Scale-free trees: The skeletons of complex networks

Dong-Hee Kim,[1] Jae Dong Noh,[2] and Hawoong Jeong[1]
[1]*Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea*
[2]*Department of Physics, Chungnam National University, Daejeon 305-764, Korea*
(Received 12 December 2003; revised manuscript received 20 July 2004; published 28 October 2004)

We investigate the properties of the spanning trees of various real-world and model networks. The spanning tree representing the communication kernel of the original network is determined by maximizing the total weight of the edges, whose weights are given by the edge betweenness centralities. We find that a scale-free tree and shortcuts organize a complex network. Especially, in ubiquitous scale-free networks, it is found that the scale-free spanning tree shows very robust betweenness centrality distributions and the remaining shortcuts characterize the properties of the original network, such as the clustering coefficient and the classification of scale-free networks by the betweenness centrality distribution.

DOI: 10.1103/PhysRevE.70.046126 PACS number(s): 89.75.Hc, 89.75.Da, 89.75.Fb, 05.10.−a

Complex network theories have attracted much attention in the last few years with advances in the understanding of the highly interconnected nature of various social, biological, and communication systems [1,2]. The inhomogeneity of network structures is conveniently characterized by the degree distribution $P(k)$, the probability for a vertex to have k edges toward other vertices. The emergence of a *scale-free* distribution $P(k)\sim k^{-\gamma}$ has been reported in many real-world networks, such as the coauthorship networks in social systems [3], the metabolic networks and protein interaction networks in biological systems [4,5], and Internet and World Wide Web in technological systems [6,7].

It is important to study the dynamics as well as the structural properties of networks because of their applications to the real world. However, the dynamical phenomena of networks such as traffic and information flow are very difficult to predict from local information due to the rich microstructures and corresponding complex dynamics. Thus, to understand the dynamical phenomena of networks, one must know the global properties of networks as well as the local properties such as the degree distribution. It is the reason why the dynamics of complex networks has not been studied systematically so far.

Due to their inhomogeneous structure, the traffic or information flow of complex networks would be also very inhomogeneous. As a simplified quantity to measure the traffic of networks, it is natural to use the betweenness centrality (BC) [8–10]. The BC of $\mathbf{G}$, either a vertex or an edge, is defined as

$$b(\mathbf{G}) = \sum_{i\neq j} b(i,j;\mathbf{G}) = \sum_{i\neq j} \frac{c(i,j;\mathbf{G})}{c(i,j)}, \tag{1}$$

where $c(i,j;\mathbf{G})$ denotes the number of shortest paths from a vertex i to j through $\mathbf{G}$, and $c(i,j)$ is the total number of shortest paths from i to j. In terms of the packet in the Internet, assuming that every vertex sends a unit packet to each of the other vertices, BC is the average amount of packets passing though a vertex or an edge.

In scale-free networks, the distribution of the vertex BC is known to follow a power law with an exponent of either 2.2 or 2.0 [11]. Though the edge BC distribution does not follow the power law exactly, the distribution of the edge BC is also very inhomogeneous in scale-free networks [12]. This indicates that there exist extremely essential edges having large edge BC's which are used for communication very frequently. Thus, one can imagine a subnetwork constructed only by these essential edges with global connectivity retained. We regard this network as a communication kernel, which handles most of the traffic in a network.

For simplicity, we define the communication kernel of a network as the spanning tree with a set of edges maximizing the summation of their edge BC's on the original networks. The constructing procedure is very similar to the minimum spanning tree algorithm [13]. We repeatedly select an edge according to the priority of the edge BC and add the edge to the tree if it does not make any loop until the tree includes all vertices [14]. Note that the residual edges can be regarded as shortcuts since they shorten the paths on the spanning tree. This concept of the spanning tree and shortcuts corresponds to that of a one-dimensional (1D) regular lattice and shortcuts, respectively, in the small-world networks [15].

In this paper, we investigate the structural and dynamical properties of the spanning tree of complex networks and the role of shortcuts in the networks. Since most real-world networks show scale-free behavior, we mainly consider various scale-free real-world and model networks and we discuss the case of homogeneous networks at the end. In various real-world and model networks, we find that the spanning trees show scale-free behavior in the degree distributions. Especially, for all scale-free networks, we find that the vertex and edge BC distributions follow a power law with the robust exponent $\eta=2.0$, regardless of the exponent value $\eta=2.0$ or 2.2 of original networks. In addition to that, it turns out that the shortcut length distribution shows either Gaussian-like or monotonically decaying behavior depending on the BC distribution exponent η of original scale-free networks.

