Underlying Scale-Free Trees in Complex Networks

D.-H. Kim, S.-W. Son, Y.-Y. Ahn, P.-J. Kim, Y.-H. Eom and H. Jeong[*)]

Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

We investigate the properties of two relatively different spanning trees of complex networks, so-called "communication kernel" and "response network". First, for the communication kernel, we construct spanning trees carrying a maximum total weight of edges that is given by average traffic, which is defined as edge betweenness centrality. It is found that the resulting spanning tree plays an important role in communication between vertices. We also find that the degree distribution of spanning trees shows scale-free behavior for many model and real-world networks and the degree of the spanning trees has strong correlation with their original network topology. For the response network, we launch an attack on a single vertex which can drastically change the communication pattern between vertices of networks. By using minimum spanning tree technique, we construct the response network based on the measurement of the betweenness centrality changes due to a vertex removal. We find that the degree distribution of the response network indicates the scale-free behavior as well as that of the communication kernel. Interestingly, these two minimum spanning trees from different methods not only have same scale-free behavior but overlap each other in their structures. This fact indicates that the complex network has a concrete skeleton, scale-free tree, as a basic structure.

§1. Introduction

Complex networks have attracted much attention recently because of the 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 $P_d(k)$, the probability for a vertex to have k edges toward other vertices. The emergence of *scale-free* (SF) or power-law degree distribution $P_d(k) \sim k^{-\gamma}$ has been reported in many real-world networks, such as coauthorship networks in social systems,[3)] metabolic networks and protein interaction networks in biological systems,[4),5)] and Internet and World Wide Web in technological systems.[6),7)]

Generally, the structures of networks include many extra edges or shortcuts for a vertex to be connected with other vertices. On the other hand, a tree has only essential edges for connections between vertices, where failure on each vertex leads to large damage on network. Fortunately, all real-world networks have plenty of shortcuts, which guarantee the robustness of the network structure. However, these shortcuts also cause mathematical difficulty in handling networks analytically. For instance, if shortcuts exist, we cannot exactly obtain dynamical properties of networks such as vertex and edge betweenness centrality because a global information of network structure is required.

The spanning tree is a special subgraph which is a tree that includes all vertices of its original network. A spanning tree can be a meaningful structure in itself

[*)] E-mail: hjeong@kaist.ac.kr

because it is the simplest structure to connect all the vertices and can be considered as a critical state of percolation problem. For example, a minimum spanning tree (MST) is a widely accepted concept in weighted networks, to find optimal networks. Although reducing networks into spanning trees may change the properties of the original networks because many edges are to be removed, it is valuable to investigate the properties of spanning trees and compare the spanning trees with the original networks.

Here we focus on the two different realizations of the weights of edges, which mainly determine the construction of spanning trees. The first one is motivated from the concept of communication kernel or backbone network. We assign "edge betweenness centrality" as weights of edges, which can be interpreted as average traffic, and construct the spanning tree which maximizes the total edge betweenness of the tree. The second one is inspired from the microarray experiments in biology, which reflect each gene's relative changes in its expression level under specific conditions. Specifically, in the gene knock-out experiments, one can reconstruct the genetic network by observing the correlation between gene expression levels after a single gene deletion. In the analogy with this, we apply perturbations of a single vertex removal on the network and construct a response network from the correlation of betweenness centrality changes of vertex by using the minimum spanning tree technique.

From these weighted network, we study the structural properties of two relatively different spanning trees of complex networks, so-called "communication kernel" and "response network". First, for the communication kernel, we construct spanning trees with a maximum total weight of edges that is given by edge betweenness centrality. We find that the degree distribution of spanning trees shows scale-free behavior for many model and real-world networks and the degree of the spanning trees has strong correlation with their original network topology. For the response network, we launch an attack on a single vertex which can drastically change the communication pattern between vertices of networks. By using minimum spanning tree technique, we construct the response network based on the measurement of the betweenness centrality changes due to a vertex removal. We find that the degree distribution of the response network also indicates the scale-free behavior as well as that of the communication kernel.

The paper is organized as follows. In §2, we describe the details of the method for finding communication kernel of original networks, and its statistical properties, including the differences and similarities of the spanning tree to its original network. In §3, we present the secondary network which can be constructed from the response of the original network under single vertex removal perturbation. A summary and conclusions are given in §4.

