Characterization of topological structure on complex networks

Ikuo Nakamura*

Sony Corporation, 2-10-14 Osaki, Shinagawa, Tokyo, Japan
and Computer Science Department, Stanford University, Stanford, California 94305, USA

(Received 30 April 2003; published 28 October 2003)

Characterizing the topological structure of complex networks is a significant problem especially from the viewpoint of data mining on the World Wide Web. ‘‘Page rank’’ used in the commercial search engine Google is such a measure of authority to rank all the nodes matching a given query. We have investigated the page-rank distribution of the real Web and a growing network model, both of which have directed links and exhibit a power law distributions of in-degree (the number of incoming links to the node) and out-degree (the number of outgoing links from the node), respectively. We find a concentration of page rank on a small number of nodes and low page rank on high degree regimes in the real Web, which can be explained by topological properties of the network, e.g., network motifs, and connectivities of nearest neighbors.

DOI: 10.1103/PhysRevE.68.045104 PACS number(s): 89.75.Fb, 05.70.Fh, 87.23.Kg, 89.20.Hh

The World Wide Web (WWW) is a gigantic collection of information which is organized by links between Web pages. It is considered as a directed graph where nodes are Web html pages and directed links are hyperlinks between pages in natural manner. One of the topological properties of the Web is a power law behavior for the degree distribution, $P(k) \sim k^{-\gamma}$, where the degree k is the number of hyperlinks for a given page [1]. Recent studies show that such a power law distribution is a common feature of many complex networks, i.e. food networks [2], metabolic network [3], citation networks [4], Internet (hardware) [5], and WWW [6]. Recent empirical measurement reports that the in-degree (the number of incoming links to the node) and out-degree (the number of outgoing links from the node) distributions of the Web are approximately scale-free in form with exponents $\gamma_{in} = 2.1$ and $\gamma_{out} = 2.7$ [6], although there is some deviation for small degree. The number of average in-degree and out-degree per node $D \approx 7.5$ is also reported. However, this global network property gives no information about identification of a given node and its neighbors. Characterizing the topological structure of the Web is motivated by some large-scale web applications, i.e., data mining on the Web. One fundamental and challenging issue is to extract what you want from the Web which consists of billions of pages. Along with query matching technique, evaluating the relevance of the pages is also important. ‘‘Page rank’’ developed by the commercial search engine Google [7,8] is such a topological measure of authority of the Web to rank all pages matching a given query. In Google, the searching result by keywords is shown in order by page rank x_n given by

$$x_n = d \sum_{m \in V_n} \frac{x_m}{j_m} + (1-d), \tag{1}$$

where $d \in (0,1)$ is called *dampening factor*, j_m is the number of links (hyperlinks) coming out from node (page) m, and V_n is a set of nodes that points to node n [9]. Equation (1) is written by asymmetric adjacency matrix A and x_n can be expressed as the dominant eigenvector of A. A diagram of consistent incoming/outgoing flows of page rank can be described from the dominant eigenvector (see Fig. 1). If there is a node with no outgoing links, page rank of the node leaks out and might go to zero. Otherwise, if an outgoing link only points to itself, all page rank could concentrate on that node. The dampening factor not only avoid these page rank leak and concentration but also takes user browsing behavior into account, where a user does not click on an infinite number of links, but gets bored sometimes and jumps to another page at random. Intuitively, a node has a high page rank if there are many nodes that point to it or if there are some nodes with high page rank that point to it. Assume that links are randomly connected, then one expects page rank might depend on their in-degree since a given link has a similar weight as other links. However, recent empirical data shows that there is very little correlation between page rank and in-degree (out-degree) distributions, except for nodes with very high in-degree [10]. It means that nonrandom characteristics of topological properties is crucial for page rank.

In this paper, we investigate the page rank distribution of the real Web and a theoretical model of growing networks to study the topological features of the networks. Using net-

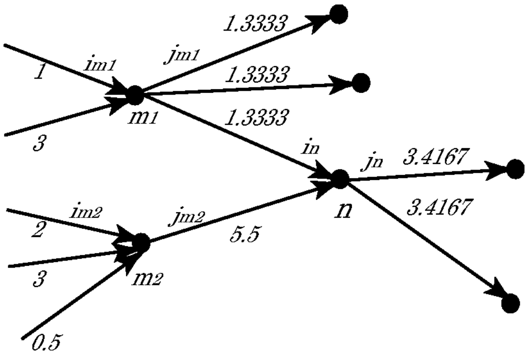

FIG. 1. Illustration of page rank transition diagram. The value shown below the directed link represents the flow of page rank. It is given by a sum of incoming flow of page rank. A sum of incoming flows is equal to a sum of outgoing ones. In this figure, the dampening factor d is set to zero.

