial network should be as an extremely close-knit community. Its close-knit nature is demonstrated by the exponentially decreasing degree distribution that decays more rapidly than power-law and has a definite cutoff. It also has an very clear assortative mixing pattern which is a major feature of social networks. We conclude that the testimonial network is the closest of the four online social networks to real-life social networks.

VIII. Conclusions

We have analyzed the complete network of an online social networking service, Cyworld. In addition, we have also analyzed sample networks from Cyworld, orkut, and MySpace in terms of degree distribution, clustering coefficient, degree correlation, and average path length.

We report a multi-scaling behavior in Cyworld's degree distribution and have substantiated our claim that heterogeneous types of users are the force behind the behavior with detailed analysis of the clustering coefficient distribution, assortativity (or disassortativity), and the historical evolution of the network size, the average path length, and the effective diameter. The observation that the scaling exponents of MySpace and orkut match those from different regions in the Cyworld network is also worthy of note.

IX. Acknowledgements

This work was supported by grant No. R01-2005-000- 1112-0 from the Basic Research Program of the Korea Science & Engineering Foundation and SK communications, Inc. We thank Jeongsu Hong and Jaehyun Lim of SK communications, Inc. for providing the Cyworld data.

References

[1] Techweb. http://www.techweb.com.

[2] Gueorgi Kossinets. Effects of missing data in social networks. Preprint, arXiv.org:cond-mat/0306335, 2003.

[3] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[4] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *SIGKDD*, 2005.

[5] Mark E. J. Newman. Scientific collaboration networks. I. network construction and fundamental results. *Phys. Rev. E*, 64:016131, July 2001.

[6] Fredrik Liljeros, Christofer R. Edling, Luis A. Nunes Amaral, H. Eugene Stanley, and Yvonne Aberg. The web of human sexual contacts. *Nature*, 411:907, 2001.

[7] Angelia Russo and Jerry Watkins. Digital cultural communication: Enabling new media and cocreation in southeast asia. *International Journal of Education and Development using Information and Communication Technology*, 1(4), 2005.

[8] Petter Holme, Christofer R. Edling, and Fredrik Liljeros. Structure and time-evolution of an internet dating community. *Social Networks*, 26:155, 2004.

[9] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic Routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, August 2005.

[10] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73:016102, 2006.

[11] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200, 2001.

[12] M. E. J. Newman. The spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, 2002.

[13] Reuven Cohen and Shlomo Havlin. Scale-free networks are ultrasmall. *Phys. Rev. Lett.*, 90:058701, 2003.

[14] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378, 2000.

[15] M. E. J. Newman. Power laws, pareto distributions and zipfs law. *Contemp. Phys.*, 46:323, 2005.

[16] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. Preprint, arXiv.org:0706.1062, 2007.

[17] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, 2002.

[18] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[19] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, 2003.

[20] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[21] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[22] C. Palmer, P. Gibbons, and C. Faloutsos. ANF: A fast and scalable tool for data mining in massive graphs. In *SIGKDD*, 2002.

[23] comscore. 694 million people currently use the internet worldwide according to comScore networks, 2006.

[24] Haewoon Kwak, Seungyeop Han, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Impact of snowball sampling ratios on network characteristics estimation: A case study of Cyworld. Technical Report CS-TR-2006-262, KAIST, 2006.

[25] R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

[26] Joern Davidsen, Holger Ebel, and Stefan Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88:128701, 2002.

[27] K. I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87:278701, 2001.

[28] G. Bianconi and A. L. Barabasi. Competition and multiscaling in evolving networks. *Europhys. Lett.*, 54:436, 2001.

[29] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Munoz. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.*, 89(25), December 2002.