Emergence in the World Wide Web

P569 Term Paper

by Zach Etienne

May 4, 2006

Abstract

The World Wide Web (WWW) is a directionally linked network of $\sim 10^9$ nodes (web pages). Unlike a simple random network in which links are distributed with uniform probability between nodes, the probability $P(k)$ that there are k (web) links to or from a given node on the WWW behaves as a power law in k over many orders of magnitude, with exponent ≈ -2.1 for links to a node, and ≈ -2.45 for outbound links [1],[5]. This emergent property implies that the WWW has evolved into a scale-free (i.e., "self-similar") network. In this paper, we will review these observations and discuss two models that attempt to explain the power-law behavior of $P(k)$.

1 Introduction

The World Wide Web (WWW, or simply "the web") is an enormous, dynamic man-made network in which web pages are the nodes, and web links comprise the directed links between the nodes. Generally speaking, web pages are distinct, man-made entities created for a multitude of reasons, but always with the purpose of conveying information via links to other web pages, text, pictures, sound, and/or video. Thus from a reductionist standpoint, full comprehension of the WWW's complex topology requires insight into the comparably complex world of human motives. Such a level of reductionism will not take us far in understanding the large-scale structure of the WWW, but all is not lost; in work reviewed by this paper, statistical techniques are instead used to comprehend the complex structure of the web. Such approaches analyze fundamental statistical quantities in the WWW in the quest for emergent behavior (i.e., patterns indicating some deviation from a random network). Once a pattern has been found, models are constructed to reproduce it. Such models lend insight into the structure and evolution of the web, which in turn may be useful to the design and improvement of web search engines, for example.

Using automated web crawling software to explore the link structure of the nd.edu (U. of Notre Dame) website in 1999 ($\approx 3.3 \times 10^5$ web pages & $\approx 1.5 \times 10^6$ web links), Barabási et al. [1],[5] showed that the probability $P(k)$ for k inbound or outbound links (i.e., links *to* or *from* a web page, respectively) on the World Wide Web significantly deviates from that of a randomly linked network (in which $P(k)$ is sharply peaked, as shown in Figure 1) and so exhibits emergent behavior. Specifically, they found

$$P(k) = k^{-\gamma}, \text{ where} \tag{1}$$

$$\gamma = 2.1 \approx 0.1 \text{ for inbound links, and} \tag{2}$$

$$\gamma \approx 2.45 \text{ for outbound links.} \tag{3}$$

Such behavior is a telltale sign that some level of self-organization has occurred, and Barabási & Albert's simple model [1] reflects this. Due in part to the fact that it involves a network of *nondirected* links however, their model predicts the WWW scaling exponent $\gamma \approx 2.9$ instead of the observed values (Eqs. 2 & 3 above).

A breakthrough in modelling $P(k)$ for inbound links occurred when Bornholdt & Ebel [4] applied a model to the WWW originally designed by Herbert Simon [6] in 1955 to explain power-law distributions in other systems. Bornholdt & Ebel demonstrated that in this model, the scaling exponent of the web is not a fundamental quantity, but may be computed directly from another empirical quantity α, given by

$$\alpha = \frac{\Delta_{\text{web pages}}}{\Delta_{\text{web links}} + \Delta_{\text{web pages}}}, \tag{4}$$

where

$$\Delta_{\text{X}} = [(\#\text{ of X at time } t + \delta t) - (\#\text{ of X at time } t)]. \tag{5}$$

For accuracy, δt should be large enough so that $\Delta_{\text{X}} \gg 1$. Finally, Bornholdt & Ebel measure α directly from WWW data and show that with this α, Simon's model yields $\gamma \approx 2.1$ for inbound links, in agreement with the raw data.

In the proceeding sections, we will show some actual data collected from the web, including network graphs and plots of $P(k)$ for inbound links, to which Eqs. 1 & 2 are fit. This section will also contain limited description as to how these data were collected. Next, we will review in detail two models used to explain the power law exponent in the data: the so-called "scale-free" model of Barabási et al. [5], [1] and Simon's 1955 model. Finally, we analyze the shortcomings of these models, indicating the gaps that future models will need to fill.