*Electronic address: ikuo@arch.sony.co.jp work motifs and a rewiring algorithm, we construct a network model which may reproduce characteristics of the real Web and follow the characteristics of empirical results.

First of all, we show the in-degree vs page rank distribution in the real Web derived from the empirical data of Notre Dame University domain [1] [Fig. 2(1a)]. Note that the average page rank of a node is 1 because their sum is normalized to the number of nodes N. The dampening factor $d = 0.15$ is chosen for both empirical data and theoretical model, which is a typical value reportedly applied in practical use. As page rank at a given node is represented by a sum of incoming flows from nearest neighbors, they are thought to depend on in-degree distribution. Out-degree at that node is also correlated since the page rank leaks from the outgoing links. The data were collected by a software agent that follows all hyperlinks on a node and recursively follows these to retrieve the related node, therefore none of the nodes has in-degree $i=0$ which cannot be retrieved from any node. Note that the nodes with in-degree $i=0$ have always zero page rank since its flow goes out from the node and never returns. The figure shows that the page rank distribution basically depends on in-degree, and most of the page rank concentrates on a small number of nodes in the network. The rate of nodes with zero page rank is $z_1 = 0.951$. The average page rank and its standard deviation [Fig. 2(2a)] indicates that they follow a power law distribution and the statistical fluctuation increases in the high in-degree region. Our result with respect to the in-degree-page-rank correlation does not agree with the observation obtained in Ref. [10]. The probability $p(i,j)$ that a randomly chosen node has in-degree i and out-degree j is a total degree of a node given by $i+j$, there is little correlation between i and j (Fig. 3). Moreover we consider some topological characteristics which we call network motifs [11], patterns of interlinks which consist of m nodes occurring at numbers significantly higher than those in random networks. In particular, the following network motifs are noticeable since they are possible candidates for the part of high page-rank nodes. For $m=1$ motif, a self-link, a directed link which ends into itself, can reduce the outgoing flow to itself which could increased its ranking. Empirical data shows self-link per link $s_1 = 0.0183$, whereas $s_1 \approx D/N$ for random wired network. The $m=2$ motifs include mutual links, a directed link and opposite directed link between two nodes, and multiple links, more than one link in the same direction between two nodes. The former can reduce the outgoing flow to itself by way of another node so as the self-link [12]. For example, the two top ranking nodes in Fig. 2(1a) have about 400 in-degree and one out-degree mutually linked to one another. The rate of multiple links per link is $s_2 = 4.77\times10^{-3}$, mutual links per link is $s_3 = 0.117$, compared with both $2D/N^2$ for a random wired network. For $m \geqslant 3$, numerous patterns of network motifs can be considered. These high ranking nodes should be considered when we build a theoretical model. In the second, we construct a growing model with directed link using an algorithm in Ref. [13] to generate a directed scale-free network such as the Web. Although there are nodes with either zero in-degree or zero out-degree in the real Web, the original growing model [13] only has nodes with zero in-degree. As the nodes with in-degree $i=0$ have no influence (the page rank is always zero) on the page-rank distribution, we transform this model to make nodes with out-degree $j=0$, ignoring nodes with in-degree $i=0$. Starting with a single node, at each step we have the following.

