PHYSICAL REVIEW E **79**, 065101(R) (2009)

Shortest path discovery of complex networks

Attila Fekete,* Gábor Vattay,† and Márton Pósfai‡
Department of Physics of Complex Systems, Eötvös University, Pázmány P. sétány 1/A., H-1117 Budapest, Hungary
(Received 8 October 2008; revised manuscript received 14 May 2009; published 23 June 2009)

In this Rapid Communication we present an analytic study of sampled networks in the case of some important shortest-path sampling models. We present analytic formulas for the probability of edge discovery in the case of an evolving and a static network model. We also show that the number of discovered edges in a finite network scales much more slowly than predicted by earlier mean-field models. Finally, we calculate the degree distribution of sampled networks and we demonstrate that they are analogous to a destroyed network obtained by randomly removing edges from the original network.

DOI: 10.1103/PhysRevE.79.065101 PACS number(s): 89.75.Hc, 64.60.aq, 89.20.Hh

Complex networks have attracted significant interest in recent years [1,2]. In most cases, the entire structure of the network is unknown and one is left with statistical samples of the original network [3,4]. The sampling of Internet topology is one of the greatest challenges due to its enormous size and decentralized structure. It motivated numerous studies on the relationship between the original and the sampled network, including the degree distribution [5–7] and the expected size of the network [8]. Recently, Internet sampling methods have emerged that rely on the measurement tool *traceroute*, which returns the sequence of IP addresses of the network nodes along the path between the measurement host and a given destination host. An abstraction of the network discovery process consists of selecting a set of source and target nodes and finding the shortest paths between source and destination pairs. A node or an edge of the network is *discovered* if it belongs to one of those shortest paths. The statistical properties of the discovered network have been studied extensively by Dall' Asta *et al.* [9]. The mean-field approximation has been developed in the limit of low source and target density $\rho_S\rho_T \ll 1$ by neglecting the correlation of different shortest paths.

In this Rapid Communication we present exact results for certain networks. A surprising finding is that the network discovery process is slower in these systems than it is predicted by the mean-field theory. While in mean-field approximation the number of discovered links scales with the product of the number of the source and target nodes, our approach predicts a scaling only with their sum. The lower number of discovered edges is a result of the high degree of overlapping between shortest paths. Our other important finding concerns the degree distribution of the discovered network. We will show that it is analogous with a destroyed network where a fraction of the edges of the original network has been randomly removed.

We investigate two main discovery strategies. In peer-to-peer (P2P) sampling each node is selected simultaneously for both source and target with probability ρ. Computer applications using the peer-to-peer principle discover the network this way, hence the name. In disjunct (DI) sampling each node is selected for source or target but not for both with probabilities ρ_S and ρ_T. This strategy is used in Internet mapping projects, where source computers belong to the measurement infrastructure, while a large number of random addresses are selected as targets.

We start our analysis with the discovery of a tree. The most important observation permitting exact calculations in this case is that an edge separates the tree into two sides. An edge is discovered only if the source and the target nodes reside on different sides of the edge. Let us denote the event that a node is selected as a source or target by S and T, respectively. Furthermore, we denote the event that at least one source or target node resides on the "left" or "right" side of the edge by $S_{L,R}$ and $T_{L,R}$, respectively. The event that a link is discovered, D, provided that its two sides L and R are known, is clearly $D=(S_LT_R)+(S_RT_L)$. Therefore, we can express the conditional probability

$$P(D|L,R) = P(S_L|L,R)P(T_R|L,R) + P(S_R|L,R)P(T_L|L,R) - P(S_LT_L|L,R)P(S_RT_R|L,R).$$

The probabilities arising in this expression can be calculated easily: $P(S_\lambda|L,R)=1-P^{N_\lambda}(\bar{S})$, $P(T_\lambda|L,R)=1-P^{N_\lambda}(\bar{T})$ and $P(S_\lambda T_\lambda|L,R)=1-P^{N_\lambda}(\bar{S})-P^{N_\lambda}(\bar{T})+P^{N_\lambda}(\bar{S}\bar{T})$, where $\lambda=L$ or R, N_L and N_R are the number of nodes on the two sides of the link, and the overlines denote complement events.