First, we confirm the spanning tree to be a communication kernel by estimating the relative importance of selected edges in the obtained spanning tree and those from the random selection. If we select the edges randomly, the fraction f of the edge BC summation over the selected edges and that over the total edge would be approximately f_0, the ratio of the number of edges in the tree and that of the network.

TABLE I. The scaling exponents and correlation coefficients of the spanning trees and original networks for various real-world networks and models. Tabulated for each network is the system size N, the mean degree $\langle k\rangle$, the ratio of edge BC summation over the edges selected for the spanning tree to total edge BC f, the ratio of the number of edges in spanning trees and original networks f_0, the degree exponent γ, the BC exponent η, the assortativity r, and the degree correlation coefficient r_p between the original network and the spanning tree. The s subscripts indicate quantities for the spanning trees. Here we consider only the largest cluster of networks when the network has several disjoint parts.

Network	N	$\langle k\rangle$	f	f_0	γ	γ_s	η	η_s	r	r_s	r_p	Ref.
NEURO	190382	12.5	0.46	0.16	2.1(1)	2.4(1)	2.2(1)	2.0(1)	0.601	−0.138	0.538	[17]
arxiv.org	44336	10.79	0.54	0.19	—	2.1(1)	—	2.0(1)	0.352	−0.119	0.497	[10]
arxiv.org/cond-mat	13860	6.43	0.61	0.31	—	2.7(1)	—	2.0(1)	0.157	−0.187	0.714	[10]
PIN	4926	6.55	0.54	0.3	—	2.3(1)	2.3(1)	2.0(1)	−0.139	−0.161	0.814	[18]
BA model	2×10^5	4	0.71	0.5	3.0(1)	2.7(1)	2.2(1)	2.0(1)	∼0	∼0	0.973	[16]
Holme-Kim model	10^4	6	0.58–0.71	0.33	3.0(1)	2.4(1)	2.2(1)	2.0(1)	−0.033	−0.117	0.947	[23]
Static model	$\sim10^4$	∼4	0.65	∼0.5	2.6–3.0	2.4–2.8	2.2(1)	2.0(1)	−0.022	−0.067	0.938	[24]
Fitness model	2×10^5	4	0.73	0.5	2.25	2.2(1)	2.2(1)	2.0(1)	∼0	∼0	0.994	[25]
Internet AS	10514	4.08	0.65	0.5	2.1(1)	2.1(1)	2.0(1)	2.0(1)	−0.185	−0.183	0.929	[19]
Adaptation model	$\sim10^5$	11.9	0.503	0.17	2.1(1)	2.1(1)	2.0(1)	2.0(1)	−0.219	−0.215	0.749	[26]
WS (p=1)	10^4	100	0.022	0.02	—	3.0(1)	—	1.7(1)	∼0	−0.176	0.549	[15]
ER	10^4	100	0.022	0.02	—	2.5(1)	—	1.7(1)	∼0	−0.209	0.453	[27]

However, it turns out that the real set of selected edges from the spanning tree possesses over ∼50% of the total edge BC in most networks (see Table I), and therefore $f \gg f_0$. For instance, the coauthorship network shows that f is nearly 3 times larger than f_0 even though the number of edges in the spanning tree is only 16% of that in the original network. Thus we can call this spanning tree of the scale-free network the communication kernel. However, for the case of homogeneous networks, the difference between f and f_0 is not significant because of the randomness of the network structure.

To find out more about this kernel, we measure the degree distribution of the spanning trees. It turns out that the degree distribution always follows the power law, which is tested for various networks including the Barabási-Albert (BA) model [16], coauthorship network in neuroscience (NEURO) [17], protein interaction networks of yeast (PIN) [18], Internet at the autonomous systems (AS) level [19], and so on (see Fig. 1 and Table I). However, the details of the degree distribution depend on each of the networks. The exponents of the power-law degree distributions of the spanning trees do not always agree with those of the original networks (see Table I). This indicates that the spanning trees are far from the random sampling of edges.

To confirm the scale-free behavior of the spanning tree, we investigate the time evolution of the degree in a growing network. Assuming that a fixed number of new vertices are introduced at each time step in the growing networks, it is well known that the degree following the power law $k_i(t) \sim t^{\beta}$ leads to the scale-free degree distribution $P(k) \simeq k^{-\gamma}$ [20], where $k_i(t)$ is the degree of the vertex i at time t and $\gamma = 1/\beta + 1$. This argument can be naturally applied for the spanning tree of the BA model since it grows constantly. At each time step of growth in the BA model, we obtain the spanning tree and measure the degree of every vertex. In Fig. 2(a), we show the time evolution of the degrees of several vertices. The degrees evolve with $\beta = 0.58$, which leads to $\gamma_s = 2.7$ of the spanning tree, which agrees with our measurement from the actual degree distribution.

