Journal of the Korean Physical Society, Vol. 56, No. 3, March 2010, pp. 990∼993

Biased Random Walk Sampling on Assortative Networks

Soon-Hyung Yook,* Yeo-kwang Yun and Yup Kim†
Department of Physics and Research Institute for Basic Sciences, Kyung Hee University, Seoul 130-701

(Received 15 September 2009, in final form 5 November 2009)

We study the statistical properties of sampled networks by using a biased random walker on assortative networks. In the biased random walk sampling, all the nodes visited by the biased random walker and the links that connect any pair of visited nodes are sampled. Here, the probability that a walker moves to one of its nearest neighbor depends on the degrees of the nearest neighbors. We compare the topological properties, such as the degree distribution, the degree-degree correlation, and the clustering coefficient of the sampled networks with those of the original networks. From the numerical results, we find that most of the topological properties of the sampled networks by the biased random walk are almost the same as those of the original networks when the network is assortative. Moreover, from the measurement of the clustering coefficient, we find that the hierarchical structures are better inherited through a biased random walk sampling when the network is highly assortative.

PACS numbers: 05.40.Fb, 89.75.Hc, 89.75.Fb
Keywords: Random walk, Complex networks
DOI: 10.3938/jkps.56.990

I. INTRODUCTION

Complex networks [1] are ubiquitous in real world. Examples of such studies include protein-protein interaction networks (PIN) [2], the world-wide web (WWW) [3], email network [4], *etc.* Empirical data or information of real networks are collected in various ways; for example, the traceroutes for the Internet [5] and high throughput experiments for protein interaction map [6]. Thus, it is natural that the empirical data should be incomplete for various reasons which come from some limitations of the experiments and experimental errors or biases. Therefore, it is very important to validate if the topological properties of the sampled networks are identical to those of the entire network or not.

Recently, several sampling methods, such as random sampling [7,8] and snowball sampling [8], were studied. Random sampling is the simplest method in which the sampled network consists of randomly selected nodes or links with a given probability p. A well known example of random sampling is a statistical survey in some social systems. In random sampling, however, many important nodes, such as hubs, are not sampled due to the even selection probability. In the snowball sampling method, all nodes directly connected to the randomly chosen starting node are selected. Then, all the nodes linked to those selected nodes in the last step are selected hierarchically. This process continues until the sampled network has the desired number of nodes [8]. Previous studies showed that the topological properties of the sampled networks strongly depend on the sampling methods [8].

More recently, we studied a random walk sampling [9] by assuming that the probability to sample a node i with degree k_i is proportional to k_i. Using the uncorrelated theoretical networks and several real networks, we showed that the random walk sampling reflected the topological properties of the original networks much better than random sampling and snowball sampling [9]. However, many real networks have degree-degree correlation [10] and the degree-degree correlation is known to affect various phenomena. For example, the percolation transition in correlated networks shows a nontrival behavior when degree-degree correlation is positive (assortative) [11]. Moreover, some networks, such as PIN, have revealed that nodes of large degree are more studied [9], which reflects some bias in network samplings. However, there have not yet been any systematic studies on the explicit relationship between the degree and the bias. Like the percolation on assortative networks, in the sampling of networks using a random walker, assortative mixing can accelerate the sampling of hubs and the nodes connected to them, which causes some non-trivial sampling. Therefore, in this paper, we study the biased random walk sampling method (BRWSM) on assortative networks for a more systematic approach to the relationship between the assortativity (positive degree-degree correlation) and generalized random walk sampling method. Using numerical simulations, we show that the sampled

*E-mail: syook@khu.ac.kr
†E-mail: ykim@khu.ac.kr

network through the biased random walk can preserve its original topology much better than an unbiased random walk sampling when there is a positive degree-degree correlation. Especially, from the measurement of the clustering coefficient, we find that if there is any hierarchical structure in the original network, then BRWSM can preserve the hierarchial structure much better than other sampling methods. We also discuss some possible hazards in the analysis of real network data that can be arise from the relationship between the BRWSM and the degree-degree correlation. As a result, we expect this study to provide better insight into understanding the important properties of real networks and to offer a systematic approach to understanding real network data.

