ELSEVIER

Physica A 281 (2000) 69–77

PHYSICA A

www.elsevier.com/locate/physa

Scale-free characteristics of random networks: the topology of the world-wide web

Albert-László Barabási*, Réka Albert, Hawoong Jeong

Department of Physics, College of Science, 225 Nieuwland Science Hall, University of Notre-Dame, Notre-Dame, IN 46556, USA

Abstract

The world-wide web forms a large directed graph, whose vertices are documents and edges are links pointing from one document to another. Here we demonstrate that despite its apparent random character, the topology of this graph has a number of universal scale-free characteristics. We introduce a model that leads to a scale-free network, capturing in a minimal fashion the self-organization processes governing the world-wide web.

PACS: 84.35.+i; 64.60.Fr; 87.23.Ge

Keywords: Disordered systems; Networks; Random networks; Critical phenomena; Scaling; World-wide web

1. Introduction

The emergence of order in natural systems is a constant source of fascination and inspiration for both physical and biological sciences. While the spatial order characterizing for example the crystals has been at the basis of many advances in contemporary physics, most complex systems in nature do not offer such high degree of order. In fact, many systems around us display rather complex topologies, that often seem random and unpredictable [1,2]. In particular, many of these systems form complex networks, whose vertices are the elements of the system and edges represent the interactions between them. For example, living systems form a huge genetic network, whose vertices

* Corresponding author. Fax: +1-219-631-5952.
E-mail address: alb@nd.edu (A.-L. Barabási)

0378-4371/00/$ - see front matter
PII: S0378-4371(00)00018-2

are proteins, the edges representing the chemical interactions between them [3]. Similarly, a large network is formed by the nervous system, whose vertices are the nerve cells, connected by axons [4]. But equally complex networks occur in social science, where vertices are individuals or organizations, and the edges characterize the social interactions between them [5], in the business world, where vertices are companies and edges represent diverse trade relationships, or describe the world-wide web (www) whose vertices are HTML documents connected by links pointing from one page to another [6,7]. Due to their large size and the complexity of the interactions, the topology of these networks is largely unknown or unexplored.

A major step in the direction of understanding the generic features of network development was the recent discovery of a surprising degree of self-organization characterizing the large scale properties of complex networks. Exploring several large databases describing the topology of large networks, that span as diverse fields as the www or the citation patterns in science, recently we demonstrated [8] that independently of the nature of the system and the identity of its constituents, the probability $P(k)$ that a vertex in the network is connected to k other vertices decays as a power-law, following $P(k) \sim k^{-\gamma}$. These results offered the first evidence that large networks self-organize into a scale-free state, a feature unexpected by all existing random network models.

In this paper we illustrate the emergence of self-organization and scaling in random networks through one important example, that of the world-wide web. We show that the incoming and outgoing link distribution of the www documents follows a power law, an indication of the scale-free nature of the network. This understanding of the network topology allows us to determine the average distance between two randomly chosen documents, or the diameter of the www. Finally, we present a model that naturally leads to a power-law distribution, explaining the mechanism responsible for the development of the scale-free state.

2. The topology of the world-wide web

Despite its increasing role in communication, the world-wide web remains the least controlled medium: any individual or institution can create websites with unrestricted number of documents and links. This unregulated growth leads to a huge and complex web, which is a large directed graph, whose vertices are documents and edges are the links (URLs) pointing from one document to another. The topology of this graph determines the web's connectivity and, consequently, our effectiveness in locating information on the www. However, due to its large size (estimated to be at least 8×10^8 documents [9,10]), and the continuously changing documents and links, it is impossible to catalogue all vertices and edges. While great efforts are made to map and characterize the Internet's infrastructure [11], little is known about what truly matters in searching for information, i.e., about the topology of the www.