2 Overview of the Data

Our goal in this section is threefold. First, we will compare the topology of a WWW-like network with that of a random network. Secondly, we will discuss the methods used for data collection on the web, and finally, we will show data measuring $P(k)$ on the web for inbound links.

2.1 Topology of the WWW

Figure 1 compares network topology and link distribution $P(k)$ for two network models: the "Scale-Free" network model of Barabási & Albert [1],[8], and the Erdös-Rényi random network model [7]. The link distribution for the random network model possesses a distinctive peaks about the average connectivity of a node, $\langle k \rangle$, whereas the "Scale-Free" model's $P(k)$ behaves as a power law, like the WWW. Both of the networks in this figure contain the same number of nodes and links. The red dots in each model indicate the 5 most connected nodes, and the green dots their nearest neighbors (1 link separation). In the random network, $\approx 27\%$ of the nodes are colored green, compared to $\approx 60\%$ in the "Scale-Free" model. Contrast these results with data collected from the WWW by Broder et al. [9], which indicate that, if the directionality of the links is neglected, 90% of all nodes connect to the single most connected node.

2.2 Data Collection Methods

Barabási et al. [1] and Albert et al. [2] used web crawling software that obtained data from the nd.edu web site in 1999 using the following two-step procedure, starting from the nd.edu homepage:

- Step 1: Download the web page, save its address, and extract the links. Go to Step 2.
- Step 2: For all links, if the link is within the nd.edu domain, save the link, follow that link and then return to Step 1 for each link.

This automated procedure accumulated a total of $\approx 3.26 \times 10^5$ web pages and $\approx 1.47 \times 10^6$ web links [2].

Broder et al. [9] on the other hand used vast WWW page/link databases collected by AltaVista in May 1999 and October 1999. These databases were constructed using web

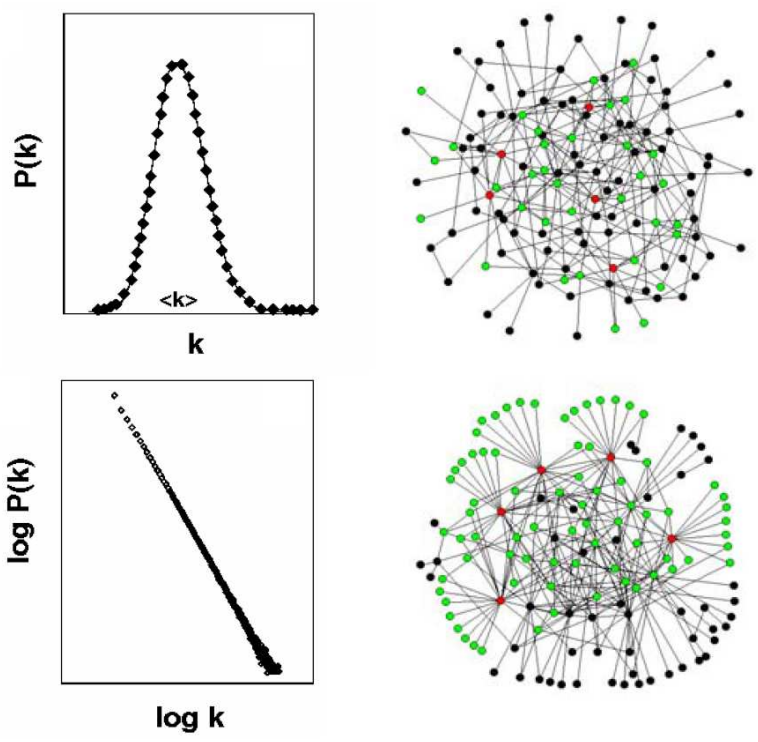

Figure 1: Comparison between network topology and link distribution $P(k)$ produced by the Erdös-Rényi random network model (top) [7] and the "Scale-Free", WWW-like, nondirectionally linked model of Barabási & Albert (bottom) [1],[8]. This figure was copied from Barabási et al. [8].

