# 스노우볼 샘플링 비율에 따른 네트워크의 특성 변화: 싸이월드의 사례 연구

곽해운[O°] 한승엽° 안용열† 문수복° 정하웅†
한국과학기술원 전산학과° 물리학과†
{haewoon[O], syhan}@an.kaist.ac.kr yongyeol@gmail.com sbmoon@cs.kaist.ac.kr
hjeong@kaist.ac.kr

# Impact of snowball sampling ratios on network characteristics estimation: A case study of Cyworld

Haewoon Kwak[O°], Seungyeop Han°, Yong-Yeol Ahn†, Sue Moon°, Hawoong Jeong†
Div. of Computer Science° Dept. of Physics†
KAIST

요 약

Today's social networking services have tens of millions of users, and are growing fast. Their sheer size poses a significant challenge in capturing and analyzing their topological characteristics. Snowball sampling is a popular method to crawl and sample network topologies, but requires a high sampling ratio for accurate estimation of certain metrics. In this work, we evaluate how close topological characteristics of snowball sampled networks are to the complete network. Instead of using a synthetically generated topology, we use the complete topology of Cyworld *ilchon* network. The goal of this work is to determine sampling ratios for accurate estimation of key topological characteristics, such as the degree distribution, the degree correlation, the assortativity, and the clustering coefficient.

## 1. Introduction

Social networking service (SNS) has been proliferating on the Internet and gaining popularity [1]. The number of users of MySpace exceeds more than 100 million users, and other popular SNSs like Orkut have also more than 10 million users[2][3]. Much research on SNS has been done [4][5]. However, the size of SNS is too large to handle, and most of the research has depended on network sampling methods.

Snowball sampling is one of the popular networking sampling methods, such as node sampling, link sampling, etc. All sampling methods take a part of the complete network by some rules. As the sampling ratio increases, the network more accurately describes the complete network. In other words, the higher sampling ratio, the more accurate estimation results one. However, for the high sampling ratio, more space and time are required. The trade-off lies between accuracy and resource.

Lee *et al.* has studied the impact of sampling ratios on parameter estimation for some network sampling methods, such as snowball sampling [6]. However, the size of the networks they have analyzed, including the Internet at the autonomous systems (AS) level, is two to three orders of magnitude smaller than today' s commercially available SNSs. Characteristics of a network topology are often not straightforward to capture by a metric or two, and the difficulty of accurate estimation lies in the diversity of metrics and their sensitivity to topology.

In this paper, we investigate the appropriate snowball sampling ratio for accurate estimation of topological characteristics, such as the degree distribution, the degree correlation, the assortativity, and the clustering coefficient. We compare the complete topology of the Cyworld *ilchon* network to sampled networks with varying sampling ratios. Our contributions lie in the methodological approach in the evaluation of the snowball sampling methods with Cyworld.

This Research was sponsored by SK Communications, Inc. and KOSEF Korea Science and Engineering Foundation through the Basic Research Program Grant No. R01-2005-000-11112-0 (2006)

## 2. Background

### 2.1 Cyworld

Cyworld is the largest online social networking service in South Korea [7], launched in September 2001. By the end of November 2005, the number of Cyworld users surpassed 12 million, 27% of the total population of South Korea. Like other SNSs, Cyworld provides the means to build and manage human relationships. In Cyworld, a member may invite other members to establish a relationship called *ilchon* between them simply by clicking the "add a friend" button. Then, if the invited member accepts it, an ilchon relationship is established. Online ilchon relationships reflect various kinds of real-world relationships: close friends and relatives, entertainment business workers and their fans, hobbyists of an interest, etc.

From a topological view, ilchon relationships form an undirected network. We get a graph structure by mapping a member into a node and an ilchon relationship into an edge. As an ilchon relationship is established with mutual acceptances, edges in an ilchon network are undirected. The number of nodes in an ilchon network thus represents the total number of Cyworld users, and the number of edges the number of established ilchon relationships.