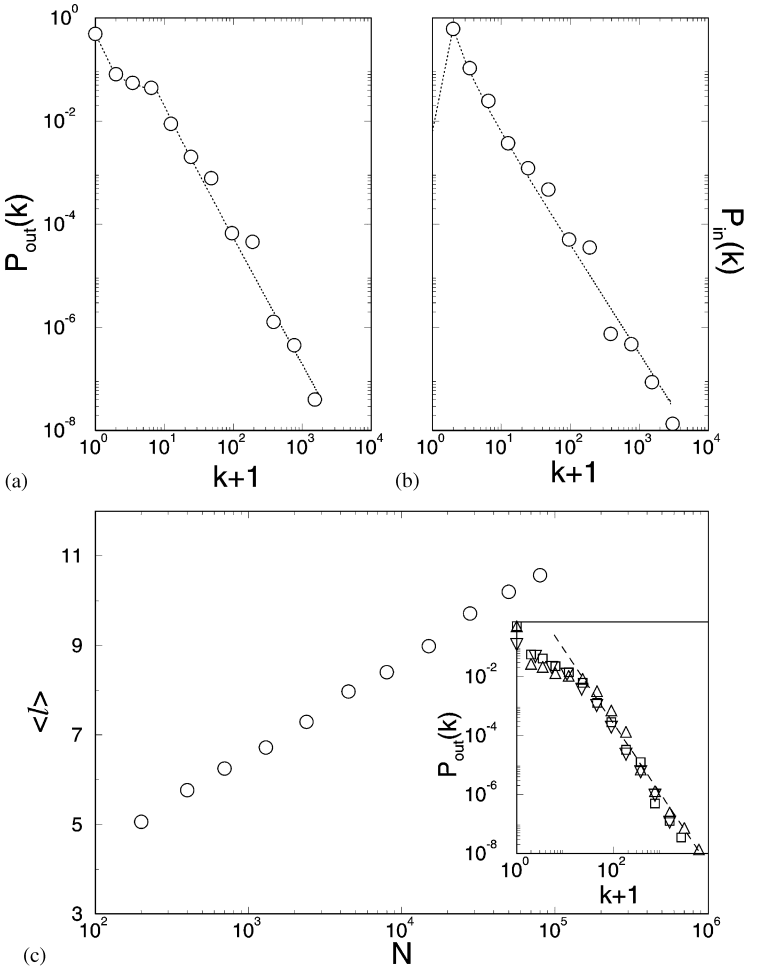

Fig. 1. The distribution of (a) outgoing links (URLs found on an HTML document) and (b) incoming links (URLs pointing to a certain HTML document). The data were obtained from the complete map of the *nd.edu* domain, that contains 325,729 documents and 1,469,680 links. The dotted lines in (a) and (b) represent the analytical fits we used as input distributions in constructing the topological model of the www, the tail of the distributions following $P(k) \sim k^{-\gamma}$, with $\gamma_{\text{out}} = 2.45$ and $\gamma_{\text{in}} = 2.1$. (c) Average of the shortest path between two documents as a function of the system size, as predicted by the model. As a check of the validity of our predictions, we have determined ℓ for documents in the domain *nd.edu*. The measured $\langle \ell_{nd.edu} \rangle = 11.2$ agrees well with the prediction $\langle \ell_{3\times 10^5} \rangle = 11.6$ obtained from our model. To show that the power-law tail of $P(k)$ is a universal feature of the www, in the inset we show $P_{\text{out}}(k)$ obtained by starting from *whitehouse.gov* (squares), *yahoo.com* (upward triangles) and *snu.ac.kr* (downward triangles). The slope of the dashed line is $\gamma_{\text{out}} = 2.45$, as obtained from *nd.edu* in (a).

