Statistical properties of sampled networks by random walks

Sooyeon Yoon, Sungmin Lee, Soon-Hyung Yook,* and Yup Kim†
Department of Physics and Research Institute for Basic Sciences, Kyung Hee University, Seoul 130-701, Korea
(Received 4 December 2006; published 25 April 2007)

We study the statistical properties of the sampled networks by a random walker. We compare topological properties of the sampled networks such as degree distribution, degree-degree correlation, and clustering coefficient with those of the original networks. From the numerical results, we find that most of topological properties of the sampled networks are almost the same as those of the original networks for $\gamma \lesssim 3$. In contrast, we find that the degree distribution exponent of the sampled networks for $\gamma > 3$ somewhat deviates from that of the original networks when the ratio of the sampled network size to the original network size becomes smaller. We also apply the sampling method to various real networks such as collaboration of movie actor, Worldwide Web, and peer-to-peer networks. All topological properties of the sampled networks are essentially the same as those of the original real networks.

DOI: 10.1103/PhysRevE.75.046114 PACS number(s): 89.75.Fb, 05.40.Fb, 89.75.Hc

I. INTRODUCTION

Since the concept of complex network [1] came into the limelight, many physically meaningful analyses for the complex networks in the real world have emerged. Examples of such studies include protein-protein interaction networks (PINs) [2], the Worldwide Web (WWW) [3], email networks [4], etc. The empirical data or information of the real networks can be collected in various ways—for example, trace routes for the Internet [5] and high-throughput experiments for the protein interaction map [6]. Thus, it is a natural assumption that the empirical data can be incomplete due to various reasons which include some limitations on the experiments and experimental errors or biases. As a result, many real networks which have been intensively studied so far can be regarded as sampled networks. Moreover, several studies have shown that the dynamical properties on the networks can be significantly affected by the underlying topology [7,8]. Therefore, it is very important and interesting to study the topological differences between sampled networks and whole networks.

Recently, several sampling methods such as random node sampling [9,10], random link sampling, and snowball sampling were studied [10]. Random node sampling is the simplest method in which the sampled network consists of randomly selected nodes with a given probability p and the original links between the selected nodes. On the other hand, in random link sampling, the links are randomly selected and the nodes connected to the selected links are kept. These two random sampling methods have been used to study the statistical survey in some social networks. In the random sampling method, however, many important nodes such as hubs cannot be sampled due to the even selection probability. Some recent studies show that in some networks such as PINs, the topological properties of randomly sampled networks significantly deviate from those of the original networks [9,10]. The idea of the snowball sampling method [10,11] is similar to the breath-first search algorithm [12,13]. In the snowball sampling method all nodes directly connected to the randomly chosen starting node are selected. Then all 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 [10]. Previous studies showed that the topological properties of the sampled networks closely depend on the sampling methods [10].

In this paper, we focus on the effect of weighted sampling on the topological properties of sampled networks. For example, the average number of references related to the most connected 20 proteins is almost 2 times larger than that related to the least connected 20 proteins in the physical interaction database of PINs [14]. This can reflect the fact that the interactions of the nodes with the large degree are more studied. Thus, the PIN can be regarded as a kind of sampled network with the degree-dependent weight. As the simplest assumption, we consider that the probability to sample a node i with degree k_i is proportional to k_i. In order to assign a degree-dependent nontrivial weight to each node, we first note the structure of the real networks. Many real networks are known to be scale-free networks in which the degree distribution follows the power law [1]

$$P(k) \sim k^{-\gamma}. \tag{1}$$

Moreover, the probability $p_v(k)$ that a random walker (RW) visits a node of degree k [7] is given by

$$p_v(k) \sim k. \tag{2}$$

The degree k causes an uneven probability of finding a node by a RW on the heterogeneous networks. Thus, by using the RW for sampling we can assign automatically a nontrivial weight to each node which is proportional to the degree of the node. Due to the uneven visiting probability, the nodes having the large degree—i.e., topologically important parts—can be easily found regardless of the starting position of the RW. Therefore, we expect that the sampling by the RW can provide a more effective way to obtain subnetworks which have almost the same statistical properties as the origi-

*Electronic address: syook@khu.ac.kr
†Electronic address: ykim@khu.ac.kr

nal one. Furthermore, we study the effects of the heterogeneity of the original networks on the RW sampling method (RWSM) by changing γ. We also apply this weighted sampling method to real networks such as the WWW [3,16], actor network [17], and peer-to-peer (P2P) network [18,19] to obtain the important information of those networks. Therefore, we expect that this study can provide a better insight to understand the important properties of real networks and offer a systematic approach to the sampling of networks with various γ.