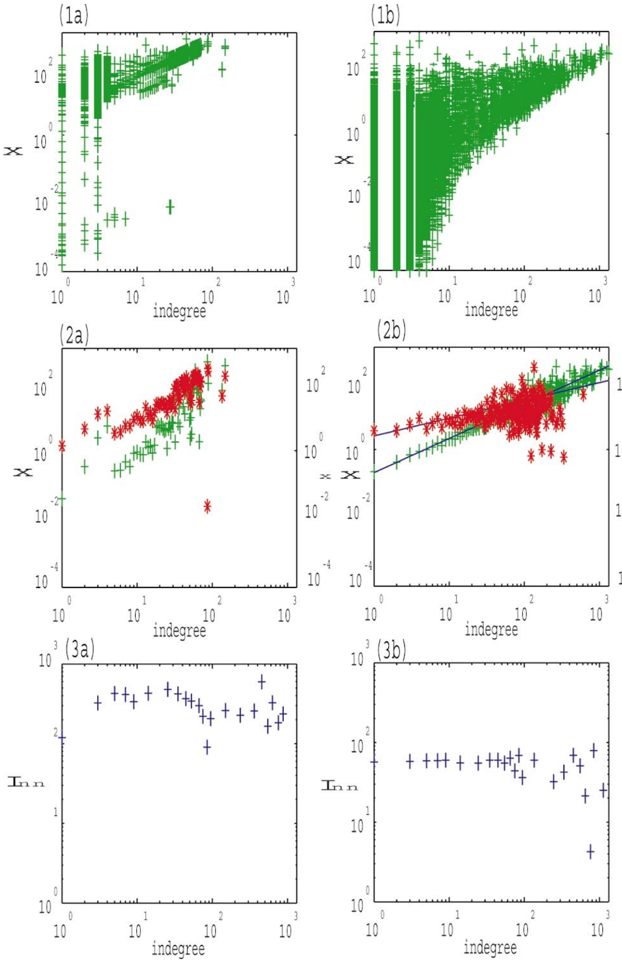

FIG. 2. (Color) (1a) Log-log plot page-rank distribution x as a function of in-degree for the real Web (dataset of Notre Dame University domain, $N=3.26\times10^5, D=4.60$, the rate of nodes with $j=0$, $z_0=0.5765$). In a large matrix computation, we sometimes need thousands of iterations to get a dominant vector, since the difference between dominant eigenvalue and other eigenvalues could be considerably small and the convergence is slow. We made 5×10^4 iterations in this case. (1b) Log-log plot of page-rank distribution x for the growing network model with $N=10^5$, $D=4.72$, $\lambda=3.20$, and $\mu=0.75$. The rate of nodes with $j=0$, $z_0=0.58$. (2a) Average page rank $\overline{x}(+)$ and standard deviation $\sigma_x(*)$ as a function of in-degree for the real Web. (2b) $\overline{x}(+)$ and $\sigma_x(*)$ for growing network model. They are fitted by the form $\overline{x}=2.1\times10^{-1}i^{1.01\pm0.01}$ (Solid line) and $\sigma_x=2.5\times10^{-1}i^{0.52\pm0.02}$ (dotted line). The scattered points of σ_x for large i are due to statistical fluctuations. (3a) In-degree—nearest neighbor average connectivities of in-degree correlation function $I_{nn}(i)$ for the real Web. (3b) $I_{nn}(i)$ for the growing network model. A specific correlation between in-degree of a node and I_{nn} is not observed except statistical fluctuation in the high in-degree regime.

(1) With probability p, a new node is created and a directed link from an existing target node to it is set up. The

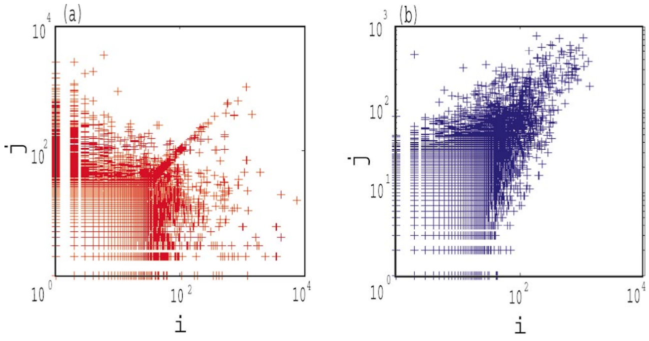

FIG. 3. (Color) The distribution of nodes with in-degree *i* and out-degree *j*. (a) Empirical data and (b) original growing model.

probability that existing node with out-degree *j* links to a newly introduced node is defined as $A_j=j+\lambda$.