Let us consider an evolving network where each new node is attached randomly to one of the nodes of the existing network. The structure of this network will be a tree. Since the network is connected the cluster sizes N_L and N_R must satisfy the relation $N_L+N_R=N$, where N is the size of the whole network. In the thermodynamic limit $N\to\infty$ we obtain $P(D|N_L)=1-\sigma^{N_L}$, where we have introduced $\sigma=P(\bar{S}\bar{T})$. The probability σ in the different sampling models is related to the source and target densities in a simple way:

$$\sigma = \begin{cases} 1-\rho & \text{P2P} \\ 1-\rho_S-\rho_T & \text{DI,} \end{cases} \quad (1)$$

where $\rho,\rho_S,\rho_T\in[0,1]$, $\rho_S+\rho_T\leq 1$. If $\rho_S+\rho_T\ll 1$ in the DI sampling model, then we can write $P(D|N_L) \approx 1-\exp(-\frac{\rho_S+\rho_T}{N}b_e)$, where $b_e=N_L(N-N_L)$ is the number of shortest paths that traverse a given link called *betweenness*

*fekete@complex.elte.hu
†vattay@elte.hu
‡posfai@complex.elte.hu

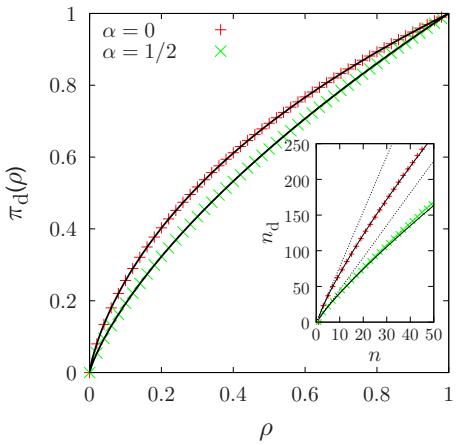

FIG. 1. (Color online) Discovery probability of edges $\pi_{\mathrm{d}}(\rho)$ as the function of the measurement node density ρ for P2P sampling of evolving trees. Data points are averaged over 100 realizations of $N=10^4$ node BA trees with $\alpha=0$ and 1/2. Solid lines show the corresponding analytic solution (3) with $\sigma=1-\rho$. Inset: the number of discovered edges n_{d} as the function of the number of the measurement nodes n. The solid line represents Eq. (4), whereas the dotted line shows its leading term $\langle l\rangle n/2$ with $\langle l\rangle=15.48$ and 9.045 for $\alpha=0$ and 1/2, respectively.

centrality. Compare this result with the mean-field model of Dall' Asta *et al.* [9]: $P(D_{\mathrm{mf}}|b_e)\approx 1-\exp(-\rho_S\rho_T b_e)$.

The probability of finding an arbitrary edge by traceroute probes can be given now straightforwardly:

$$\pi_{\mathrm{d}}=\sum_{N_L=0}^{\infty}P(D|N_L)P(N_L)=1-H_1(\sigma), \tag{2}$$

where $H_1(z)=\Sigma_{N_L}P(N_L)z^{N_L}$ is the generating function of the cluster size distribution $P(N_L)$.

Expression (2) has been tested on the Dorogovtsev-Mendez (DM) network growth model [10], a generalization of the Barabási-Albert (BA) model [11], where new nodes with m new links are attached to old nodes with degree-dependent probability $\Pi(k_i)=\frac{k_i-m+am}{\Sigma_i(k_i-m+am)}$, where $a\geq 0$. The growing tree corresponds to $m=1$. We calculated the distribution $P(N_L)$ for this model analytically in Ref. [12]. The generating function can be expressed in terms of hypergeometric functions $H_1(z)=z_2F_1(1-\alpha,1,2-\alpha;z)-z\frac{1-\alpha}{2-\alpha}{}_2F_1(2-\alpha,1,3-\alpha;z)$ and $\alpha=\frac{1}{1+a}$. At $a=1$ we recover the original BA preferential attachment model with scale-free degree distribution and at $a=+\infty$ we obtain uniform attachment probability with exponential degree distribution. In these cases π_{d} can be expressed with elementary functions