II. MODEL

We now introduce the RWSM. First, we generate original scale-free networks (SFNs) by use of the static model in Ref. [15] from which various sizes of subnetworks are sampled. The size or number N_o of nodes of the original network with each γ is set to be $N_o=10^6$. The typical values of γ used in the simulations are $\gamma=2.23$, 2.51, 3.05, 3.40, and 4.2. We set the average degree $\langle k\rangle=4$ for each network. After preparation of the original networks, a RW is placed at a randomly chosen node and 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. We use $N_s=10^3$, 10^4, 2×10^4, 4×10^4, 6×10^4, 8×10^4,10^5, and $1,2,3,\ldots,9\times10^5$. For the sampling of the real network, we use the WWW, actor network, and Gnutella snapshot provided by Refs. [3,17,18], as the original networks. Then, we apply the above procedure.

III. NUMERICAL SIMULATIONS

A. Degree distribution

The degree distribution is one of the most important measures for the heterogeneity of networks [1]. In Fig. 1, we compare the degree distributions of the sampled networks to those of the original networks for various γ. We find that the degree distribution of the sampled network also satisfies the power law $P(k)\sim k^{-\gamma_s}$.

Especially, from the data in Figs. 1(a)–1(d) we find that the γ_s of the sampled networks with $N_s/N_o\geq0.01$ is nearly equal to γ of the original network, even though the γ_s for the small N_s has a relatively larger error bar. It shows that the sampling method by RW does not change the heterogeneity in degree for networks with $2<\gamma\lesssim3$. Since most of the real networks have $2<\gamma<3$ [1], this result is practically important.

We summarize the obtained γ_s's for various N_s's and γ's in Table I.

In contrast to the case $\gamma\lesssim3$, γ_s for $\gamma>3$ slightly deviates from the γ of the original networks if $N_s/N_o\lesssim0.1$. [See the data for $\gamma=4.2$ in Figs. 1(e) and 1(f) or in Table I.] Numerically we find that γ_s is nearly equal to the original γ for $N_s/N_o>0.1$ when $\gamma\lesssim4.2$. Of course, one can expect the substantial deviation of γ_s from γ as γ increases further from $\gamma=4.2$.

This γ-dependent behavior of $P(k)$ can be understood from Eqs. (1) and (2). Equation (1) indicates that $\langle k^2\rangle$ diverges with finite $\langle k\rangle$ for $\gamma\leq3$. This implies that the topology of a network has several dominant hubs which have an extraordinary large number of degrees when $\gamma<3$. Since the probability of visitation of the RW follows Eq. (2), the RW can more effectively find the central part of the network around the hubs when $\gamma<3$. Thus the sampled networks can inherit easily the topological properties of the original networks.