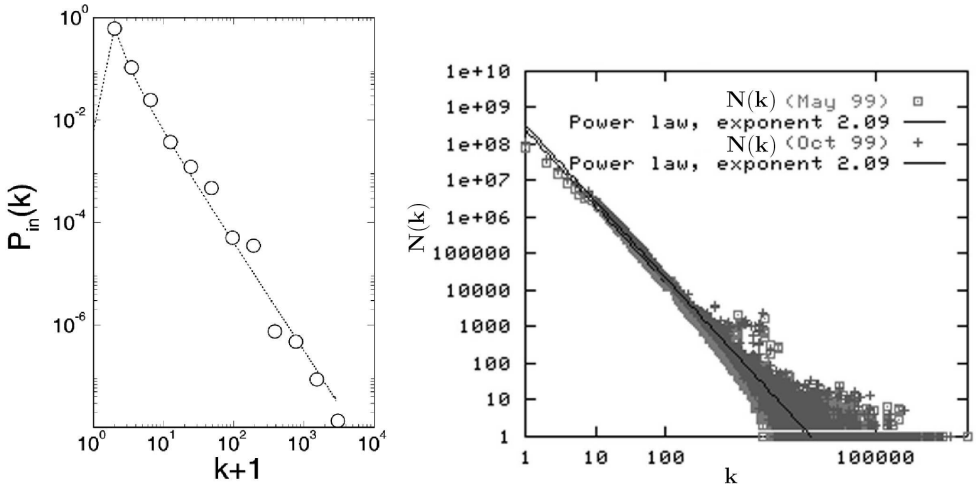

Figure 2: Left plot: the inbound link probability distribution $P(k)$, as measured by Barabási et al. [5], and Albert et al. [2]. Right plot: the total number of web pages as a function of k inbound links $N(k)$, as derived from AltaVista web crawling data by Broder et al. [9]. The best fit to the data in the left plot yields $\gamma = 2.1$, and, as indicated in the right plot, the inbound link scaling exponent is given by $\gamma = 2.09$ in both the October and May 1999 AltaVista data sets. Note that the left plot is copied from [5], and the right plot is copied from [9] (labeling modified for notational consistency in the right plot).

crawling technology similar to that by Barabási et al. [8], but with some important exceptions. First, it was not restricted to the nd.edu web site, and second, it contained filters designed to remove duplicate web sites, spam web sites, etc. Since it was not restricted to the nd.edu domain, the AltaVista data set was about three orders of magnitude larger than the Barabási et al. [1] and Albert et al. [2] data set. The AltaVista database included $\approx 2.03 \times 10^8$ web pages and $\approx 1.47 \times 10^9$ web links in May 1999, and $\approx 2.71 \times 10^8$ web pages and $\approx 2.13 \times 10^9$ web links in October 1999.

2.3 P(k) Measured from the WWW

Figure 2 shows a plot by Barabási et al. [5] of $P(k)$ and a plot of $N(k)$ by Broder et al. [9]. $P(k)$ is the probability that a given node on the WWW has k links to it, and $N(k)$ is the number of pages with k inbound links. Since $N(k)$ is equivalent to $P(k)$ without a normalization constant, we conclude from these plots that regardless of the database size, web crawling algorithm, or time in which the data was collected, the inbound link power law exponent on the WWW is given by $\gamma \approx 2.1$.

3 Model Analysis & Discussion

3.1 "Scale-Free" Model by Barabási & Albert

Introduced by Barabási & Albert in 1999 [1], this model is intended to explain the fact that $P(k)$ behaves as a power law. It is based on two observed deviations between the WWW and random network models. First, unlike the static network size in random network models (such as the Erdös-Rényi model [7]), the web is a dynamic network that has increased in size at a staggering rate since its inception in the early 1990s. Secondly and most importantly, random network models link nodes with fixed probability, whereas web pages appear to link preferentially to those pages that already have the most links.

The Scale-Free model comes in two flavors: discrete and continuous. The discrete flavor is used to determine properties of the model through numerical simulations, and the continuous flavor (the continuum limit of the discrete model) is used for an analytical analysis of the model and to verify the validity of the simulation algorithm.