§2. Communication kernel of complex network

In order to extract the communication kernels from networks, we reconstruct the network which consists of relatively important edges in communication between vertices, i.e., edges with high average traffic which is quantified by the *edge betweenness* *centrality*.[8] For the simplicity of the algorithm, we study only undirected networks. In order to select important edges from the communication perspective, we choose the edges according to the priority of their edge betweenness centralities (BCs),[8] the average number of packets passing through the edge. The edge BC is defined as

$$b(i \to j) = \sum_{k \neq l} b_{k \to l}(i \to j) = \sum_{k \neq l} \frac{g_{k \to l}(i \to j)}{g_{k \to l}}, \tag{2·1}$$

where $g_{k\to l}(i \to j)$ denotes the number of shortest paths from the vertex k to l through the edge from the vertex i to j, and $g_{k\to l}$ is the total number of shortest paths from k to l. We construct the spanning trees with maximum total edge BC by using the following procedures:[9] (i) Calculate the edge BC of the network. (ii) Select the edge with the highest edge BC from the unmarked edges in the network and mark it. (iii) Add the selected edge if the selected edge does not create any loop in the tree, otherwise reject it. Unless the tree contains all vertices, return to step (ii).

Following the methods described above, we obtain spanning trees from various networks including the Barabási-Albert (BA) model, the coauthorship network in neuroscience, Internet at autonomous systems (AS) level, and the protein interaction network (PIN) of yeast. We measure the degree distribution of each spanning tree and compare the degree of the spanning tree with that from its original network.

Regardless of details of the construction method, it turns out that all spanning trees show *scale-free* (SF) behavior in their degree distributions [see Fig. 1]. However, the details of degree distribution depend on each original network. The exponents of degree distributions of spanning trees are similar to those of original networks (Table I), but they do not always agree with those of the original networks. The spanning trees of coauthorship networks exhibit truncation or exponential decay similar to the original networks. For the protein interaction network, SF behavior of the spanning tree is relatively clear, even though the original network shows exponential cutoff.

In order to investigate degree correlation between spanning trees and their original networks, we plot degrees from a spanning tree and its original networks (Fig. 2). We find that the degrees of spanning trees (k_s) and their original networks (k) roughly follow the simple relation $k_s \sim k^{\alpha}$, which leads to the degree distribution of the span-

Table I. The scaling exponents and correlation coefficient between the spanning trees and original networks for the BA model, the coauthorship network in neuroscience, Internet at autonomous systems (AS) level, and the protein interaction network (PIN) of yeast. Tabulated for each network is the system size N, the mean degree $\langle k \rangle$, the degree exponent of the spanning trees γ_s and the degree correlation coefficient r between spanning trees and original networks. The ratio of the number of edges between spanning trees and original networks f_0; the ratio of edge BC summation over the edges selected for the spanning tree to total edge BC f_{mst}.

Network	N	$\langle k \rangle$	γ	γ_s	r	f_0	f_{mst}	Ref.
BA model	2×10^5	4	3.0(1)	2.7(1)	0.973	0.5	0.71	11)
Coauthorship	190382	12.5	2.1(1)	2.4(1)	0.538	0.16	0.46	12)
Internet AS	10514	4.08	2.1(1)	2.1(1)	0.929	0.50	0.63	13)
PIN	4926	6.55	3.2(2)	2.3(1)	0.814	0.30	0.54	14)

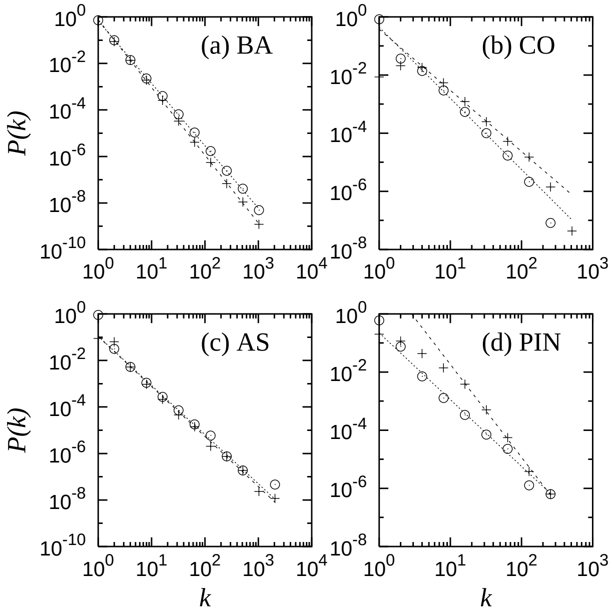

Fig. 1. Spanning trees obtained by MST technique. The degree distribution of spanning trees (○) and original networks (+) for (a) the BA model, (b) the coauthorship network, (c) Internet AS, and (d) PIN.