### 2.2 Snowball Sampling Method

The process of snowball sampling is very similar to the breadth-first search. First, we randomly choose a *seed* node. Next, we visit all the nodes directly connected to the seed node. Here, we define a *layer* as the nodes which have the same hop count to reach a seed node. In other words, all nodes directly connected to the seed node are said to be on the first layer. Then, we get the second layer by visiting all nodes directly connected to the nodes on the first layer. This process continues until the number of visited nodes reaches a limit determined by the sampling ratio. In order to control the number of visited nodes, we stop sampling even if not all the nodes on the last layer are visited. By definition, *hub* nodes are connected to many nodes, and likely to be visited in few layers in snowball sampling. Snowball sampled networks have a bias towards hub nodes. While this bias leads overestimation of some metrics like degree distribution, whether the initial node is a hub or not does not make a noticeable difference in characterizing the sampled network [6].

Nevertheless, the snowball sampling method is the one of only a few available solutions when we crawl the web. When we crawl an SNS site for topology information gathering, we can only follow the links from a node each time. We do not have random access to the overall topology of the network, and cannot directly access nodes or links which are not adjacent to each other. This is the reason why the link and node sampling methods are infeasible in crawling.

## 3. Characteristics of Sampled Networks on Cyworld

In this section, we compare the degree distribution, degree correlation, and assortativity of the complete Cyworld network with those of sampled networks with various sampling ratios.

We use an anonymized snapshot of the entire ilchon network topology in November 2005. From this complete network topology, we construct sampled networks with varying sampling ratios. The sampling ratios vary from 0.1% to 1.0% in steps of 0.1%. Below the sampling ratio of 0.1% and beyond the sampling ratio of 1.0%, we use the following sampling ratios only: 0.01%, 0.05%, 0.25%, 0.75%, and 2.0%. For each sampling ratio, we take 5 sampled networks, each with a different seed in order to avoid any seed-dependent artifacts in analysis.

### 3. 1 Degree Distribution

The degree of a node is defined as the number of edges incident to the node. Many real networks, including online SNSs, are known to have a power-law degree distribution $p(k) \sim k^{-\gamma}$ [11][12][13], where $p(k)$ is the empirical probability of a node having degree k.

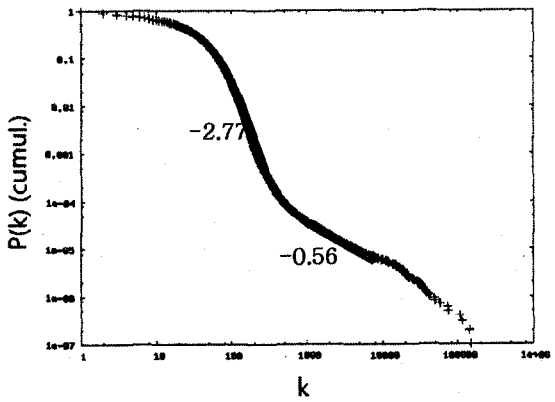


Fig. 1. Cumulative probability distribution of the complete Cyworld network

Figure 1 plots the cumulative degree distribution of the complete Cyworld ilchon network. The plot exhibits a multi-scaling behavior: the first part below the degree of 500 has an estimated exponent of -2.77 and the second part has -0.5. A well-known Hill estimator is used to estimate

the tail behavior of a distribution [14]. Crovella *et al.* have proposed an empirically sound method to estimate the scaling exponent of a heavy tail distribution [15]. However, both approaches are not directly applicable to distributions of multi-scaling behavior. We resort to manual identification of two different scaling regions and line-fitting in our exponent estimation.

We plot the degree distributions of sample networks with sampling ratios of 0.1%, 0.2%, 0.25%, and 0.5% in Figure 2. The multi-scaling behavior observed in the complete network appears only in 1 out of 5 sampled networks with the sampling ratio of 0.1%, 3 out of 5 with the sampling ratio of 0.2%, and in all 5 with the sampling ratio of 0.25% and 0.5%. Although one sampled network in Figure 2 (c) has a different scale exponent from other four sampled networks, it basically comes from the sampling methodology which takes a different part of the complete network.

From Figure 2, we observe a multi-scaling behavior from all five sampled networks when the sampling ratio is larger than 0.25%. In contrast to ratios below 0.25%, we note that the multi-scaling behavior is present in only one or two sampled networks. When the sampling ratio is 0.5% or higher, we observe the multi-scaling behavior with exponents of -1.56 and -0.57, different from -2.77 and -0.56 of the complete network.