3.1.1 Discrete Version

As initial data, the discrete algorithm begins with a small number of nodes m_o ($= 1, 3, 5$, or 7) with an unspecified structure of $m = m_o$ *undirected* links (allowing for a node to link to itself). Next, a node with m links is connected to the network one link at a time, where the probability that any one of these links connects to another node i with k_i links is given by

$$P(k_i) = \frac{k_i}{2N}, \tag{6}$$

where N is the total number of undirected links. The factor of 2 arises since each link has two ends, which makes

$$2N = \sum_i k_i \tag{7}$$

the appropriate normalization constant for $P(k_i)$.

Figure 3 shows results from simulations of this algorithm, by Barabási et al. [1] [8] with a variety of initial conditions. Notice that the scaling exponent γ (from Eq. 1) obtained from these simulations is independent of m (the only initial parameter). Thus γ is a universal feature of this model, but is found to be 2.9, not between the observed $\gamma = 2.1$ (for inbound links) or $\gamma \approx 2.45$ (for outbound links) on the web, as one might naïvely expect.

3.1.2 Continuum Version

The continuum version of the Scale-Free model by Barabási et al. [1],[8] assumes that the variables related to the number of links to/from the ith node – k_i, and $P(k_i)$ – are continuous. Further, it introduces a new continuous variable: t (for "time"), corresponding to the iteration step number in the discrete model. In constructing their continuum Scale-Free model, Barabási et al. start with the observation that the number of links connected to a

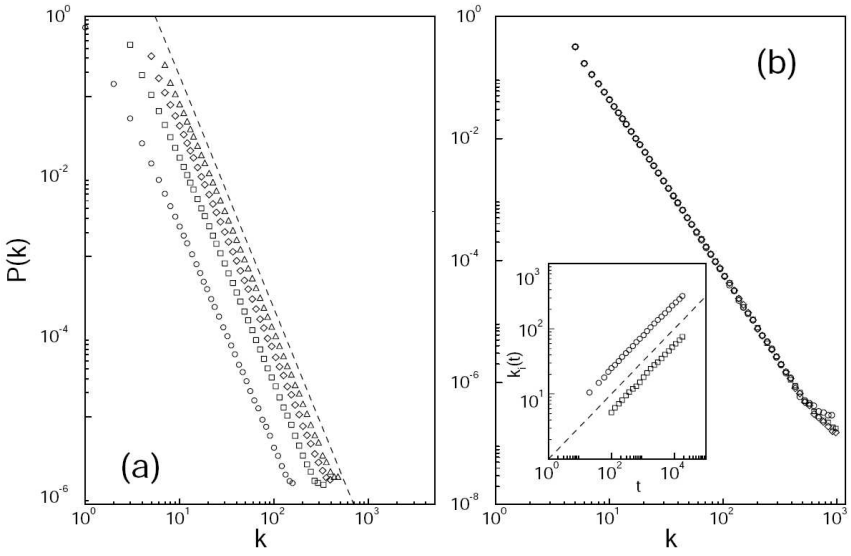

Figure 3: Results from discrete "Scale-Free" model simulations by Barabási et al. (figure reproduced from [8]). The left plot (Plot a) shows the link probability distribution $P(k)$ after $3\times10^5-m$ timesteps for a variety of initial conditions. In particular, the open circles, squares, diamonds, and triangles plot $P(k)$ with parameters $m = m_o = 1, 3, 5$, and 7 respectively. The dashed line in this plot represents the best fit to the slope, $\gamma = 2.9$, of these data sets. The right plot (Plot b) shows the time evolution of the $m = m_o = 5$ case after $1 \times 10^5 - m$ (circles), $1.5 \times 10^5 - m$ (squares), and $2 \times 10^5 - m$ (diamonds) iterations. The inset to Plot b shows the evolution of the number of links $k(t)$ for two nodes. The dashed line in this inset possesses slope 0.5, the best-fit slope to these data sets.

new node i, k_i, should increase in time t proportionally to $P(k_i)$. Further, since the discrete model adds m links at each timestep, the proportionality constant must be given by m. Thus Eq. 6 implies

$$\partial_t k_i = m\frac{k_i}{2N}. \tag{8}$$

Since $N = mt$, the above equation may be written

$$\partial_t k_i = \frac{k_i}{2t}. \tag{9}$$

This equation has general solution $k_i(t) = \alpha t^{1/2}$, where α is given by $k_i(t_i) = m$, since each node has m links at t_i, the time at which it is added to the network. Thus we have found the solution to Eq. 9, and it is

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{1/2}. \tag{10}$$

Notice that the above expression agrees with the results plotted in the inset to Plot b of Figure 3. Next, Barabási et al. use this result to find that $P(k)$, the probability that a given node possesses k links in their model, has the form

$$P(k) = 2m^2k^{-3} \propto k^{-3} \tag{11}$$

as $t \to \infty$. This power law exponent agrees with the simulated results of Figure 3, where γ was found to be ≈ 2.9.