ning trees $P(k_s) \sim k_s^{-\gamma_s}$, $\gamma_s = (\gamma + \alpha - 1)/\alpha$. α is estimated as 1.0 ± 0.1. In addition, we calculate Pearson's correlation coefficient r between k and k_s,

$$r = \frac{\overline{kk_s} - \bar{k}\bar{k_s}}{\sqrt{(\overline{k^2} - \overline{k}^2)(\overline{k_s^2} - \overline{k_s}^2)}}. \tag{2·2}$$

Most networks exhibit strong correlation between the degree of the spanning tree and its original network. In particular, in the case of spanning trees obtained by using the minimum spanning tree technique, the BA model, Internet and protein interaction network show a very high correlation coefficient, higher than 0.9 [see Table I].

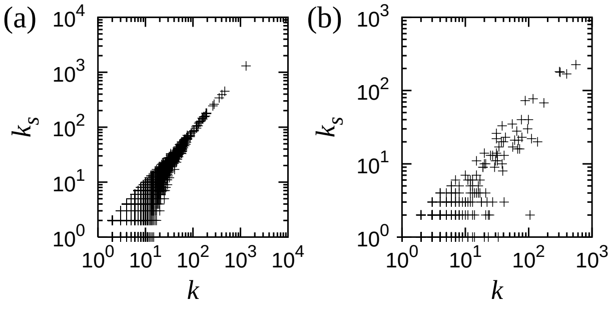

Fig. 2. Scattered plot of the degree of the original network (k) and the spanning tree (k_s) for (a) the BA model and (b) Internet AS.

Because we know that edge BC represents the average traffic over the network, we use the minimum spanning tree technique to find the communication kernel of the network. To verify this, we calculate the ratio f of edge BC summation over selected edges from spanning trees to the summation over all edges in the original networks, and we compare these quantities for the minimum spanning trees [see Table I]. If we select the edges randomly, the ratio f_0 between the edge BC summation over the selected edges and total edge BC summation would be approximately the ratio of the number of edges in the tree to the number of edges in the network, $f_0 = (N-1)/M$, where N is the number of vertices and M is the total number of edges. However, the real set of selected edges from the kernel spanning tree using the minimum spanning tree technique possesses over 50% of the total edge BC of the network and, therefore $f_{mst} \gg f_0$. Thus we can call the spanning trees constructed by using the minimum spanning tree method *the communication kernels*.

Finally, we would mention that the present study of spanning trees only reflects the partial structure of the network, because many redundant shortcuts are not considered. Therefore, to fully understand the whole structure of the network, we should investigate the role of the shortcuts, including loop structures in the networks.[10)]

§3. Response network from perturbation

Typically, the genetic network can be obtained by using the correlation between gene expression levels in the microarray experiment. For instance, a single gene deletion cause the changes of whole gene expression levels, and thus through many gene deletion experiments, one can obtain the correlation between genes. By regarding the genes and the correlation as the vertices and the weight of edges in networks, respectively, a weighted network is defined and often simplified to the binary network using the minimum spanning tree techniques.

We consider the gene deletion as the vertex removal in networks and choose the vertex BC[8)] for the correspondence to the gene expression level because BC is well-defined global quantity which can be affected by any small change of the network structure. Precisely, the vertex BC of vertex k is defined as

$$b(k) = \sum_{i,j} b_{i\to j}(k) = \sum_{i,j} \frac{g^k_{i\to j}}{g_{i\to j}}. \tag{3·1}$$

In Eq. (3·1), $g_{i\to j}$ is the number of geodesic paths from i to j and $g^k_{i\to j}$ is the number of paths from i to j that pass through k.