The high correlation between the degrees from spanning trees and the original networks also guarantees the preserved scale-free behavior of the spanning trees. The correlation coefficient between the degree of the original network k and the degree of its spanning trees k_s is defined as the Pearson's correlation coefficient between k and k_s, $r_p = (\langle k k_s\rangle - \langle k\rangle \times \langle k_s\rangle) / \sqrt{(\langle k^2\rangle - \langle k\rangle^2)(\langle k_s^2\rangle - \langle k_s\rangle^2)}$. Most networks exhibit a strong degree correlation between the spanning tree and its original network (see Table I). We find that the degrees of

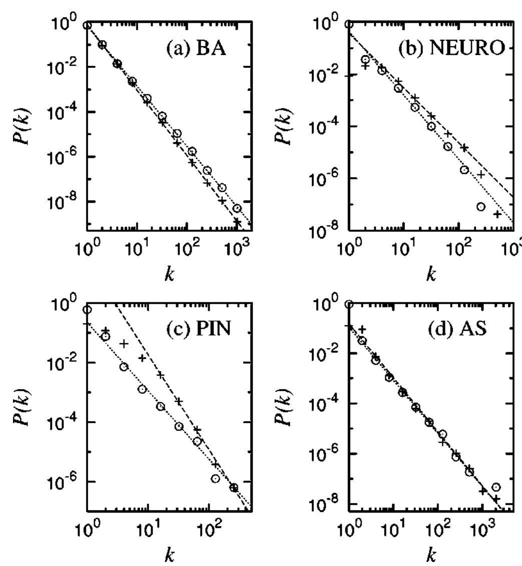

FIG. 1. Degree distributions of the spanning trees (○) and their original networks (+), (a) BA model with $m=2$, (b) coauthorship network, NEURO, (c) PIN, and (d) Internet AS. The data points are shifted vertically to enhance the visibility.

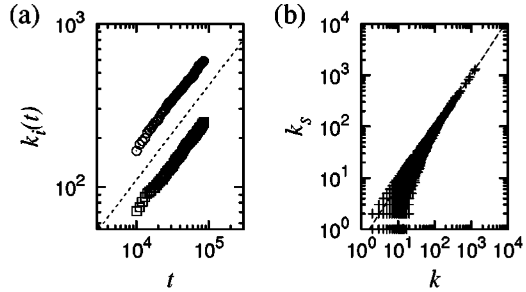

FIG. 2. (a) Time evolution for the degree of two vertices added to system at $t=5$ (○) and $t=55$ (□), where the dashed line is a linear fit with slope 0.58. (b) Scattered plot for degree of the original network (k) and the spanning tree (k_s). The dashed line has a slope of 1.08.

networks (k) and their spanning trees (k_s) roughly follow $k_s \sim k^{\alpha}$, which leads the degree distribution of the spanning trees, $P(k_s) \sim k_s^{-\gamma_s}$, with $\gamma_s = (\gamma + \alpha - 1)/\alpha$. In Fig. 2(b), we show that the BA model has $\alpha = 1.08$, which leads to $\gamma_s = 2.75$, in good agreement with the result obtained by direct measurement.

The assortativity is another interesting feature of the spanning trees. The assortativity r [21], which measures the degree correlation of vertices directly connected by an edge, is defined by $r = (\langle jk \rangle - \langle j \rangle \langle k \rangle)/(\langle k^2 \rangle - \langle k \rangle^2)$, where j and k are the remaining degrees at the end of an edge and the angular brackets indicate the average over all edges. We find that all spanning trees show dissortative or neutral behavior regardless of the assortativity of the original networks (see Table I). Thus, we can propose that it is a general characteristic of the spanning trees of scale-free networks. We need further study to prove our conjecture.

We find that the BC distribution of the spanning tree is robust regardless of its original network in scale-free real-world and model networks. The vertex and edge BC distributions of the spanning trees follow a power law with the robust exponent $\eta_s = 2.0$ (see Fig. 3 and Table I). This is consistent with the numerical and analytical results for the growing scale-free tree model [11]. The same BC distribution for vertices and edges is the general feature of trees. In the mean-field picture, the largest BC of edges belonging to a vertex gives a dominant contribution to the BC of the vertex [22]. For our obtained spanning trees, we verify numerically that the largest edge BC of a vertex almost equals the vertex BC for most of the vertices [see Fig. 3(c)].