3.2 Herbert Simon's Model

3.2.1 Original Formulation of Simon's Model

In 1955, Herbert Simon introduced a model [6] designed in part to describe the distribution of word frequencies in literature. Given the set of words in a given document $\mathcal{R}$, it was found that the probability of a word from $\mathcal{R}$ occurring i times behaves as a power law, as with $P(k)$ for the World Wide Web.

Next we reconstruct the basis of Simon's model ([6], pgs. 427-429), in our own words. Define $f(i,k)$ as the number of different words, occurring i times each, at the kth word in the document ($k \gg 1$). Also define $\mathcal{D}$ as the set of words up to and including the kth word in the document. Simon's model is then based on the following two assumptions:

1. **Assumption 1**: There is a fixed probability α that the next, $(k+1)$st word $\mathcal{W}$ is not in the set $\mathcal{D}$ ($\mathcal{W} \notin \mathcal{D}$).

2. **Assumption 2**: If $\mathcal{W} \in \mathcal{D}$ (probability $1-\alpha$), it must be chosen from the set with probability P_i proportional to the number of words in $\mathcal{D}$ occurring i times. That is, $P_i \propto if(i,k)$.

From Assumption 2, we see that P_i (the probability that, assuming $\mathcal{W} \in \mathcal{D}$, a word occurring i times will be picked) must be normalized to the total number of words in $\mathcal{D}$:

$$P_i = \frac{if(i,k)}{\sum_j jf(j,k)} = \frac{if(i,k)}{k}. \tag{12}$$

Now define a quantity $E_k(i) = \langle f(i,k+1)\rangle - f(i,k)$, where $\langle f(i,k+1)\rangle$ is the expectation value for the number of words occurring i times at position $k+1$ in the document. Therefore, $-1 \leq E_k(i) \leq 1$.

Assuming that $\mathcal{W} \in \mathcal{D}$ and $i > 1$, let us consider two cases:

- Case 1: $E_k(i) > 0 \rightarrow$ the $k+1$ word must have been a word appearing $i-1 > 0$ times in $\mathcal{D}$.
- Case 2: $E_k(i) < 0 \rightarrow$ the $k+1$ word must have been a word appearing i times in $\mathcal{D}$.

According to Assumption 2, Case 1 will occur with overall probability $(1-\alpha)P_{i-1}$ and Case 2 will occur with overall probability $(1-\alpha)P_i$. Since $E_k(i)$ involves the expectation value $\langle f(i,k+1)\rangle$, $E_k(i)$ for $i > 1$ must be equivalent to the following sum of probabilities:

$$\begin{aligned} E_k(i) &= [\text{Probability of Case 1}] - [\text{Probability of Case 2}] && (13)\\ &= (1-\alpha)(P_{i-1} - P_i) \qquad (i > 1). && (14) \end{aligned}$$

To complete our expression for $E_k(i)$, we must specify $E_k(1)$. By the same line of reasoning, Assumptions 1 and 2 yield

$$\begin{aligned} E_k(1) &= [\text{Probability that a new word is picked}] - P_1 \\ &= \alpha - P_1. \end{aligned} \tag{15}$$

With the goal of uncovering a steady state distribution of words appearing i times, $P(i)$, Simon solves the steady state version of the above equations for $E_k(i)$ (where $\langle f(i,k+1)\rangle \rightarrow f(i,k+1)$). He finds

$$P(i) \propto \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(i+\rho+1)} \tag{16}$$

where

$$\rho = \frac{1}{1-\alpha}. \tag{17}$$

Note that the above ratio of Γ's may be written approximately as $i^{-(1+\rho)}$, so

$$P(i) \sim i^{-(1+\rho)} = i^{-\gamma}. \tag{18}$$

Therefore the above set of "evolution equations" for $E_k(i)$ in the steady state limit has yielded a one-parameter power law distribution in word count. For more information on how well this model applies to actual documents, see Simon's original paper on the subject ([6]). Next, we will discuss Bornholdt & Ebel's application of Simon's model to networks, including the WWW.