$$\pi_{\mathrm{d}}=\begin{cases}-\dfrac{1-\sigma}{\sigma}\ln(1-\sigma) & \text{if } a=+\infty(\text{i.e., } \alpha=0),\\ \dfrac{1-\sigma}{2\sqrt{\sigma}}\ln\dfrac{1+\sqrt{\sigma}}{1-\sqrt{\sigma}} & \text{if } a=1(\text{i.e., } \alpha=1/2).\end{cases} \tag{3}$$

Figure 1 shows simulations for the P2P sampling model at $\alpha=0$ and 1/2. The analytic results (3), plotted with dashed lines, fit the simulation data excellently.

From the point of view of the efficiency of the discovery process, it is important to calculate how many edges can be discovered with a given number of source n_S and target nodes n_T. For the Internet discovery the disjunct sampling model is relevant, where $\varrho_T+\varrho_S=(n_T+n_S)/N=n/N=1-\sigma\ll 1$. The series expansion of Eq. (2) yields $\pi_{\mathrm{d}}=1-\Sigma_{N_L}P(N_L)(1-\frac{n}{N})^{N_L}$. We can rearrange the series by adding and subtracting the terms $1-n\frac{N_L}{N}$ and averaging them separately $\pi_{\mathrm{d}}=\frac{n\langle N_L\rangle}{N}-\Sigma_{N_L}P(N_L)[(1-\frac{n}{N})^{N_L}-1+n\frac{N_L}{N}]$.

Several authors have pointed out that the distribution of $b_e=N_L(N-N_L)$ follows a universal power-law tail in trees with exponent -2 [12–14]. It also implies that asymptotically $P(N_L)\approx cN_L^{-2}$ in an arbitrary tree for $N_L\gg 1$. Specifically, $c=1-\alpha$ in the DM model. Using this asymptotic form we can calculate the leading behavior in the $N\to\infty$ limit $\pi_{\mathrm{d}}=\frac{n\langle N_L\rangle}{N}-c\mathrm{Li}_2(1-n/N)+c\frac{\pi^2}{6}-c\frac{n}{N}(\ln N-\gamma)$, where $\mathrm{Li}_2(x)$ is the dilogarithm function and $\gamma\approx 0.5772$ is the Euler constant. For small argument $\mathrm{Li}_2(1-x)$ can be expanded by using Euler's reflection formula $\mathrm{Li}_2(1-x)=-\mathrm{Li}_2(x)+\frac{\pi^2}{6}-\ln(x)\ln(1-x)\approx -x+\frac{\pi^2}{6}+x\ln(x)+\cdots$. Finally we get $\pi_{\mathrm{d}}=\frac{n\langle N_L\rangle}{N}+c\frac{n}{N}-c\frac{n}{N}\ln n-c\frac{n}{N}\gamma$.

To process this further, let us express the term $\langle N_L\rangle$ more straightforwardly. The sum of b_e for all edges clearly equals the total length of the shortest paths between all possible pairings of nodes: $\Sigma_{e\in E}b_e=\Sigma_{i,j\in V}l_{i,j}$. Since $\langle b\rangle=\frac{1}{N-1}\Sigma_{e\in E}b_e$ and $\langle l\rangle=\frac{2}{N(N-1)}\Sigma_{i,j\in V}l_{i,j}$ we can write $\langle l\rangle N/2=\langle b\rangle$. Therefore, the average branch size can be given as $\langle N_L\rangle=\langle l\rangle/2+\langle N_L^2\rangle/N$, where $\langle N_L^2\rangle/N=\frac{1}{N}\Sigma_{N_L=1}^{N}\frac{c}{N_L^2}N_L^2=c$. For a large, but finite network the average number of discovered edges is $\langle n_{\mathrm{d}}\rangle=(N-1)\pi_{\mathrm{d}}$, that is