It has been reported that the scaling exponent of snowball sampled networks are smaller than that of the complete network [6], and our results also present it. In addition, the scaling exponent of the part of high degree nodes is closer to that of the complete network than that of the part of low degree nodes. It is explained by the characteristics of a snowball sampling method. In snowball sampling, the nodes including part of high degree nodes are more selected than others, so the scaling exponent of the part of high degree nodes quickly approach that of the complete network.

The estimated scaling exponents approach those of the complete network only when the sampling ratio goes above 0.5%. To detect multi-scaling behavior from sampled networks from any sampled network, at least 0.25% or higher sampling ratio is required. For scaling exponent values within the range of 1 from the those of the complete network, 0.5% sampling ratio is needed.

The multi-scaling behavior has been observed in Internet backbone traffic [16][17], but not in any online networks, such as bloggers' networks and other SNSs like MySpace [2]. To our best knowledge, the Cyworld ilchon network is the first to exhibit this kind of behavior among SNSs. We suspect heterogeneous makeup of Cyworld users to be attributable to this multi-scaling behavior, as one-time commercial events and famous entertainers attract anomalously large numbers of users than most friends and family relationships. However, further work is needed to confirm our conjecture and explicate the causes.

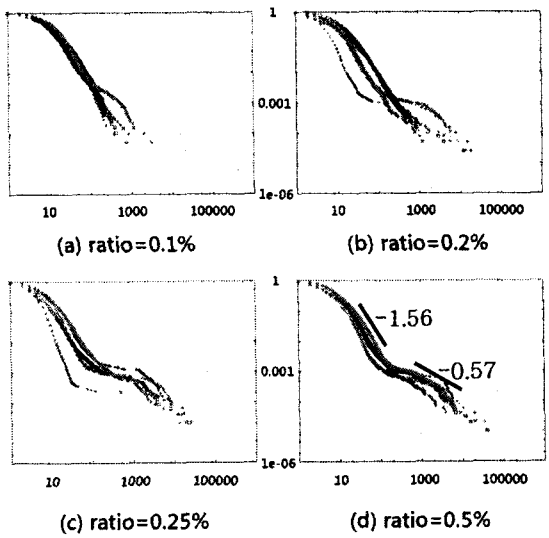


Fig. 2. Cumulative probability distribution from sampled networks of Cyworld.

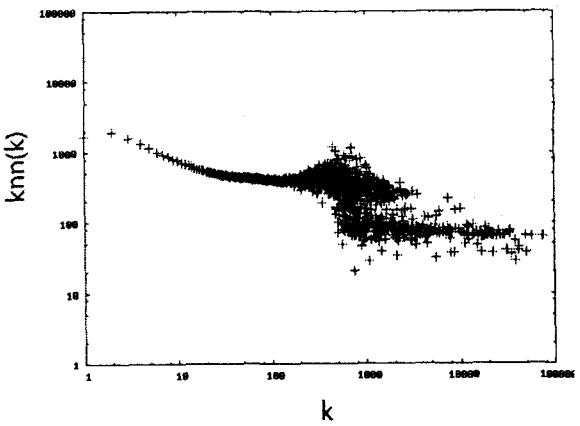


Fig. 3. Degree correlation from complete network

### 3.2 Degree Correlation

The degree correlation is a mapping between a node degree *k* and the mean degree of neighbors of nodes having degree *k*.

We plot the degree correlation of the complete network in Figure 3. Although complex pattern emerges, It is clear that the overall correlation is negative. It is known that if the trend of a degree correlation of a network is negative, a hub node tends to be connected non-hub nodes [18]. Therefore, the plot in Figure 3 tells that the hub nodes of Cyworld is more likely to be connected to non-hub nodes than to hub nodes.

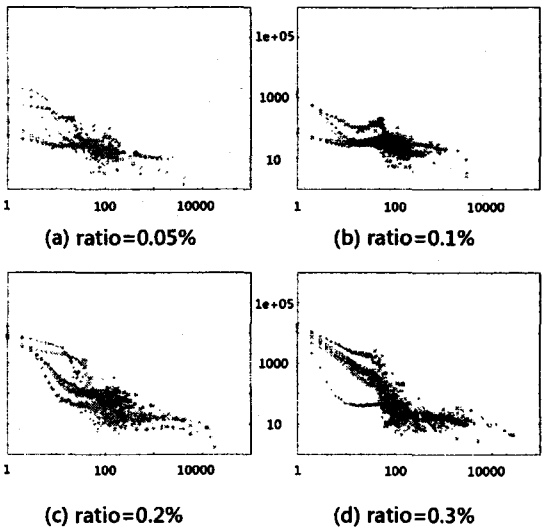


Fig. 4. Degree correlation from sampled networks of Cyworld.