The spanning trees show robust features, such as a scale-free degree distribution, robust BC distribution, and dissortative or neutral degree correlation. Here one can ask what is the role of shortcuts which are not included in the spanning tree. To answer this question, we focus on the length of the shortcuts on the spanning trees. The length of a shortcut between vertices i and j is defined as the minimum number of hops from i to j on the spanning tree. The nonzero clustering coefficient of the original networks can now be explained by short-length shortcuts. Obviously, shortcuts with length 2 build triangles of vertices and hence increase the clustering coefficient. All networks with nonvanishing clustering coefficient have a significant amount of shortcuts with length 2 (see Fig. 4).

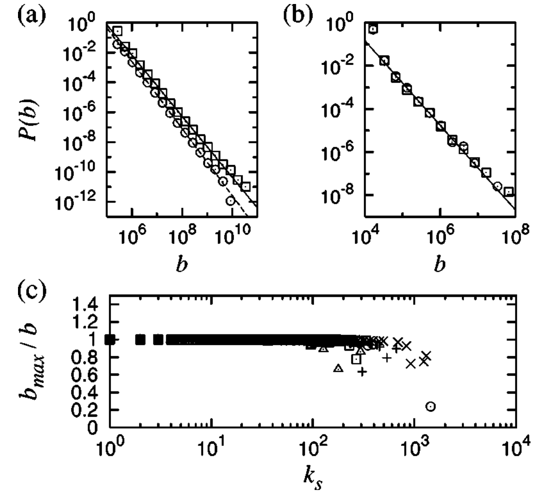

FIG. 3. The vertex BC distribution of the original networks (○) and the spanning tree (□) for (a) the BA model averaged over ten ensembles and (b) Internet AS. In (a), the solid and dashed lines have slopes of 2.0 and 2.2, respectively. The lines in (b) are linear fits with slope 2.0. The data points are shifted vertically to enhance the visibility. (c) The ratio of the largest value of edge BC ($b_{\max}$) to vertex BC (b) of a vertex with degree k_s for the BA tree (+) and the spanning trees of the BA model (×), NEURO (□), Internet AS (○), and PIN (△) networks.

Interestingly, in scale-free networks, we find that there are two types in the shortcut length distribution (see Fig. 4). In one distribution (type I), most shortcuts distribute near a large mean value, similar to the Gaussian distribution, which shows that the network is the longer-loop dominant structure. In the other distribution (type II), the number of shortcuts monotonically decreases as the length increases, which indicates that the network is tree like. Most of the networks including the BA model, coauthorship networks, and PIN belong to type I. On the other hand, Internet AS and the adaptation model are type II. We find that our classification exactly agrees with the grouping by exponent of the BC

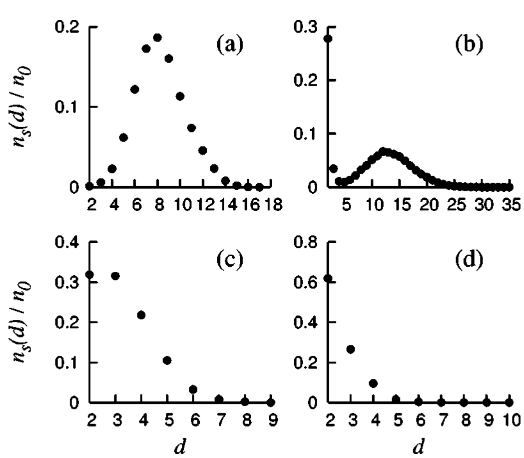

FIG. 4. The length distribution of shortcuts for (a) the BA model ($m=2$), (b) coauthorship network of neuroscience, (c) Internet AS, and (d) adaptation model with 10^5 vertices. $n_s(d)$ and n_0 are the number of shortcuts with length d and total number of shortcuts, respectively.

distribution [11]. The networks belonging to type I or type II show vertex BC distributions with an exponent of 2.2 or 2.0, respectively. Goh *et al.* [11] characterized the networks with a BC exponent of 2.0 as the linear mass-distance relation, which shows that the shortest paths of the networks are similar to trees. Our result also supports the treelike structures of the type-II networks and gives an intuitive explanation of the reason why the BC exponents of the type-II networks are as same as those of the scale-free trees. Because there exist mostly short-length shortcuts in the type-II networks with monotonically fast-decaying shortcut length distributions, the structure of the original networks is not significantly different from their spanning trees. Therefore, the BC exponents of the type-II networks are unchanged at 2.0 of their spanning trees.