3.2.2 Application of Simon's Model to Networks, by Bornholdt & Ebel [4]

The two assumptions in Simon's model are incompatible with the vocabulary of networks, so Bornholdt & Ebel rewrite Simon's assumptions in the context of a growing network, where $f(i,k)$ is the number of nodes with i inbound links at iteration k. Specifically, they rewrite Simon's assumptions as follows:

1. **Assumption 1'**: At iteration $k+1$ there is fixed probability α that a new node is added.

2. **Assumption 2'**: With probability $1-\alpha$, a single, directed link pointing to node i is added at iteration $k+1$ with probability P_i given by Eq. 12. Ignore the source of the directed link.

Notice from Assumption 2' that Bornholdt & Ebel's application of Simon's model considers only inbound links. As defined above for networks, all quantities behave in exactly the same way as in Simon's model, yielding the same "evolution equations" as Simon's model (Eqs. 14 & 15) and resulting $P(i)$ (Eq. 18).

Therefore, the probability distribution of inbound links $P(i)$ (identical to $P(k)$ as defined in Eq. 1) goes as a power law in the number of inbound links i, with exponent given by a single parameter α. As stated in Assumption 1', α is the probability that a node is added instead of a link. For the web, the total number of web pages and web links increases with time, so α may be determined by:

$$\alpha = \frac{N_{\text{pages}}(\delta t) - N_{\text{pages}}(0)}{[N_{\text{links}}(\delta t) - N_{\text{links}}(0)] + [N_{\text{pages}}(\delta t) - N_{\text{pages}}(0)]}, \tag{19}$$

where $N_X(t)$ is the number of X measured at time t, and for accuracy, δt should be large enough so that both the numerator and denominator of the above expression is $\gg 1$. Using the AltaVista data sets from May & October, 1999 [9] (also see Section 2.2), we find

$$\alpha \approx \frac{68 \times 10^6}{732 \times 10^6} \approx 0.1, \tag{20}$$

which yields an inbound link distribution power law exponent that agrees with actual measurements: $\gamma \approx 2.1$.

4 Conclusion

Although the generalization of Simon's model to networks accurately predicts the power law exponent of $P(k)$ for inbound links, this model has its faults. For example, it does not specify the origin of directed links, so its usefulness as a model for topological parameters other than $P(k)$ (e.g., average path length between pairs of nodes) is severely limited. Viewed in this way, Barabási & Albert's model is more robust, since it constructs a network in which each link has two well-defined ends, allowing the model to predict network quantities other than

$P(k)$. However, we know that the World Wide Web's structure is more complex than that assumed by either of the models we have presented in this paper. As other, more robust models are developed [3], the web will continue to evolve and grow, one web page at a time.

References

[1] A.-L. Barabási and R. Albert, Science 286, 509-512 (1999).

[2] R. Albert, H. Jeong, and A.-L. Barabási, Nature 401, 130 (1999).

[3] For a review of models contained in this paper and other current models of networks with power-law link distributions, see R. Albert and A.-L. Barabási, Reviews of Modern Physics 74, 47 (2002).

[4] Stefan Bornholdt and Holger Ebel, Phys. Rev. E 64, 035104 (2001).

[5] A.-L. Barabási, R. Albert, and H. Jeong, Physica A 281, 69-77 (2000).

[6] H. A. Simon, Biometrika 42, 425 (1955).

[7] P. Erdös and A. Rényi, Publ. Math. Debrecen 6, 290 (1959).

[8] A.-L. Barabási, Z. Deszö, E. Ravasz, S.-H. Yook, and Z. Oltvai, Scale-free and hierarchical structures in complex networks (to appear in Sitges Proceedings on Complex Networks, 2004).

[9] A. Broder et al., Comput. Netw., 33, 309 (2000).