To construct a response network, we calculate the BC of all vertices in the network and store these results in memory. We denote this original BC of vertex k as $b^{(0)}(k)$. Then, we choose one vertex i from the network and remove this vertex along with all edges that are connected to vertex i. After removing vertex i, we again calculate the BC of all remaining vertices. We denote $b_i(k)$ for the BC of vertex k after vertex i removal and $\Delta b_i(k)$ for the BC difference before and after vertex removal. Because we are only interested in single vertex removal, we restored the removed vertex and repeated this procedure for each vertex in the network.

$$\Delta b_i(k) = b_i(k) - b^{(0)}(k). \tag{3·2}$$

After we calculate $\Delta b_i(k)$ for every vertex k with vertex i removal, we construct a matrix $\boldsymbol{\Delta b} = (b_{ij})$ [see Eq. (3·3)] from the results:

$$\boldsymbol{\Delta b} = \begin{bmatrix} \Delta b_1(1) & \cdots & \Delta b_1(j) & \cdots \\ \vdots & \ddots & \vdots & \\ \Delta b_i(1) & \cdots & \Delta b_i(j) & \\ \vdots & & & \ddots \end{bmatrix}. \tag{3·3}$$

The matrix dimension is $N \times N$. It is similar to the adjacency matrix of a weighted graph with N vertices, edges of which are connected to every vertex in the network with corresponding weight b_{ij}. The weight b_{ij} represents how two vertices i and j are related indirectly and is interpreted as the influence of the removal of vertex i to the vertex j, which is very analogical to the gene expression level change caused by the single gene knock-out in microarray experiments.

From this adjacency matrix $\boldsymbol{\Delta b}$, here we build the *secondary network* by using the minimum spanning tree technique.[9] Because it is reasonable to connect two vertices with higher correlation (Δb_{ij}), in our simulation we choose to connect the edge with largest weight first, to find the substructure which represents the maximum influential network. We put an edge between a vertex and its most influential vertex, with a constraint that every set of N vertices must be connected with only $(N-1)$ edges by following the regular scheme of the minimum spanning tree construction.

From the degree distribution of secondary networks, we find that secondary networks also show *scale-free* behavior [see Fig. 3]. However, the exponents are very different from the degree exponent of the original network. The network constructed by using MST method shows the exponent of 2.2, which is far from 3.0, the degree exponent of the original BA network. The degree exponent of secondary networks is similar to the exponent of BC distribution of the BA model. This might be related

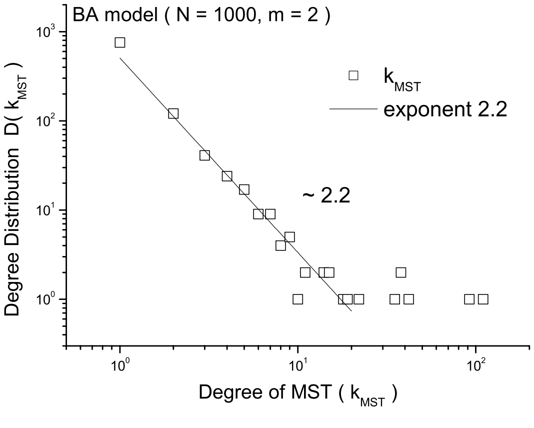

Fig. 3. Degree distribution of the secondary network constructed by using the MST method.

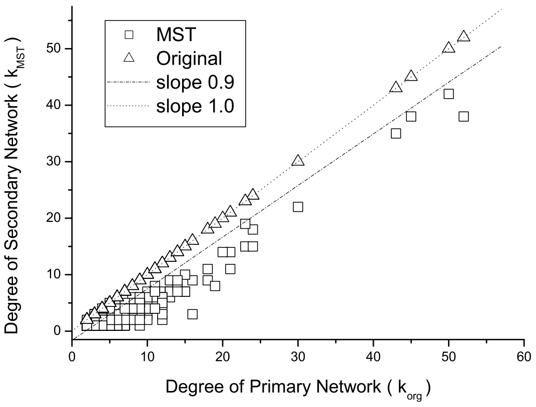

Fig. 4. Relation between degrees of the original network and the secondary network.

to the fact that BC changes governing secondary networks are nearly proportional to the BC of the original networks.

We find that the resulting secondary networks have structural similarity to the original network on comparing local properties of secondary networks, such as degree and nearest neighbors, to those of the original network. The degrees of those networks show strong correlation in Fig. 4.