Finally, we also study the spanning tree of homogeneous networks, the Erdös-Rényi (ER) random network [27] and the Watts-Strogatz (WS) small-world network [15] (see Fig. 5 and Table I). The sparse versions of networks, which have 10^4 vertices and average degree 4, show no scale-free behavior. They have no hub because the degree of the spanning tree has to be smaller than that of the original network. But interestingly, if the network gets much denser by increasing the average degree to 100, scale-free behavior appears in the degree distribution of the spanning tree up to $k < \langle k \rangle$. However, it does not share the properties with the trees from scale-free networks such as the BC exponent [28].

In summary, we have studied the properties of the spanning trees with maximum total edge betweenness centrality. We find that a complex network can be decomposed into scale-free trees and addition shortcuts on it. In various scale-free real-world and model networks, the scale-free spanning trees represent the communication kernels on networks. The scale-free spanning trees show robust characteristics in the degree correlation and the betweenness centrality distribution. The remaining shortcuts are responsible for detailed characteristics such as the clustering property and the BC distribution of scale-free networks. The distribution of the shortcut length clearly distinguishes the scale-free networks into two types, which coincides with the classes determined from the BC exponents [11]. We also find that the scale-free behavior can be found in homogeneous networks. We note that the property of the spanning tree generated from other methods was independently studied by other groups [29,30].

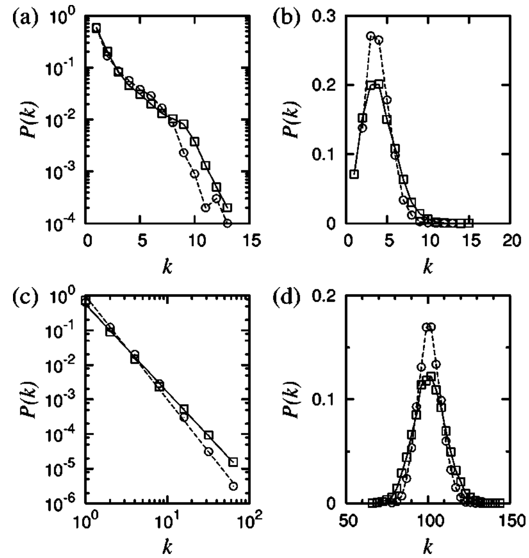

FIG. 5. Degree distribution of (a) the spanning trees of the sparse homogeneous networks with 10^4 vertices and average degree 4 and (b) their original networks. For dense homogeneous networks with 10^4 vertices and average degree 100, the degree distributions of (c) the spanning tree and (d) their original networks are drawn. The two popular homogeneous model networks, the ER network (□, solid line) and the WS network, with the rewiring probability $p = 1.0$ (○, dashed line), are tested.

The authors are grateful to Byungnam Kahng, Kwang-Il Goh, and Mark Newman for fruitful discussions and supporting the real-world network data. This work was supported by Grant No. R14-2002-059-01002-0 from the KOSEF-ABRL program and by Korean Systems Biology Research Grant No. M10309020000-03B5002-00000 from the Ministry of Science and Technology.

[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] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[4] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).
[5] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2000).
[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[7] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[8] M. L. Freeman, Sociometry **40**, 35 (1977).
[9] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[10] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001); **64**, 016132 (2001).
[11] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001); K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001); M. Barthélemy, *ibid.* **91**, 189803 (2003); K.-I. Goh, C.-M. Ghim, B. Kahng, and D. Kim, *ibid.* **91**, 189804 (2003).
[12] D.-H. Kim and H. Jeong (unpublished).
[13] J. B. Kruskal, Proc. Am. Math. Soc. **7**, 48 (1956).
[14] We study only undirected networks for simplicity of the algorithm. However, our analysis can be generally applied to directed networks.
[15] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[16] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[17] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[18] K.-I. Goh, B. Kahng, and D. Kim, e-print q-bio.MN/0312009.
[19] D. Meyer, University of Oregon Route Views Archive Project, http://archive.routeviews.org
[20] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[21] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[22] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **66**, 026101 (2002).
[23] P. Holme and B. J. Kim, Phys. Rev. E **65**, 026107 (2002).
[24] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[25] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2001).
[26] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **88**, 108701 (2002).
[27] P. Erdös and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[28] The BC exponent 1.7 of the spanning tree of the homogeneous network is different from 2.0 from the scale-free network.
[29] G. Szabó, M. Alava, and J. Kertész, Physica A **330**, 31 (2003).
[30] A. Clauset and C. Moore, e-print cond-mat/0312674.