$$\langle n_{\mathrm{d}}\rangle\approx n\left(\frac{\langle l\rangle}{2}-c\ln n+2c-c\gamma\right) \tag{4}$$

in the limit $1\ll n=n_S+n_T\ll N$, The above result shows that $\langle n_{\mathrm{d}}\rangle$ depends on the sum of n_S and n_T. This is in contrast to the mean-field model, which predicts that $\langle n_{\mathrm{d}}\rangle$ scales with the product of n_S and n_T. The logarithmic term of Eq. (4) accounts for the possibility that a new measurement node is placed at a node discovered by previous measurement nodes. The inset of Fig. 1 displays simulation results and the formula corresponding to the P2P sampling.

We continue with the analysis of a static model where nodes are randomly connected with a prescribed degree distribution p_k. This "configuration model" is a generalization of the Erdős-Rényi (ER) model [15], where the degree distribution is Poissonian. It has been shown in [16] that the generating function of branch sizes $H_1(z)$ satisfies the implicit equation $H_1(z)=zG_0'[H_1(z)]/\langle k\rangle$, where $G_0(z)=\Sigma_k p_k z^k$ is the generating function of the degree distribution. In the configuration model loops become irrelevant in the thermodynamic limit $N\to+\infty$ and each edge is a part of a tree. Here, N_L and N_R are independent and the joint probability function has a product form $P(N_L,N_R)=P(N_L)P(N_R)$. The summation in π_{d} can be carried out separately for N_L and N_R, which yields

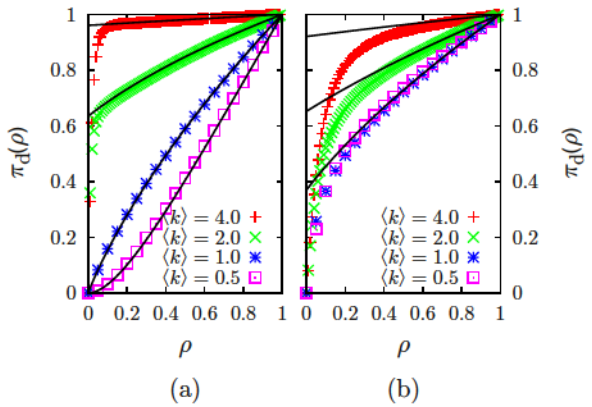

FIG. 2. (Color online) Discovery probability of edges as the function of the measurement node density ρ for static (a) Poissonian and (b) power-law networks. 100 P2P samplings were averaged in $N=10^4$ size networks with average degrees $\langle k\rangle=0.5$, 1, 2, and 4. Solid lines show analytic formula (6). Note that the analytic results for $\langle k\rangle=0.5$ and 1 are the same in the case of the power-law model. This result is confirmed by the numerical simulations, where data points collapse almost completely.

$$\pi_d = 2\{1 - H_1[P(\bar{S})]\}\{1 - H_1[P(\bar{T})]\} - \{1 - H_1[P(\bar{S})] - H_1[P(\bar{T})] + H_1[P(\bar{S}\bar{T})]\}^2. \quad (5)$$

In the case of P2P discovery this can be reduced to

$$\pi_d = [1 - H_1(1-\rho)]^2. \quad (6)$$

This formula can be tested on the ER model, with $G_0(z) = e^{\langle k\rangle(z-1)}$. The cluster size distribution can be given by the Lambert W function $H_1(z) = -W(-\langle k\rangle e^{-\langle k\rangle z})/\langle k\rangle$. Simulation results are presented in Fig. 2(a). The analytic result (6) is also shown for comparison. One can see that it is discontinuous at zero density if $\langle k\rangle > 1$, when a giant component emerges in the network. The simulation data deviate from the analytic solution around the discontinuity due to finite-scale effects. The size of the jump is $P_0 = [1 - H_1(1)]^2$, which is precisely the probability of infinitely large branches being attached to both sides of an edge. If P_0 is regarded as an order parameter, the observed phenomenon resembles a phase transition at $\langle k\rangle = k_c = 1$.