To determine the local connectivity of the www, we constructed a robot, that adds to its database all URLs found on a document and recursively follows these to retrieve the related documents and URLs. From the collected data we determined the probability $P_{\text{out}}(k)$ ($P_{\text{in}}(k)$) that a document has k outgoing (incoming) links. As Figs. 1a and b

illustrate, we find that both $P_{\text{out}}(k)$ and $P_{\text{in}}(k)$ follow a power law over many orders of magnitude [1]

$$P_{\text{out}}(k) \sim k^{-\gamma_{\text{out}}} \tag{1}$$

and

$$P_{\text{in}}(k) \sim k^{-\gamma_{\text{in}}} . \tag{2}$$

This distribution is remarkably different not only from the Poisson distribution predicted by the classical theory of random graphs by Erdős and Rényi [12,13], but also from the bounded distribution found in recent models of random networks [14]. The power-law tail indicates that the probability of finding documents with a large number of links is rather significant, the network connectivity being dominated by highly connected web pages. The same is true for the incoming links: the probability of finding very popular addresses, to which a large number of other documents point, is non-negligible, an indication of the flocking sociology of the www. Furthermore, while the owner of each web page has complete freedom in choosing the number of links on a document and the addresses to which they point, the overall system obeys scaling laws characteristic only of highly interactive self-organized systems and critical phenomena [15].

To investigate the connectivity and the large-scale topological properties of the www, we construct a directed random graph consisting of N vertices, assigning to each vertex k outgoing links, such that k is drawn from the power-law distribution shown in Fig. 1a. These links are randomly connected to the other vertices, with the constraint that the number of incoming links per document follows the distribution shown in Fig. 1b. A particularly important quantity in a search process is the shortest path between two documents, ℓ, defined as the smallest number of URL links one needs to follow to navigate from one document to the other. As Fig. 1c shows, we find that the average of ℓ over all pairs of vertices follows

$$\langle \ell \rangle = 0.35 + 2.06 \log(N) , \tag{3}$$

indicating that the web forms a small-world network [14,16–18], known to characterize social or biological systems. Using $N = 8 \times 10^8$ [9,10], we find $\langle \ell_{www} \rangle = 18.59$, i.e., two randomly chosen documents on the web are on average 19 clicks away from each other. Since for a given N, ℓ follows a Gaussian distribution, $\langle \ell \rangle$ can be interpreted as the *diameter* of the web, a measure of the shortest distance between any two points in the system. Despite its huge size, our results indicate that the www is a highly connected graph of average diameter of only 19 links. The logarithmic dependence of $\langle \ell \rangle$ on N is important to the future potential of the www: we find that the expected 1000% increase in the size of the web over the next few years will change $\langle \ell \rangle$ from 19 to only 21. The relatively small value of ℓ suggests that an intelligent agent, i.e., who can interpret the links and follow only the relevant one, can find in a short time the desired information by navigating the www. However, this is not the case for a robot,

[1] Note that after the completion of this work a number of other groups [19–21] have arrived independently to the same conclusion, observing power-law scaling in the connectivity distribution.

that locates the information based on matching strings: we find that such a robot, aiming to identify a document at distance $\langle \ell \rangle$, needs to search $M(\langle \ell \rangle) \simeq 0.53N^{0.92}$ documents which, using $N = 8 \times 10^8$ [9,10], leads to $M = 8 \times 10^7$, i.e., to 10% of the full www. This indicates that robots cannot benefit from the highly connected nature of the web, their only successful strategy being indexing as large a fraction of the www as possible.

3. The scale-free model

While the model we used to estimate the diameter of the web has the same scale-free nature as the www, it does not answer an important question: what is the mechanism that leads to these power-law distributions in the first place? To answer this question we need to design a model that does not have as an input the power-law scaling, but through some dynamical processes leads to a network that has the same scale-free properties as the www.

A common feature of the earlier network models, such as the Erdős–Rényi (ER) [12,13] or the Watts–Strogatz (WS) [14] model is that they both predict that the probability distribution of the vertex connectivity, $P(k)$, has an exponential cutoff, and has a characteristic size $\langle k \rangle$, that depends on p. In contrast, as we demonstrated in the previous section, for the www $P(k)$ is free of scale, following a power-law distribution over many orders of magnitude. To understand the origin of this discrepancy, we have recently suggested that there are two generic aspects of real networks that are not incorporated in these models [8]. First, the current network models assume that we start with a fixed number (N) of vertices, that are then randomly connected or reconnected, without modifying N. In contrast, most real world networks are *open*, i.e., they form by the continuous addition of new vertices to the system, thus the number of vertices, N, increases throughout the lifetime of the network. For example, the www grows exponentially in time by the addition of new web pages. Consequently, the *network continuously expands by the addition of new vertices* that are connected to the vertices already present in the system.