The paper is organized as follows: In section II, we introduce the network models and BRWSM. In Section III we present the simulation results. Summary and discussion are given in Section IV.

II. MODEL

1. Degree-degree Correlation

Since many real networks are known to be scale-free (SF) networks [1], we only consider SF networks in this study. A SF network is characterized by a power-law degree distribution, $P(k) \sim k^{-\gamma}$. In order to generate the original SF network of size N_o, we use the static model suggested by Goh *et al.* [12]. In the static model, a weight $w_i = i^{-\sigma}$ is assigned to each node i ($i = 1, 2, \cdots, N_o$ and $0 \leq \sigma < 1$). By adding a link between unconnected nodes i and j with probability $w_i w_j / (\sum_{n=1}^{N} w_n)^2$, one can obtain a SF network. The value of γ of the resulting SF network is known to satisfy the relation $\gamma = (1+\sigma)/\sigma$. Thus, by adjusting σ, we obtain a network with any γ (> 2).

When there are M edges, the degree-degree correlation is generally measured by using the Pearson coefficient. The Pearson coefficient is defined as [10],

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

where j_i and k_i are the degrees of the vertices at the ends of the ith edge, with $i = 1, \cdots, M$. When $r > 0$ ($r < 0$) the network is said to be assortative (disassortative). For the systematic generation of correlated networks, we use the edge exchange method [13]. In this edge exchange method, two randomly selected edges are shuffled with probability p. The degree-degree correlation of the original network can be adjusted by choosing a proper value of p [13]. For $\gamma < 3$, it is not easy to obtain an assortative network. Thus, in most of the following simulations, we consider SF networks with $\gamma > 3$.

2. Biased Random Walk Sampling Method

We now introduce the BRWSM. After the preparation of the original network, a walker is placed at a randomly chosen node. At each time step, the walker takes a biased random walk. The probability that a walker at a node i moves to one of its nearest neighbors, j, is given by

$$P_{ij} = \frac{{k_j}^{\alpha}}{\sum_{l \in \Gamma_i} {k_l}^{\alpha}}, \quad (2)$$

where k_j is the degree of node j and Γ_i represents the set of i's nearest neighbors. α controls the degree of bias. The walker moves until it visits N_s distinct nodes. Then, we construct subnetworks with these N_s visited nodes and the links which connect any pair of nodes among the N_s visited nodes in the original network. In the biased random walk, the probability to find a walker at a node of degree k is [14]

$$P_{BRW}(k) \sim k^{\alpha+1}. \quad (3)$$

From Eq. (3), we expect that if $\alpha > -1$ and the network is assortative then the core region which consists of nodes of large degrees is more easily sampled by using the BRWSM than other sampling methods. Thus, the network sampled by using the BRWSM is expected to preserve the topological properties of the original network.

III. NUMERICAL SIMULATIONS

1. Degree Distribution

The degree distribution is one of the most important measures for the heterogeneity in the network topology [1]. In Fig. 1, we compare the degree distributions between the sampled networks to those of the original networks with $\gamma = 4.5$ for various α and the Pearson coefficient of the original network, $r(N_o)$. We find that the degree distribution of the sampled network also satisfies the power-law. However, if α or $r(N_o)$ is small (Figs. 1(a)-(c)), then γ of the sampled network noticeably deviates from that of the original network, $\gamma = 4.5$. On the other hand, when both $r(N_o)$ and α are large (see Fig. 1(d)), γ of the sampled network remains the same as that of the original network. In Fig. 1(e), we summarize the change in $\gamma(N_s)$ for various values of $r(N_o)$ and α. The result indicates that when the assortativity of an original network increases, then the BRWSM with large α preserves its degree distribution much better.