We plot the degree correlation of sampled networks with sampling ratio of 0.05%, 0.1%, 0.2%, and 0.3% in Figure 4. All plots show complex patterns, and it is very hard to tell which is the closest to Figure 3. In close examination of Figures 4(a) and 4(b), we observe that some graphs do not have a negative trend for $1 < k < 100$.

Thus, from this sampling experiment, we recommend a sampling ratio of 0.2% or higher for degree correlation estimation.

### 3.3 Assortativity

Assortativity is summarized the distribution of degree correlation. Like overall trend of degree correlation in Section 3.2, the value of assortativity is negative, a hub node tends to be connected non-hub nodes. The assortativity $r$, can be written as below,

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

where $j_i$ and $k_i$ are the degrees of the vertices at the ends of the $i$th edge, with $i = 1, 2, \cdots$, and $M$. $M$ is the number of all edges [18].

Positive values of $r$ exhibit a positive degree correlation. It means that a hub node tends to connect to other hub nodes. In other words, negative value stands for the negative degree correlation.

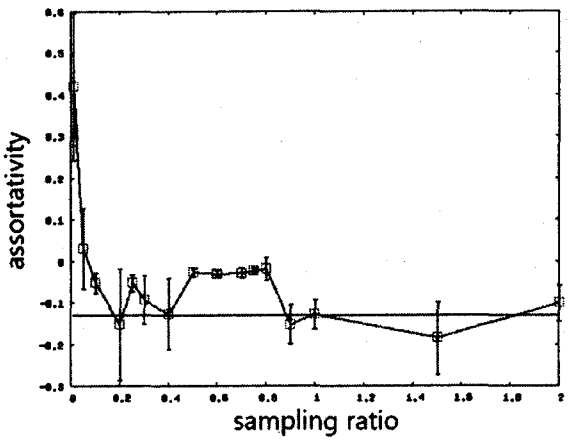


Fig. 5. Assortativity of sampled networks of Cyworld. The red line is from complete network of Cyworld.

Figure 5 presents that changes in assorativity for sampled networks. The horizontal line is the assortativity of the complete network of Cyworld. At the sampling ratio larger than 0.9%, the assorativity value of complete network falls in the range of sampled networks.

In [6], sampled networks from snowball sampling were shown to be more disassortative than the original network whether the original network is assortative or disassortative, but our results are very different. We cannot find some evidences which support that sampled networks are more disassortative than the original network in our results. It seems that the assortativity of sampled networks has nothing to do with that of the complete network. We are not sure why disassortative and assortative sampled networks are mixed up in our results.

### 3.4 Clustering Coefficient

The clustering coefficient $C_i$ of node $i$ is the ratio of the total number of the existing links between its neighbors, y, to the total number of all possible links between its neighbors [13], and can be written as below,

$$C_i = \frac{2y}{k_i(k_i - 1)}, \quad (2)$$

where $k_i$ is the degree of node $i$.

Figure 6 depicts average clustering coefficients of the sampled Cyworld networks with various sampling ratio. The minimum sampling ratio is 0.15%, and the maximum sampling ratio is 2.0%. We observe a gap between the

sampled network and complete network closes as the sampling ratio increases.

In addition, all clustering coefficient values of sampled networks are larger than that of the complete network. It means that the sampled networks from snowball sampling are more likely to be clustered than the complete network.