Second, the random network models assume that the probability that two vertices are connected is random and uniform. In contrast, most real networks exhibit *preferential connectivity*. For example, a newly created webpage will more likely include links to well known, popular documents with already high connectivity. This example indicates that the probability with which a new vertex connects to the existing vertices is not uniform, but there is a *higher probability to be linked to a vertex that already has a large number of connections.*

A simple model incorporating only these two ingredients naturally leads to the observed scale invariant distribution. The model is defined in two steps:

(1) *Growth*: Starting with a small number (m_0) of vertices, at every timestep we add a new vertex with $m(\leqslant m_0)$ edges (that will be connected to the vertices already present in the system).

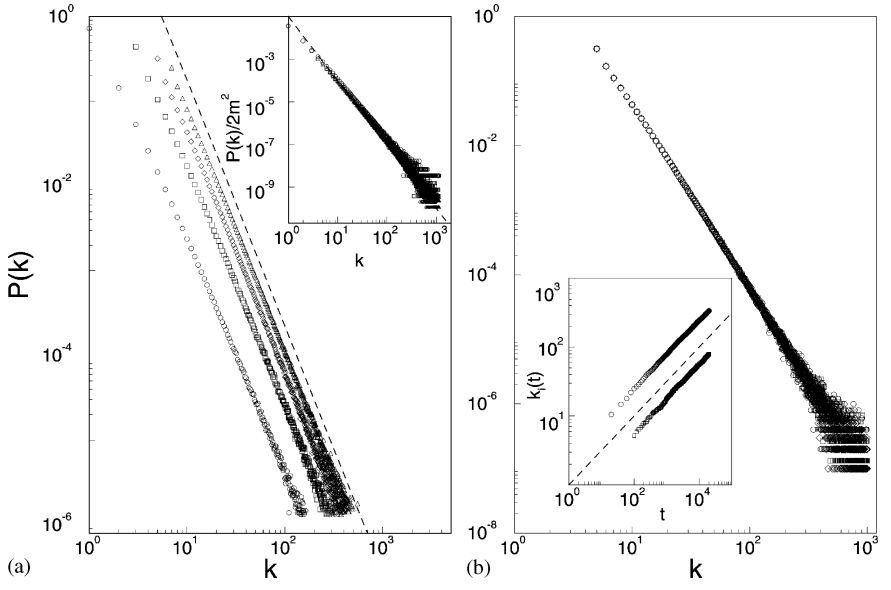

Fig. 2. (a) Connectivity distribution of the model, with $N = m_0 + t = 300\,000$ and $m_0 = m = 1$ (circles), $m_0 = m = 3$ (squares), $m_0 = m = 5$ (diamonds) and $m_0 = m = 7$ (triangles). The slope of the dashed line is $\gamma = 2.9$. The inset shows the rescaled distribution (see text) $P(k)/2m^2$ for the same values of m, the slope of the dashed line being $\gamma = 3$. (b) $P(k)$ for $m_0 = m = 5$ and system sizes $N = 100\,000$ (circles), $N = 150\,000$ (squares) and $N = 200\,000$ (diamonds). The inset shows the time-evolution for the connectivity of two vertices, added to the system at $t_1 = 5$ and $t_2 = 95$. Here $m_0 = m = 5$, and the dashed line has slope 0.5, as predicted by Eq. (5).