2. Degree-degree Correlation

The degree-degree correlation can also be measured from the average degree of the nearest neighbors of nodes

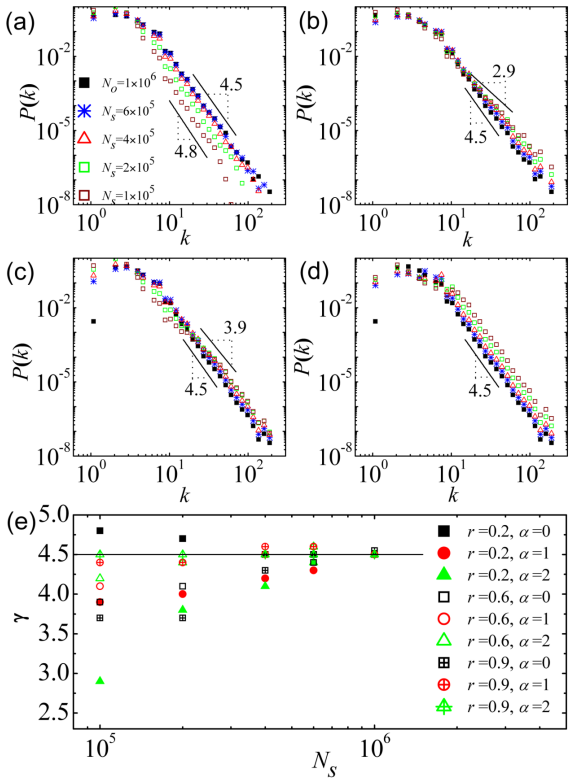

Fig. 1. Plot of $P(k)$ for a small value of $r(N_o)(= 0.2)$ with (a) $\alpha = 0$ and (b) $\alpha = 2$. Plot of $P(k)$ for large $r(N_o)(= 0.9)$ with (c) $\alpha = 0$ and (d) $\alpha = 2$. (e) Plot of $\gamma(N_s)$ for various $r(N_o)$ and α. The solid line in (e) represents γ of the original network, $\gamma(N_o) = 4.5$.

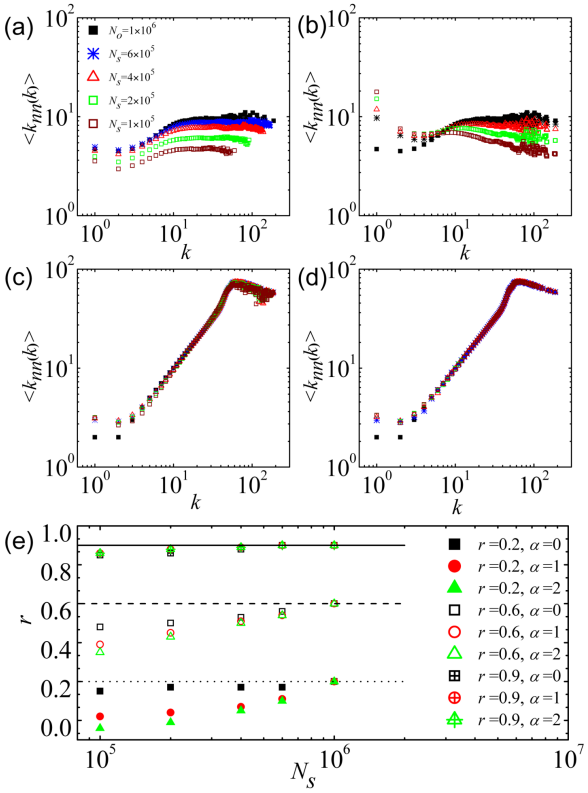

Fig. 2. Plot of $\langle k_{nn}(k)\rangle$ for a small value of $r(N_o)(= 0.2)$ with (a) $\alpha = 0$ and (b) $\alpha = 2$. Plot of $\langle k_{nn}(k)\rangle$ for large $r(N_o)(= 0.9)$ with (c) $\alpha = 0$ and (d) $\alpha = 2$. (e) Plot of $r(N_s)$ for various values of α and $r(N_o)$. The Pearson coefficient of the original networks, $r(N_o) = 0.9$, $r(N_o) = 0.6$, and $r(N_o) = 0.2$, are denoted by a solid line, a dashed line, and a dotted line, respectively.