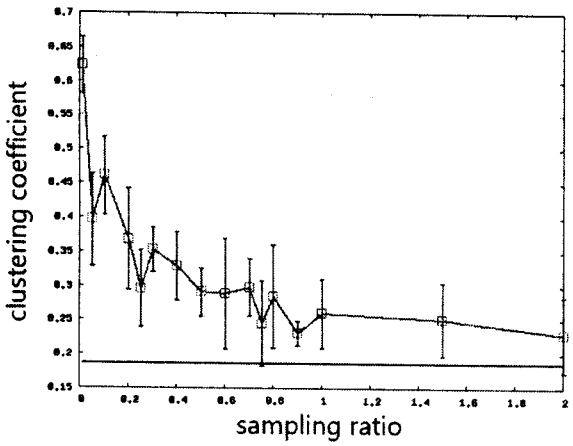


Fig. 6. Average clustering coefficients of sampled networks of Cyworld. The red line is from complete network of Cyworld

## 4. Discussion

We have described the characteristics of sampled networks derived from the Cyworld ilchon network. From the results, we could get some insights on snowball sampling. However, the results in this paper come from only one real network, Cyworld, and may not apply to other networks. To generalize our results, we need to investigate if Cyworld is a representative of other social networks. In future, we plan to compare other SNSs with Cyworld. A simple qualitative analysis gave us some intuition about this problem. In Section 3.1, Cyworld has two kinds of ilchon networks corresponding to different scaling behavior. From the relationship between members perspective, there are two kinds of ilchon in Cyworld. We conjecture that one reflects offline friendship, the other does online community. Except the difference of cultures of different countries, basically other SNSs support the similar feature with ilchon called *friend*, etc. Therefore, we expect that the friend network would be similar to ilchon network of Cyworld, because the members of other SNSs also make their offline friends of friends in online, and the other SNSs support many features to meet attractive member online, for example, conditional search and form a new online community. Therefore, members in other SNSs also have on-line and off-line friends. This is the reason why we conjecture our results can be generalized.

## 5. Summary

In this work, we have studied the appropriate snowball sampling ratio of real social networking service, Cyworld. For good representation of multi-scaling behavior the degree distribution, we need a sampling ratio of 0.25% or larger. For accurate estimation of a negative trend of the degree correlation from all sampled networks, the sampling ratio of 0.2% or larger is required. For precise estimation of assortativity, we need a sampling ratio of 0.9% or larger. Finally, we can observe a gap between the sampled network and complete network closes as the sampling ratio increases for estimation of clustering coefficient.

## Acknowledgements

We thank Jeongsu Hong and Jaehyun Lim of SK Communications, Inc. for providing us with the Cyworld data.

## References

[1] Tech web, http://www.techweb.com
[2] MySpace, http://www.myspace.com
[3] Orkut, http://www.orkut.com
[4] W. Bachnik, S. Szymczyk, P. Leszczynski, R. Podsiadlo, E. Rymszewicz, L. Kurylo, D. Makowiec, and B. Bykowska, Quantitative and sociological analysis of blog networks, Acta Physica Polonica B 36, 2435 (2005)
[5] S. Wasserman and K. Faust, Social Network Analysis, Cambridge University Press, Cambridge (1994)
[6] S. H. Lee, P.-J. Kim, and H. Jeong, Statistical properties of sampled networks, Physical Review. E 73. 016102 (2006)
[7] Rankey.com, http://www.rankey.com
[8] Steven K. Thomson, Sampling (2003)
[9] M. E. J. Newman, Soc. Networks 25, 83 (2003)
[10] Wikipedia, http://www.wikipedia.org
[11] M. E. J. Newman, SIAM Review. 45, 167 (2003)
[12] R. Albert and A-L. Barabasi, Reviews of Modern Physics. 74, 47 (2002)
[13] S. N. Dorogovtsev and J. F. F. Mendes, Advances in Physics. 51, 1079 (2002)
[14] B. M. Hill, A Simple general approach to inference about the tail of a distribution, The Annals of Statistics, 3:1163-1174, 1975
[15] M. E. Crovella and M. S. Taqqu, In Methodology and Computing in Applied Probability, Vol 1 No. 1 (1999)
[16] A. Feldmann, A. C. Gilbert, and W. Willinger, Data networks as cascades: Investigating the multifractal nature of internet wan traffic, ACM/Sigcomm'98 (1998)
[17] Z. - L. Zhang, V. Ribeiro, S. Moon, C. Diot, Small-Time scaling behaviors of internet backbone traffic: An empirical study, INFOCOM 2003 (2003)
[18] M. E. J. Newman, Physical Review Letter. 89, 208701 (2002)