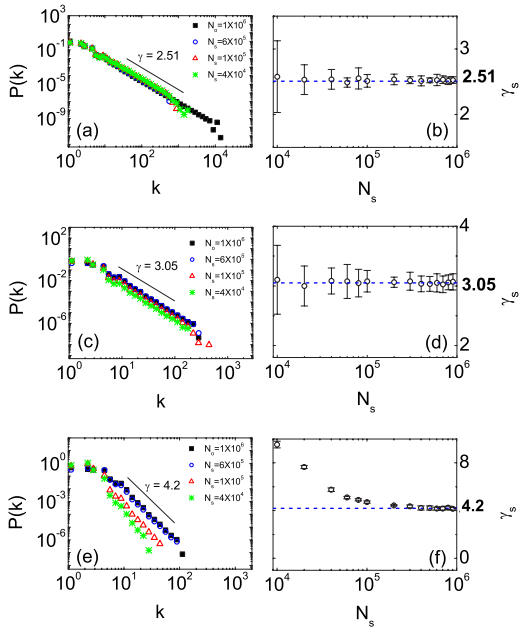

FIG. 1. (Color online) Degree distributions for sampled networks of static scale-free networks with (a) $\gamma=2.51$, (c) 3.05, and (e) 4.2. Degree exponents γ_s for the sampled networks extracted from the original network for the network size $N_o=10^6$ with (b) $\gamma=2.51$, (d) $\gamma=3.05$, and (f) $\gamma=4.2$. The slopes of solid lines in (a), (c), and (e) and the values of the dashed lines in (b), (d), and (f) are the degree exponents of the original networks.

The RWSM is also applied to real networks. In Fig. 2, we show the $P(k)$ of the WWW [3], the actor network [17], and the P2P networks (Gnutella) [18]. The number of nodes in the original real networks is $N_o=392340$, 325729, and 1074843 for the actor network, the WWW, and the Gnutella, respectively. The degree distribution for the WWW follows the power law with $\gamma=2.6$ (WWW) [3]. The data in Fig. 2(a) show that $P(k)$ of the sampled WWW also follows the power law with $\gamma\simeq2.6$ even for small N_s $(=10^3)$. For the actor network and the Gnutella, $P(k)$'s of the original networks are broad or fat-tailed and do not follow the simple power law (1). However, as one can see in Figs. 2(b) and 2(c), the actor network and Gnutella network also have big hubs which cause high heterogeneity in degree, and the sampled networks show nearly the same degree distribution as the original one. These results also provide evidence that the nodes with large degrees play an important role in the RWSM.

B. Degree-degree correlation

Another important measure to characterize the topological properties of the complex network is the degree-degree correlation. Many interesting topological properties such as the

TABLE I. The changes of the degree distribution exponents γ_s depending on sampled network size N_s. γ's are the degree exponents of the original network with $N_o=10^6$.

	$10^{-6}\ N$									
γ	0.8	0.6	0.4	0.2	0.1	0.08	0.06	0.04	0.02	0.01
2.23(5)	2.23(3)	2.24(3)	2.24(2)	2.24(3)	2.3(1)	2.2(2)	2.3(2)	2.3(3)	2.3(5)	2.3(5)
2.51(7)	2.51(6)	2.53(8)	2.51(8)	2.54(8)	2.5(1)	2.6(2)	2.49(7)	2.5(1)	2.5(1)	2.6(5)
3.05(9)	3.1(1)	3.1(2)	3.0(1)	3.06(9)	3.1(2)	3.1(2)	3.1(3)	3.1(2)	3.0(3)	3.1(6)
3.40(8)	3.37(7)	3.40(9)	3.4(1)	3.4(1)	3.4(2)	3.5(3)	3.4(2)	3.7(4)	3.8(4)	4.4(3)
4.2(1)	4.2(1)	4.2(1)	4.2(2)	4.44(5)	4.71(9)	4.91(8)	5.1(1)	5.8(2)	7.7(1)	9.5(3)

self-similarity [20] can be affected by the degree-degree correlation. The degree-degree correlation can be characterized by $\langle k_{nn}(k)\rangle$, the average degree of the nearest neighbors of nodes with degree k [21,22]. If the $\langle k_{nn}(k)\rangle$ increases (decreases), the network is characterized as assortative (disassortative) mixing. As shown in Fig. 3(a), for the static SFN with $2<\gamma<3$ the original network and the sampled networks all show the disassortative mixing. This can be explained by the dynamical properties of RWs on complex networks. In the networks showing disassortative mixing, the RW on a hub should go through a node of small k to move to another hub. Thus, many nodes having small k can be connected to the hubs in the sampled networks and the sampled networks remain disassortative. If the networks have neutral degree correlation, then the networks sampled by the RW also show neutral degree correlation. [see Figs. 3(b) and 3(c).] In Figs. 3(d)–3(f), we plot $\langle k_{nn}(k)\rangle$ of real networks. $\langle k_{nn}(k)\rangle$'s of the sampled networks show the same degree correlations as those of the original networks. As shown in Figs. 3(d)–3(f), the degree correlations are assortative, disassortative, and neutral for the actor, WWW, and Gnutella networks, respectively.