(2) With probability $q=1-p$, a directed link is created between existing nodes. The link creation rate is assumed to depend only on the out-degree of the node *j* which it emanates and the in-degree of the target node *i*, which is defined as $C(j,i)=(j+\lambda)(i+\mu)$. The probability $p(i,j)$ is, in general, not equal to the product $p_i p_j$ of the separate distribution. However, its numerical distribution is not close to the empirical result which seems to be rather a separate distribution (Fig. 3). Hence we adapted the separate distribution $p_i p_j$ by making the out-degree of nodes reconfigured. The page-rank distribution for in-degree is shown in Fig. 2(1b) with relevant parameters, $\lambda=3.20$, $\mu=0.75$, $p=0.21$, to match the empirical values. The maximum value of page rank is almost independent of in-degree whereas the minimum limit increases as growing in-degree. No concentration on small numbers of nodes is found, which is different from empirical data where most of the nodes have page rank of zero. The average page rank $\overline{x}$ and its standard deviation σ_x follow a power law distribution $\overline{x}\sim i^{1.00\pm0.01}$ and $\sigma_x\sim i^{0.52\pm0.02}$ [Fig. 2(2b)]. We here prohibit self-link connection and multilink connection, which would be discussed later when the network motifs are introduced. To evaluate the difference between the real Web and theoretical model, we add some topological features to the original model by embedding network motifs, i.e., self-links, mutuallinks, multiple links, and nearest neighbor power law distribution obtained by the link rewiring algorithm. The probability $P_1(j,i)$ that a randomly chosen link starts from a node with out-degree *j* and ends into the one with in-degree *i* has a random link distribution, i.e., separate distribution of *i* and *j*. Assume that $P_1(j,i)$ is randomly distributed, given by Bayiesan statistics, $P_1(j,i)=P(j)P(i|j)=Cj^{-\gamma_{out}}i^{1-\gamma_{in}}$ with normalized constant *C*. The conditional probability $P(i|j)$ denotes that a randomly selected link ends into a node with in-degree *i*, provided that it emanates from node with out-degree *j*. However, the probability $P(j,i)$ is not randomly distributed in the real Web, for example, a trusted page (with many in-degrees) tends to link to other trusted ones, a few but selected links tend to link to trusted pages rather than many but nonselective links link to worthless pages. As is pointed out in Ref. [14], the nearest neighbor's average connectivities of undirected link exhibits a power law dependence on the connectivity degree. But no discussion about directed links has been made so far. Hence we have measured in-degree vs average in-degree of nearest neighbors, which is crucial from the page rank point of view since the page rank at a given node is determined by those from nearest neighbors. We see in Fig. 2(3a) that there is a small enhancement around intermediate regime $i\sim10$ and suppression among low and high in-degree regime. To provide these nonrandom link distribution, we introduce a rewiring algorithm of existing two links between nodes. A correlation function $P_2^{(0)}(i_m,i_n)$ between the in-degree of origin node *m* and the in-degree of target node *n* is written as

$$P_2^{(0)}(i_m,i_n)=\sum_j p(i_m,j)P(i_n|j). \qquad (2)$$

Other pairs of nearest neighbor correlations (in-out, out-in, and out-out degree correlation) can be also deduced from similar definitions. First of all, we randomly choose pairs of links $a\rightarrow b$ and $c\rightarrow d$. If $i_a>i_c$ and $i_b>i_d$, or if $i_a<i_c$ and $i_b<i_d$, these two links are rewired as $a\rightarrow d$, $c\rightarrow b$, provided that none of these links already exist. Otherwise, we rewire with probability $\exp(-\Delta H/T)$, where $\Delta H=-(i_a-i_c)(i_b-i_d)$ and *T* is an annealing parameter to prevent the system from getting trapped in a local minimum. When a rewiring algorithm is performed *l* times per node, the link probability distribution is approximately given by

$$P_2^{(l)}(i_a,i_d)=\sum_{i_b,i_c} P_2^{(l-1)}(i_a,i_b)P_2^{(l-1)}(i_c,i_d)[\theta(-\Delta H) + \theta(\Delta H)\exp(-\Delta H/T)], \qquad (3)$$

where $\theta(x)$ is a step function. This rewiring algorithm conserves the distribution of $p(i,j)$. The nearest neighbor average in-degree connectivity of node *n* given by $i_{nn}(n)=(1/j_n)\Sigma_{m\in\mathcal{V}}i_m$, which can be reformulated in the probabilistic representation as

$$I_{nn}^{(l)}(i_1)=\frac{1}{P(i_1)}\sum_{i_2} i_2P_2^{(l)}(i_1,i_2). \qquad (4)$$