§4. Summary

We investigate the structural properties of the spanning trees of various model and real-world networks. We construct spanning trees by using the minimum spanning tree technique resulting in communication kernels of the original networks. We find that the degree distribution of the spanning trees shows scale-free behavior for many model and real-world networks. In addition, we find that the degree of the spanning trees has strong correlation with their original network topology and the scale-free behavior does not depend on details of the construction method employed for the spanning trees. And we study BC changes under a vertex deletion in the

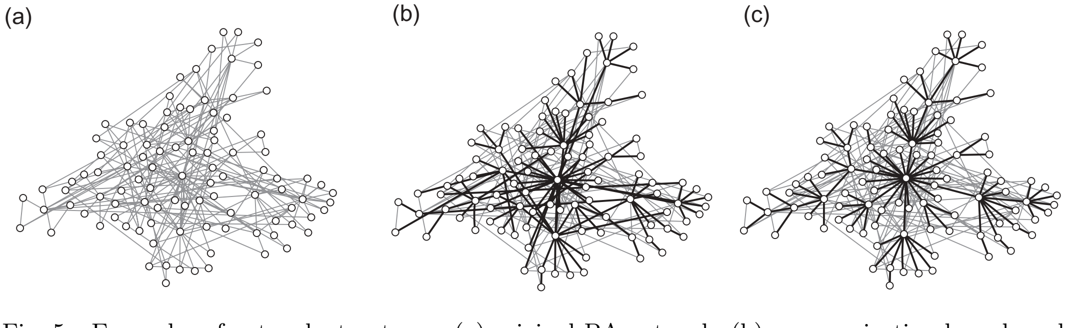

Fig. 5. Examples of network structures: (a) original BA network, (b) communication kernel, and (c) response network. The thick and thin lines indicate the edges included and excluded in spanning trees, respectively.

BA model. We find that BC changes follow the power-law distribution, and the secondary networks constructed by using MST method have similar local structure to their original networks. Strong correlation between unperturbed BC and BC changes of a vertex gives rise to the power-law distribution of BC changes. Local similarity of secondary networks and original networks indicates that the deletion of a vertex greatly affects BCs of nearest neighbors. Moreover, it is interesting to note that MSTs from two different methods, communication kernel and secondary network overlap significantly if we start from the same original network. In Fig. 5, overlapping between two MST is about 72%, which indicates that secondary network constructed from vertex removal perturbation is indeed important network, communication kernel of original network. In many biological system, we do not have complete information about the topology of the underlying network, therefore we have to use indirect method to find out the topology of the network. Typical examples include gene knock-out experiment using microarray, which is exactly what we did to find out the underlying genetic network for constructing secondary network. We believe that our theoretical method can shed a light on uncovering the structure of the biological networks.

Acknowledgements

This work was supported in part by grant No. R08-2003-000-10285-0 from the Basic Research Program of KOSEF and Ministry of Science and Technology through Korean Systems Biology Research Grant (M10309020000-03B5002-00000).

References

1) R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74** (2002), 47.
2) S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51** (2002), 1079.
3) M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98** (2001), 404.
4) H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási, Nature **407** (2000), 651.
5) H. Jeong, S. P. Mason and A.-L. Barabási and Z. N. Oltvai, Nature **411** (2000), 41.
6) M. Faloutsos, P. Faloutsos and C. Faloutsos, Comput. Commun. Rev. **29** (1999), 251.
7) R. Albert, H. Jeong and A.-L. Barabási, Nature **401** (1999), 130.
8) M. L. Freeman, Sociometry **40** (1977), 35.
U. Brandes, J. Math. Sociol. **25** (2001), 163.
M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99** (2002), 7821.
9) J. B. Kruskal, Proc. Amer. Math. Soc. **7** (1956), 48.
R. C. Prim, Bell Syst. Tech. J. **36** (1957), 1389.
10) D.-H. Kim, J. D. Noh and H. Jeong, Phys. Rev. E **70** (2004), 046126.
11) A.-L. Barabási and R. Albert, Science **286** (1999), 509.
12) A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek, Physica A **311** (2002), 590.
13) D. Meyer, *University of Oregon Route Views Archive Project*, http://archive.routeviews.org.
14) K.-I. Goh, B. Kahng and D. Kim, q-bio.MN/0312009.