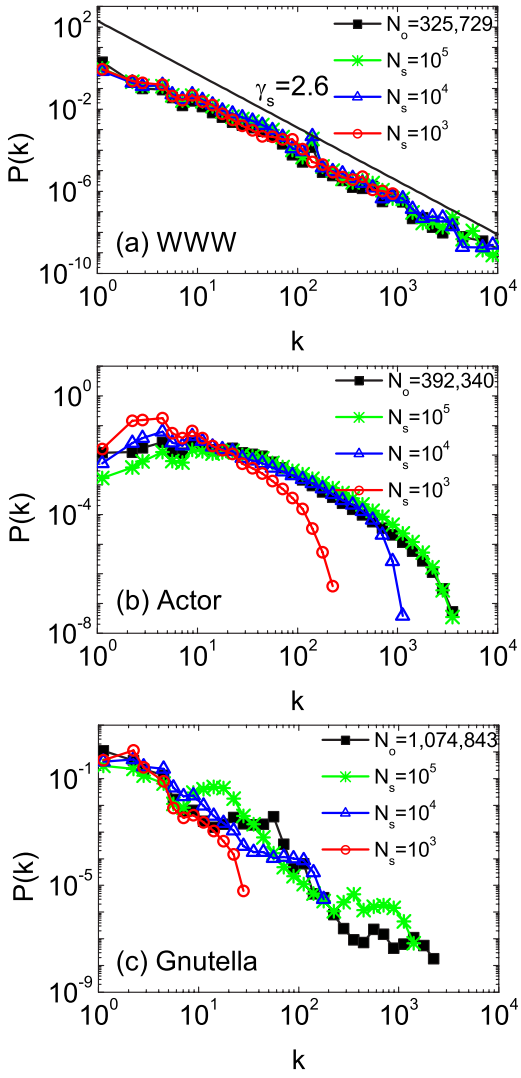

FIG. 2. (Color online) Degree distributions for sampled networks of three real networks. (a) WWW (N_o=325,729, γ=2.6) [3], (b) collaboration network of movie actors (N_o=392,340) [17], and (c) Gnutella (N_o=1,074,843) [18]. The slopes of the solid line in (a) is the value of degree exponents obtained from the simple linear fitting for degree distributions of the sampled networks.

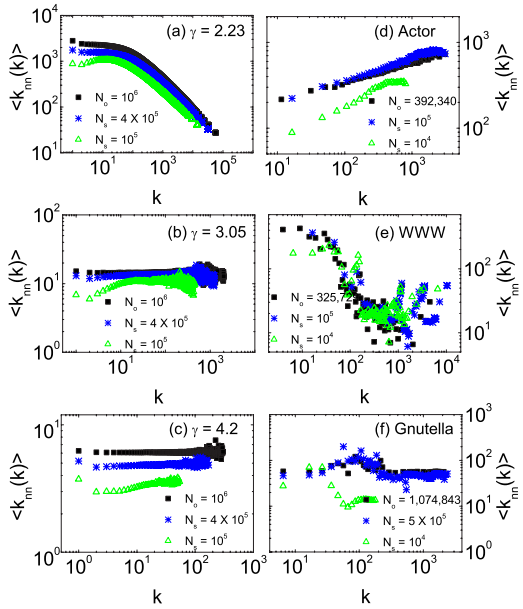

FIG. 3. (Color online) Distributions of $\langle k_{nn}\rangle$ for subnetworks extracted from the original networks with (a) γ=2.23, (b) γ=3.05, and (c) γ=4.2. (d) Collaboration network of movie actors. (e) WWW. (f) Gnutella.