with degree k, $\langle k_{nn}(k)\rangle$ [15]. When the network is neutral, $\langle k_{nn}(k)\rangle$ does not depend on k. On the other hand, if the network is assortative (disassortative), then $\langle k_{nn}(k)\rangle$ increases (decreases) as k increases. Figs. 2(a) and (b) show $\langle k_{nn}(k)\rangle$'s for $\alpha = 0$ and $\alpha = 2$, respectively, when $r(N_o) = 0.2$. As shown in Figs. 2(a) and (b), the sampled network shows two different behaviors depending on α. When $\alpha = 0$, although the value of $\langle k_{nn}(k)\rangle$ for $r(N_o) = 0.2$ decreases as N_s decreases, its functional dependency on k is not changed from that of the original network. On the other hand, if we increase α, then $\langle k_{nn}(k)\rangle$ of the sampled network becomes a decreasing function of k even for $N_s = 0.6N_o$ (Fig. 2(b)). The value of $\langle k_{nn}(k)\rangle$ for large $r(N_o)$ are also displayed in Figs. 2(c) and (d). From the data, we find that $\langle k_{nn}(k)\rangle$ is almost the same as that of the original network for $\alpha \geq 0$. For a more detailed study, we also measure the Pearson coefficient of the sampled network with N_s nodes, $r(N_s)$, as shown in Fig. 2(e). From the data in Fig. 2(e), we find that $r(N_s)$ shows the same behavior as those for the analysis of $\langle k_{nn}(k)\rangle$ when the original network is weakly assortative ($r(N_o) = 0.2$) or highly assortative ($r(N_o) = 0.9$). For an intermediate value of $r(N_o)(= 0.6)$, we find that $r(N_s)$ significantly deviates from $r(N_o)$ for any $\alpha(> 0)$.

3. Clustering Coefficient

The clustering coefficient C_i of a node i is defined by

$$C_i = \frac{2y_i}{k_i(k_i - 1)} , \qquad (4)$$

where k_i is the degree of node i and y_i is the number of connections between its nearest neighbors [1]. C_i physically means the fraction of connected pairs among pairs of node i's neighbors. C_i is one if all neighbors are completely connected whereas C_i becomes zero on an infinite-sized random network [1]. By averaging C_i's over the same k, we obtain $C(k)$. The measurement of $C(k)$ is also important because $C(k)$ is known to reflect the hierarchical modular structure of networks [16]; $C(k)$ does not depend on k if the network does not have any well-defined hierarchical modules [16]. In Fig. 3, we compare the $C(k)$'s for the original network and the sampled networks. From the data in Fig. 3(a), we find that, if both

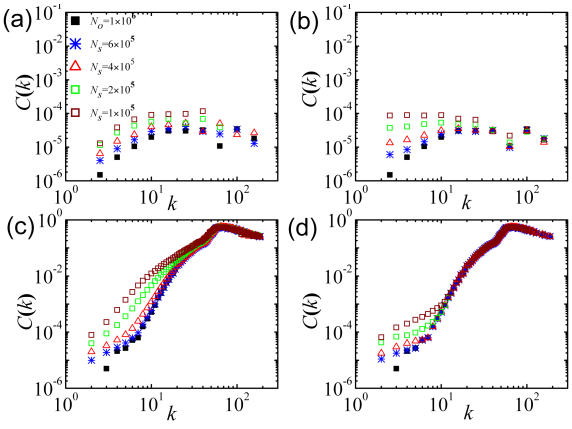

Fig. 3. Plot of $C(k)$ for a small value of $r(N_o)(= 0.2)$ with (a) $\alpha = 0$ and (b) $\alpha = 2$. Plot of $C(k)$ for large $r(N_o)(= 0.9)$ with (c) $\alpha = 0$ and (d) $\alpha = 2$.