(2) *Preferential attachment*: When choosing the vertices to which the new vertex connects, we assume that the probability Π that a new vertex will be connected to vertex i depends on the connectivity k_i of that vertex, such that

$$\Pi(k_i) = k_i / \sum_j k_j \,. \tag{4}$$

After t timesteps the model leads to a random network with $N = t + m_0$ vertices and mt edges. As Fig. 2a shows, this network evolves into a scale-invariant state, the probability that a vertex has k edges following a power law with an exponent $\gamma_{\text{model}} = 2.9 \pm 0.1$. The scaling exponent is independent of m, the only parameter in the model. Since the power law observed for real networks describes systems of rather different sizes at different stages of their development, one expects that a correct model should provide a distribution whose main features are independent of time. Indeed, as Fig. 2b demonstrates, $P(k)$ is independent of time (and, subsequently, independent of the system size $N = m_0 + t$), indicating that despite its continuous growth, the system organizes itself into a *scale-free stationary state*.

We have recently developed a continuum theory to calculate analytically the probability $P(k)$, allowing us to determine exactly the scaling exponent γ [22]. The theory predicted two major results:

(1) First, the connectivity k_i of a vertex i depends on time as

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{0.5}, \tag{5}$$

where the vertex i was added to the system at time t_i with connectivity $k_i(t_i) = m$ (see Fig. 2b).

(2) Second, the probability density for $P(k)$ follows

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2 t}{m_0 + t}\frac{1}{k^3}, \tag{6}$$

predicting

$$\gamma = 3, \tag{7}$$

independent of m. Furthermore, Eq. (6) also predicts that the coefficient A of the power-law distribution, $P(k) \sim Ak^{-\gamma}$, is proportional to the square of the average connectivity of the network, i.e., $A \sim m^2$. All these results have been verified using numerical simulations.

4. Relationship between the scale-free model and the www

It is far from us to suggest that the scale-free model introduced above describes faithfully the topology of the www. Naturally, the www has a much richer structure, that cannot be captured by such simple ingredients. For example, the links are not invariant in time, they constantly change, being either eliminated or rewired to other documents. Similarly, the www documents are not stable, they are often removed, and change address. Furthermore, the web pages are structured in domains, that by themselves have a rather complex hierarchical structure. In order to obtain a faithful model of the www we need to incorporate these ingredients. Nevertheless, we believe that our model captures in a minimalist way the main ingredients that are responsible for the development of the scale free state observed for the www.

Our model predicts $\gamma = 3$, while for the www we obtained the $\gamma_{\text{out}} = 2.45$ and $\gamma_{\text{in}} = 2.1$, significantly different values. We do not have a clear answer to this discrepancy yet. The solution to this problem is expected to come once we gain a better understanding of the possible universality classes characterizing random networks. For example, we have recently found that if the vertices are added not sequentially (as we do in the scale free model) but in a parallel fashion, the scaling exponent changes from $\gamma = 3$ to $\ln 3/\ln 2$. This indicates that our model is not unique, but there are other universality classes describing the development of random networks. Once the factors determining these universality classes will be understood, we can proceed in understanding the particular exponents describing the www.

A major assumption in the model was the use of a linear relationship between $\Pi(k_i)$ and k_i, given by (4). However, at this point there is nothing to guarantee that $\Pi(k)$ is linear, i.e., in general we could assume that $\Pi(k) \sim k^{\alpha}$, where $\alpha \neq 1$. The precise form

of $\Pi(k)$ could be determined numerically by comparing the topology of real networks at not too distant times. In the absence of such data, the linear relationship seems to be the most efficient way to go. In principle, if nonlinearities are present (i.e., $\alpha \neq 1$), that could affect the nature of the power-law scaling. This problem will be addressed in future work [23].

In the model we assumed that new links appear only when new vertices are added to the system. In many systems, including the www, links are added continuously. Our model can be easily extended to incorporate the addition of new edges. Naturally, if we add too many edges, the system becomes fully connected. However, in most systems the addition of new vertices (and the growth of the system) competes with the addition of new internal links. As long as the growth rate is large enough, we believe that the system will remain in the universality class of our model, and will continue to display scale-free features.