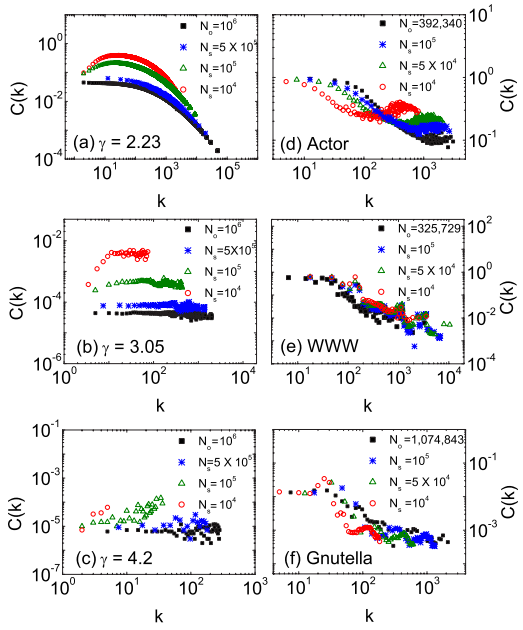

FIG. 4. (Color online) $C(k)$ for subnetworks from the original networks with (a) γ=2.23, (b) γ=3.05, and (c) γ=4.2. (d) Collaboration network of movie actors. (e) WWW. (f) Gnutella.

C. Clustering coefficient

We also measure a clustering coefficient of the sampled networks. The clustering coefficient C_i of a node i is defined by

$$C_i = \frac{2y_i}{k_i(k_i-1)}, \tag{3}$$

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 a infinite-sized random network [1].

In Fig. 4, we plot the clustering coefficient $C(k)$ against degree k. The shifts in the value of $C(k)$ as changing N_s can be understood from the local topology of the networks. For example, in the static SFN $C(k)$ $[\sim(\ln N)^2/N]$ for γ=3 is larger than $C(k)$ $(\sim 1/N)$ for $\gamma > 3$ [23]. This implies that the SFN with γ=3 has more triangular loops (loops of length 3) around the nodes of large k than the networks with $\gamma > 3$. Thus in SFN with γ=3 the triangular loops can be sampled more than the tree like regions by RW, which causes the large shift in the value of $C(k)$ as changing N_s for γ=3. In real networks, the degree correlation also seems to play an important role. For example, the actor network shows an assortative mixing [see Fig. 3(d).] Considering the assortativity and $C(k)$ for large k in Fig. 4(d), we expect that the actor network has more interconnections between hubs than WWW or Gnutella. Thus $C(k)$ of the sampled actor networks can deviate more from that of the original network compared to WWW or Gnutella [Figs. 4(d) and 4(e)].

More importantly, $C(k)$ is also known to reflect the hierarchical modular structure of networks [21,24]. $C(k)$ does not depend on k if the network does not have any well-defined hierarchical modules [21,24]. As shown in Fig. 4, $C(k)$ of both the original networks and the sampled networks shows a tendency to decrease with increasing k for SFN with γ=2.23 and real networks. [See Figs. 4(a) and 4(d)–4(f)]. This implies that the sampled networks have the same modular structure as the original networks. On the other hand, the topology of networks with $\gamma \gg 3$ resembles closely the random graph; thus, $C(k)$ does not depend on the degree k [24]. The dependence of $C(k)$ on k for the sampled SFNs with $\gamma \geq 3$ is also nearly the same as the original SFNs. [See Figs. 4(b) and 4(c).]

IV. DISCUSSION AND CONCLUSIONS

We study the topological properties of sampled networks by the RWSM with SFNs and several real networks. From the numerical simulations, we find that the $P(k)$ of the sampled network follows the power law $P(k) \sim k^{-\gamma_s}$. We also find that the $\gamma_s \simeq \gamma$ for all N_s when $2 < \gamma \lesssim 3$. Even though γ_s somewhat increases as decreasing N_s for $\gamma > 3$, the γ_s's with $N_s/N_o \gtrsim 0.1$ still follow the original one. We also study the degree-degree correlation and clustering coefficient by measuring $\langle k_{nn}(k) \rangle$ and $C(k)$. The sampled networks have the same degree correlation and modular structure as the original networks for all values of γ. The RWSM is also applied to the actor, WWW, and Gnutella networks. By measuring $P(k)$, $\langle k_{nn}(k) \rangle$, and $C(k)$, we confirm that the topological properties of the sampled networks are well maintained after sampling and the RWSM is an efficient sampling method for the real networks. A similar degree-based weighted sampling has already been applied to web crawlers [16,25], sampling the P2P networks for studying the size-dependent behavior [19], and a model-based testing strategy of communication systems such as internet router protocol [26].