We also generated networks with power-law degree distribution using the hidden-variable model introduced in [17–20]. Simulations are shown in Fig. 2(b) with degree exponent $\gamma=3$. Note that the analytic solution is discontinuous at zero density, i.e., $P_0 > 0$, for all $\langle k\rangle > 0$. The phase transition can be observed again, since the analytic solution—and P_0—is independent of $\langle k\rangle$ below a critical point $k_c(\gamma) = \frac{\zeta(\gamma-1)}{\zeta(\gamma)}$. Indeed, data points almost collapse at $\langle k\rangle = 0.5$ and 1 which are below $k_c(\gamma=3) \approx 1.3684$. The phenomenon occurs when the degree generating function $G_0'(z)$ depends linearly on $\langle k\rangle$. This is characteristic of pure power-law distributions until $\langle k\rangle$ is below the critical value k_c.

Now we turn our attention to the degree distribution $P_d(k')$ of the discovered nodes. In our analysis we consider only the contribution of those shortest paths to k', which *traverse* a given node. We will show that $P_d(k')$ is analogous to the degree distribution of a partially severed network obtained by random edge pruning [21,22]. This duality between the sampling and the destruction of networks is very surprising considering the striking differences between the two processes, e.g., the explored network is surely connected in contrast to the destroyed one.

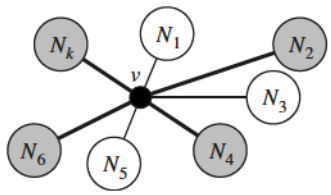

FIG. 3. Sketch of an arbitrary vertex v with degree k and the emerging branches with sizes $N_1, N_2, \ldots, N_k$. Shaded circles represent branches where measurement nodes can be found in. Thick lines symbolize the discovered edges of node v.

Let us consider a node v with original degree k. If every link is removed independently with probability p, then k', the degree of the node after random edge removal, will follow a binomial distribution: $P(k'|k) = \binom{k}{k'}(1-p)^{k'}p^{k-k'}$. Consequently,

$$P_{\text{pruned}}(k') = \sum_{k=k'}^{\infty} \binom{k}{k'}(1-p)^{k'}p^{k-k'}P_0(k). \quad (7)$$

Regarding the sampling process we examine a randomly selected node of the discovered network $v \in V_d$ in the static model first. Let us suppose that the sizes of the branches with original degree k are $N_1, N_2, \ldots, N_k$ (see Fig. 3). For the sake of simplicity we discuss only the P2P sampling model, where the probability of placing a measurement node in branch i is simply $(1-\sigma^{N_i})$. Since branch sizes are independent we can average over N_i separately. The results we obtain indicate that measurement nodes can be found in different branches with probability $1-H_1(\sigma)$.

We can see from Fig. 3 that the degree of a discovered node k' equals the number of branches where measurement nodes can be found in. It follows that $P_d(k'|k) = \frac{1}{P(v\in V_d|k)}\binom{k}{k'}[1-H_1(\sigma)]^{k'}H_1^{k-k'}(\sigma)$, where $2 \le k' \le k$. The subscript of P_d refers to the probability distribution restricted to the discovered network. In order to obtain the distribution of k' one should average this probability over $P_d(k)$, the distribution of *the original degrees of the discovered nodes*. This distribution can be obtained by $P_d(k) = \frac{P(v\in V_d|k)P_0(k)}{P(v\in V_d)}$, so

$$P_d(k') = \frac{\sum_{k=k'}^{\infty}\binom{k}{k'}[1-H_1(\sigma)]^{k'}H_1^{k-k'}(\sigma)P_0(k)}{P(v \in V_d)}, \quad (8)$$

where $k' \ge 2$ and $P(v \in V_d) = 1 - G_0[H_1(\sigma)] - [1 - H_1(\sigma)]G_0'[H_1(\sigma)]$. It is evident from Eqs. (7) and (8) that $P_d(k')$ equals $P_{\text{pruned}}(k')$—normalized properly for $k' \ge 2$—if $p = H_1(\sigma)$. In other words the discovered network is equivalent with an edge destroyed one.