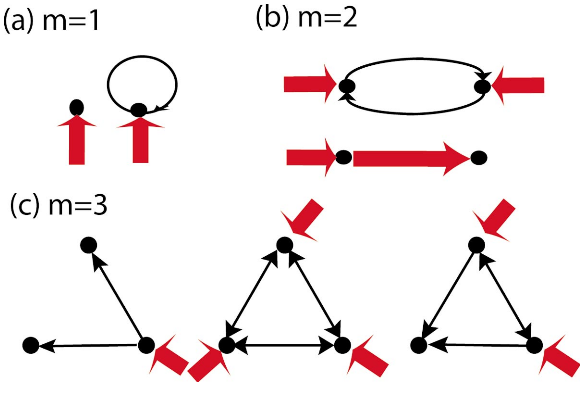

FIG. 4. (Color) Possible high page-rank nodes with *m* network motifs. Thick line with arrow shows dozens of incoming links.

As it is of no importance whether the system is in global minimum since the Web does not always reside in the minimum point of some dynamics, we here take $T=0$. From Eq. (4) $I_{nn}^{(0)}(i) \sim$ const and one rewiring per node makes the curve of distribution $I_{nn}^{(1)}(i)$ a similar shape to that observed in the real Web.

We use the rewiring technique shown above and other rewirings for generating the network motifs, of which the rate of appearance s_1, s_2, s_3 follow the empirical data. Though we repeated more than 10^2 times the simulations for $N=10^4$ network, the concentration of page rank on small nodes is rarely observed as far as nodes which become a part of motifs are randomly chosen. Instead, we find that the concentration occurs when originally high page-rank nodes have a small number of outgoing links and mutual links to selected high page rank nodes, put up mutual link exclusively or outgoing links to node with no out-degree. Most of the top 100 ranking nodes observed in the empirical data have such a topological characteristic. Some typical patterns of very high score which consists of *m* nodes are described in Fig. 4. In other words, page rank is robust when originally low page rank nodes try to get high ranking by putting up arbitrary links to itself, but become vulnerable to intentional manipulation of originally high rank nodes. Using the growing model, some limiting cases are also investigated. For example, as the rate of mutual links s_1 is growing, σ_x is suppressed. A weight of a given link becomes similar to other links as the mutual-link rate increases, since the flow of page rank goes all over the network and returns to itself. In the limit $s_1=1$, σ_x has minimum value (it goes to zero if *d* is zero) and *x* is exactly proportional to *i*.

In conclusion, we have investigated the page-rank distribution of the real Web and the theoretical network model. In the real Web, we find that the page rank specifically depends on in-degree and they may concentrate on a small number of nodes. This phenomenon can be explained by link manipulation of high ranking node, such as self-link, multiple-link, mutual-link connections. As there are still higher degrees of motifs in the real Web, the influence of higher degree motifs on the network properties should be discussed in future work. To characterize the nodes in the network, page rank is well known to be valid in ranking order of the Web. This method could be also useful for other applications, i.e., metabolic flow network and propagational investment currency system (in which some value can propagate from node to node in the network), to rank the specific quantity of given nodes.

[1] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).

[2] J.M. Montoya and R.V. Sole, J. Theor. Biol. **214**, 405 (2002).

[3] D.A. Fell and A. Wagner, Nat. Biotechnol. **18**, 1121 (2000).

[4] S. Redner, Eur. Phys. J. B **4**, 131 (1998).

[5] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[6] A. Broder *et al.*, Proceedings of the Ninth World Wide Web Conference, 2000 (unpublished).

[7] L. Page and S. Brin, Proceedings the Seventh International World Wide Web Conference, 1998 (unpublished).

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, Stanford Digital Libraries Working Paper, 1998 (unpublished).

[9] In Google, page rank only cannot get us high rankings. Assume that a search on Google returns 10^5 result. The search engine does not calculate every factor for each of them. It first gets a subset of documents that are most likely to be related to the query. That is, a number of subsets of documents are stored in database beforehand and it queries the whole database using a few factors. Second, it applies all the factors to those, e.g., 10^3 pages in a given subset.

[10] G. Pandangan, P. Raghavan, and E. Upfal, Proceedings of the Eighth World Wide Web Conference, 2002 (unpublished).

[11] R. Milo *et al.*, Science **298**, 824 (2002).

[12] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[13] P.L. Krapivsky, G.J. Rodgers, and S. Redner, Phys. Rev. Lett. **86**, 5401 (2001).

[14] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001); A. Vazquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).