The numerical simulations indicate that γ plays very important role in the RWSM. γ dependent behavior of the sampled networks can be understood from the dynamical property of RWs. Since most of the networks in the real world have $2 < \gamma < 3$, the results imply very important meaning in practice. Based on our results, we expect that if we obtain the empirical data by weighted sampling in which the weight is proportional to the degree, then the sampled networks can share the same topological properties with the whole network. Especially, the weighted sampling method becomes very efficient as the heterogeneity of networks increases. At the same time, we also expect that our study can provide a systematic way to extract subnetworks from the empirical data to study the size-dependent behavior of various dynamical properties on many real networks satisfying $\gamma < 3$ [19].

On the other hand, if γ is sufficiently large (or in the limit $\gamma \to \infty$), the original network becomes homogeneous. For large γ, we thus expect that the RWSM becomes equivalent to random sampling, since γ_s increases as N_s decreases as shown in Fig. 1 and Table I. In practice, the RWSM has a limited applicability for $\gamma > 3$ when the size of subnetworks becomes smaller.

As mentioned in Sec. III, the RWSM is very efficient to sample a node of larger degree rather than a node of smaller degree, which means that all the nodes of small degree around the hubs can not be sampled. This is the main difference between the RWSM and the snowball algorithm. As a result, some important nodes of small degree (e.g., the connecting nodes between hubs in disassortative networks) have a chance not to be sampled even though the subnetworks preserve the topological properties of the original network well. Restrepo *et al.* recently reported that some nodes of smaller degree can be more important than a node of larger degree in several real networks [27]. In future studies it is desirable to combine the RWSM with the method in which important nodes of small degree are effectively sampled.

ACKNOWLEDGMENTS

This work is supported by Grants No. R01-2006-000-10470-0 from the Basic Research Program of the Korea Science & Engineering Foundation and No. KRF-2004-015-C00185 from the Korea Research Foundation.

[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 (London) **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 (London) **403**, 623 (2000).
[7] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004); J. Lahtinen, J. Kertész, and K. Kaski, Phys. Rev. E **64**, 057105 (2001); B. Tadić, Eur. Phys. J. B **23**, 221 (2001).
[8] S. Lee, S. H. Yook, and Y. Kim, Phys. Rev. E **74**, 046118 (2006).
[9] M. Stumpt and C. Wiuf, Phys. Rev. E **72**, 036118 (2005).
[10] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong, Phys. Rev. E **73**, 016102 (2006).
[11] M. E. J. Newman, Soc. Networks **25**, 83 (2003).
[12] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[13] R. Sedgwick, *Algorithms* (Addison-Wesley, Reading, MA, 1988).
[14] http://mips.gsf.de; http://www.expasy.ch
[15] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[16] http://en.wikipedia.org/wiki/Web_crawler
[17] A. -L. Barabási and R. Albert, Science **286**, 509 (1999).
[18] D. Stutzbach and R. Rejaie (unpublished).
[19] S. Lee, S. H. Yook, and Y. Kim, eprint arXiv:physics/0703040.
[20] S. H. Yook, F. Radicchi, and H. Meyer-Ortmanns, Phys. Rev. E **72**, 045105(R) (2005).
[21] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[22] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[23] J.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim, Eur. Phys. J. B **49**, 231 (2006).
[24] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002); E. Ravasz and A.-L. Barabási, *ibid.* **67**, 026112 (2003); A. Vázquez, *ibid.* **67**, 056104 (2003); J.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim, Eur. Phys. J. B **49**, 231 (2006).
[25] Ziv Bar-Yossef and Maxim Gurevich (unpublished); R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez (unpublished).
[26] A. Denise, M.-C. Gaudel, S.-D. Gouraud *et al.* (The RsST group) (unpublished).
[27] J. G. Restrepo, E. Ott, and B. R. Hunt, Phys. Rev. Lett. **97**, 094102 (2006).