In the case of an evolving network at least one of the branches, say N_k, tends to infinity as $N \to \infty$, so the probabil-

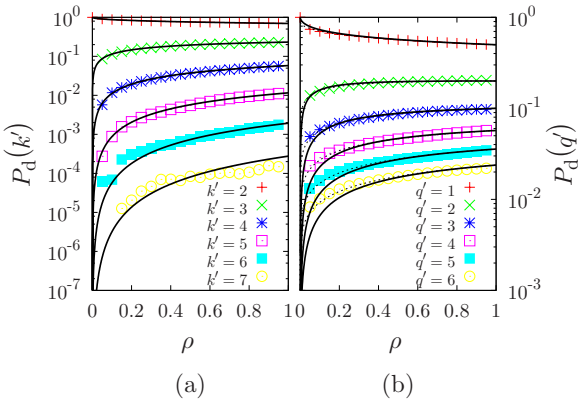

FIG. 4. (Color online) The probability of discovered degree $P_d(q')$ and in-degree $P_d(q')$ as the function of ρ in P2P sampling model. The original networks are $N=10^4$ node (a) static ER and (b) evolving BA graphs. Data points are averaged for ten networks with ten samplings in each realization. Solid lines consist of analytic solution (8). Exact solution for the evolving model is shown with dotted lines for comparison.

ity that a measurement node can be found in the kth branch tends to 1. In order to circumvent this effect let us redefine the network in such a way that every link should be directed toward the gigantic side of the network. Let $q=k-1$ denote the in-degree of nodes in this directed network. It is easy to see that the discovered *in-degree* q' will be equal to the number of branches where measurement nodes can be found in. We can follow the same procedure as in the case of the static model. We only need to replace k' and k in Eq. (8) with the corresponding in-degrees q' and q, and the normalization constant with $P(v \in V_d)=1-G_0^{(in)}[H_1(\sigma)]$.

Simulation results are shown for both static and evolving networks in Fig. 4. Note that we have assumed above that $H_1(\sigma)$ is independent of q. This is only an approximation in the case of the evolving network model. However, $H_1(\sigma|q)$ can be calculated exactly for the DM model, which is shown with dotted lines [23].

In conclusion we presented a study of network discovery processes. We derived analytically the probability of finding an arbitrary link of the network via shortest-path network discovery. We considered both static and evolving random networks with various sampling scenarios. We also demonstrated an important duality between the discovery of networks by shortest paths and the destruction of the same network by edge removal.

The authors thank the partial support of the National Science Fund Hungary (Grant No. OTKA 77779), the National Office for Research and Technology (Grant No. NAP 00635/2005), and the EU IST FET Onelab2 Project (Grant No. FP7-224263).

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] J. Leskovec and C. Faloutsos, in *Proceedings of the SIGKDD* (ACM, New York, NY, 2006), pp. 631–636.
[4] S. H. Lee, P.-J. Kim, and H. Jeong, Phys. Rev. E **73**, 016102 (2006).
[5] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie, *INFOCOM'03* (IEEE, New York, NY, 2003), Vol. 1, pp. 332–341.
[6] T. Petermann and P. De Los Rios, Eur. Phys. J. B **38**, 201 (2004).
[7] A. Clauset and C. Moore, Phys. Rev. Lett. **94**, 018701 (2005).
[8] F. Viger, A. Barrat, L. Dall'Asta, C.-H. Zhang, and E. D. Kolaczyk, Phys. Rev. E **75**, 056111 (2007).
[9] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vazquez, and A. Vespignani, Phys. Rev. E **71**, 036135 (2005).
[10] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[11] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[12] A. Fekete, G. Vattay, and L. Kocarev, Phys. Rev. E **73**, 046102 (2006).
[13] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **66**, 026101 (2002).
[14] D.-H. Kim, J. D. Noh, and H. Jeong, Phys. Rev. E **70**, 046126 (2004).
[15] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[16] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[17] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[18] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[19] B. Söderberg, Phys. Rev. E **66**, 066121 (2002).
[20] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[21] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[22] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).
[23] A. Fekete and G. Vattay (unpublished).