Naturally, we need to include the reconnection or rewiring of the existing links. Thus some links, that were added when a new vertex was added to the system, will break and reconnect with other vertices, probably still obeying preferential attachment[2] If reattachment dominates over growth (i.e., addition of new links by new vertices), the system will undergo a process similar to ripening: the very connected sites will acquire all links. This will destroy the power-law scaling in the system. However, as long as the growth process dominates the dynamics of the system, we expect that the scale-free state will prevail.

The above discussion indicates that there are a number of "end-states" or absorbing states for random networks, that include the scale-free state, when power-law scaling prevails at all times, the fully connected state, which will be the absorbing state of the ER model for large connection probability p, and the ripened state. The precise nature of the transition between these states is still an open question, and will be the subject of future studies [23].

Finally, the concept of universality classes has not been properly explored yet in the context of random network models. For this we have to define scaling exponents that can be measured for *all* random networks, whether they are generated by a model or a natural process. The clustering of these exponents for different systems might indicate that there are a few generic universality classes characterizing complex networks. Such studies have the potential to lead to a better understanding of the nature and growth of random networks in general.

Growth and preferential attachment are mechanisms common to a number of complex systems, including business networks [24], social networks (describing individuals or organizations), transportation networks [25], etc. Consequently, we expect that the scale-invariant state, observed in all systems for which detailed data has been available to us, is a generic property of many complex networks, its applicability reaching far beyond the www. A better description of these systems would help in understanding

[2] Note that a model with such ingredients has been proposed by L.A.N. Amaral and M. Barthélémy, Private communication.

other complex systems as well, for which so far less topological information is available, including such important examples as genetic or signaling networks in biological systems. Similar mechanisms could explain the origin of the social and economic disparities governing competitive systems, since the scale-free inhomogeneities are the inevitable consequence of self-organization due to the local decisions made by the individual vertices, based on information that is biased towards the more visible (richer) vertices, irrespective of the nature and the origin of this visibility.

Acknowledgements

This work was supported by the NSF Career Award DMR-9710998. We wish to thank L.A.N. Amaral and P. Schiffer for useful discussions.

References

[1] R. Gallagher, T. Appenzeller, Science 284 (1999) 79.
[2] R.F. Service, Science 284 (1999) 80.
[3] G. Weng, U.S. Bhalla, R. Iyengar, Science 284 (1999) 92.
[4] C. Koch, G. Laurent, Science 284 (1999) 96.
[5] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, 1994.
[6] Members of the Clever project, Sci. Am. 280 (June 1999) 54.
[7] R. Albert, H. Jeong, A.-L. Barabási, cond-mat/9907038.
[8] A.-L. Barabási, R. Albert, preprint.
[9] S. Lawrence, C.L. Giles, Science 280 (1998) 98.
[10] S. Lawrence, C.L. Giles, Nature 400 (1999) 107.
[11] K. Claffy, T.E. Monk, D. McRobb, Nature Web Matters, January 7, 1999 (http://helix.nature.com/webmatters/tomog.html).
[12] P. Erdős, A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17.
[13] B. Bollobás, Random Graphs, Academic Press, London, 1985.
[14] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[15] A. Bunde, S. Havlin, Fractals in Science, Springer, Berlin, 1994.
[16] S. Milgram, Psychol. Today 2 (1967) 60.
[17] M. Kochen (Ed.), The Small World, Ablex, Norwood, NJ, 1989.
[18] M. Barthélémy, L.A.N. Amaral, Phys. Rev. Lett. 82 (1999) 15.
[19] B.A. Huberman, L.A. Adamic, cond-mat/9901071.
[20] R. Kumar, P. Raghavan, S. Rajalopagan, A. Tomkins, Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.
[21] M. Faloutsos, P. Faloutsos, C. Faloutsos, SIGCOMM99, Harvard University, Cambridge, MA, 1999.
[22] A.-L. Barabási, R. Albert, H. Jeong, Physica A 272 (1999) 173.
[23] R. Albert, A.-L. Barabási, submitted.
[24] W.B. Arthur, Science 284 (1999) 107.
[25] J.R. Banavar, A. Maritan, A. Rinaldo, Nature 399 (1999) 130.