$r(N_o)$ are small ($r(N_o) = 0.2$ and $\alpha = 0$), the functional form of $C(k)$ of the sampled networks is relatively well preserved; $C(k)$ increases for $k < 30$ and decreases when $k > 30$. However, if α increases ($\alpha = 2$), then the functional form of $C(k)$ noticeably deviates from that of the original networks when $N_s < 0.2N_o$ (Fig. 3(b)). The α-dependent behavior of $C(k)$ for large $r(N_o)$ is completely different from that for small $r(N_o)$. As shown in Figs. 3(c) and (d), we find that the value of $C(k)$ of the sampled networks coincides with that of the original networks when α increases. This result implies that if the network is highly assortative and has any kind of hierarchical structure, then only the BRWSM can preserve its hierarchical structure.

IV. CONCLUSION AND DISCUSSION

We study the topological properties of sampled networks by using the BRWSM with assortative SF networks. From the numerical simulations, we find that the $P(k)$ of the sampled network follows the power-law $P(k) \sim k^{-\gamma}$, even for a relatively small $N_s(= 0.1N_o)$. We also show that the BRWSM with large α preserves the topological properties, such as the degree distribution, the degree-degree correlation, and the hierarchical structure, of an original network very well when the original network is highly assortative.

Based on our measurement, we now address some remarks on a possible hazard in the analysis of real data. From the measurement of $\langle k_{nn}(k)\rangle$ and $r(N_s)$, we find that if the network has a relatively weak degree-degree correlation ($r(N_o) < 0.6$), then the BRWSM can decrease the degree-degree correlation, and sometimes sampled networks can show a disassortative mixing (for example, see Fig. 2(b)). This indicates that the local topology around the hubs of the sampled networks becomes star-like. Therefore, if the sampling methods are highly biased, for example, the traceroutes [5], and the degree-degree correlation of an original network is close to neutral, then the degree-degree correlation of the sampled networks can be significantly changed by the sampling process. Therefore, in this case, a very careful investigation of the topological properties of the obtained networks is necessary.

ACKNOWLEDGMENTS

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government(2009-0073939), by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (Grant No. 2009-0052659), and by the Korea Research Foundation grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund)(KRF-2007-313-C00279).

REFERENCES

[1] R. Albert and A. -L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[2] S. H. Yook, Z. Oltvai and A.-L. Barabási, Proteomics **4**, 928 (2003).
[3] R. Albert, H. Jeong and A.-L. Barabási, Nature **401**, 130 (1999).
[4] H. Ebel, L. -I. Mielsch and S. Bornholdt, Phys. Rev. E **66**, 035103(R) (2003).
[5] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez and A. Vespignani, Phys. Rev. E **71**, 036135 (2005); A. Clauset and C. Moore, Phys. Rev. Lett. **94**, 018701 (2005).
[6] P. Uetz *et al.*, Nature **403**, 623 (2000).
[7] M. P. H. Stumpt and C. Wiuf, Phys. Rev. E **72**, 036118 (2005).
[8] S. H. Lee, P.-J. Kim and H. Jeong, Phys. Rev. E **73**, 016102 (2006).
[9] S. Yoon, S. Lee, S.-H. Yook and Y. Kim Phys. Rev. E **75** 046114 (2007).
[10] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[11] S.-W. Kim and J. D. Noh, J. Korean Phys. Soc. **52**, S145 (2008).
[12] K.-I. Goh, B. Kahng and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[13] R. Xulvi-Brunet and I. M. Sokolov, Phys. Rev. E **70**, 066102 (2004).
[14] S. Kwon, S. Yoon and Y. Kim, Phys. Rev. E **77**, 066105 (2008).
[15] R. Pastor-Satorras, A. Vázquez and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[16] R. Pastor-Satorras, A. Vázquez and A. Vespignani, Phys. Rev. E **65**, 066130 (2002); Z. E. Ravasz and A. -L. Barabási, Phys. Rev. E **67**, 026112 (2003); A. Vázquez, Phys. Rev. E **67**, 056104 (2003); J. -S. Lee, K. -I. Goh, B. Kahng and D. Kim, Eur. Phys. J. B **49**, 231 (2006).