PHYSICAL REVIEW E, VOLUME 64, 035104(R)

# World Wide Web scaling exponent from Simon's 1955 model

Stefan Bornholdt* and Holger Ebel
*Institut für Theoretische Physik, Universität Kiel, Leibnizstrasse 15, D-24098 Kiel, Germany*
(Received 1 September 2000; published 27 August 2001)

The statistical properties of the World Wide Web have attracted considerable attention recently since self-similar regimes were first observed in the scaling of its link structure. One characteristic quantity is the number of (in-)links $k$ that point to a particular web page. Its probability distribution $P(k)$ shows a pronounced power-law scaling $P(k) \sim k^{-\gamma}$ that is not readily explained by standard random graph theory. Here, we recall a simple and elegant model for scaling phenomena in general copy- and growth-processes as proposed by Simon in 1955. When combined with an experimental measurement of network growth in the World Wide Web, this classical model is able to model the in-link dynamics and predicts the scaling exponent $\gamma = 2.1$ in accordance with observation.

DOI: 10.1103/PhysRevE.64.035104 PACS number(s): 05.65.+b, 64.60.Fr, 89.20.Ff

Recently, a broad range of scaling phenomena has been observed in natural and artificial network structures, motivating new research on the dynamics of complex networks. One interesting example of a complex network is the World Wide Web (WWW), which forms a directed graph consisting of hypertext documents (nodes) and hyperlinks (edges). The properties of the WWW are of particular interest to new methods of search and retrieval of information. Search engines, for example, often face the problem of selecting or ranking the results of a keyword search from a vast number of hits. Many current search engines use simple text matching, though more advanced approaches use the specific link structure of the WWW [1]. One important quantity in this respect is the number $k$ of links that point to a particular web page (in-links). Its probability distribution $P(k)$ as observed at present in the internet exhibits a pronounced power-law scaling [2,3]

$$P(k) \propto k^{-\gamma}. \tag{1}$$

This remarkable feature is not readily explained by standard random graph theory [4] which for random networks predicts an exponential decay of the connectivity distribution, suppressing the occurrence of highly connected vertices. However, such highly connected nodes are frequently seen in measurements and form the power-law tail of $P(k)$.

Recent network models explain such scale-free link distributions as a result of network growth processes. These models are generally based on (i) constant network growth and (ii) the preferential addition of links to nodes that already have a large number of links. For example Huberman and Adamic [5,6] assume that the number of new links pointing to a node in one time step is a random fraction of the number of links already pointing to it. In the approach of Barabási *et al.* [2,7] the probability of connecting a new link to a node is directly proportional to the number of links that node already has. Both models are quite successful in explaining the emergence of power law statistics in the link structure of growing random networks. However, when applied to the World Wide Web, they fail to predict a scaling exponent that agrees with the observed value. They calculate an exponent of $\gamma = 3$ [2,7] which is too steep, or arrive at a range of $1 < \gamma < \infty$ depending on a free parameter [5,6]. Similarly, an extended model by Albert and Barabási [8] predicts a range of exponents between 2 and $\infty$, depending on parameters in the model. While these models only consider the distributions of in-links (and in this Rapid Communication we will focus only on such models), there are also extended versions that can accommodate more general scenarios, including modeling the generation of out-link distributions. However, such advanced models for the combined distributions of in-links and out-links arrive at the same range for the exponent of the in-link distribution [9–13]. Still, their exponent depends on at least one free parameter and, as the models above, they fail to explicitly predict the WWW in-link exponent without further tuning. One approach to an independent prediction of this exponent from experimental data is described below. When applying the above models to the WWW, a second problem emerges from the preferential linking assumption: The way it is defined in these models it correlates age and connectivity of nodes. However, this is not observed in the link structure of the WWW [5].

In the following, we address the problem of WWW growth by sketching a simple stochastic process for adding new nodes and in-links, based on a classical, however in this context, almost unnoticed model for scaling phenomena. It will allow us to calculate the exponent of the in-link distribution $\gamma$ for the WWW from experimental data of internet growth, and further solves the age correlation problem exhibited by the approaches mentioned above. The main idea has been formulated in the well-known model for scaling phenomena in copy- and growth-processes by Simon in 1955 [14]. Originally it was proposed to explain the scaling behavior observed in distributions of word frequencies in texts or city population figures (Zipf's law [15]). It models the dynamics of a system of elements with associated counters. New elements are constantly added while the counters are incremented at a rate depending on their current values. In Ref. [14], the model is formulated in terms of words in a text. In each iteration step $t$ the text grows by one word. The

*Email address: bornholdt@theo-physik.uni-kiel.de

 ©2001 The American Physical Society

$(t+1)$st word will be either a new one (with probability $\alpha$) or an old word (with probability $1-\alpha$) that has already appeared in the preceding text. Old words are copied from the existing text, i.e., if the $(t+1)$st word is an old word, the probability that it has appeared $k$ times is proportional to the total number of occurrences of all words that have appeared exactly $k$ times. This assumption is weaker than supposing that the probability of the $(t+1)$st word being a particular word which already occurred $k$ times is proportional to $k$. In order to model network growth consider a network with $n$ nodes with connectivities $k_i$, $i=1,\ldots,n$. The nodes are divided into connectivity classes $[k]$. A class $[k]$ is defined as the set containing all nodes with identical connectivity $k$. The cardinality of such a class $[k]$, i.e., the number of all nodes with connectivity $k$, is denoted by $f(k)$. For the growth process, the following steps are iterated:

(i) With probability $\alpha$ add a new node and attach a link to it from a node chosen in an arbitrary way.

(ii) Else add one link from an arbitrary node to a node $j$ of class $[k]$ chosen with probability

$$P_{[k]}=\frac{kf(k)}{\sum_i if(i)}. \tag{2}$$

Note that it is not specified how to choose the node $j$, which will receive the link, from class $[k]$. Also, it is not specified where the links originate from. Therefore, this model does not include modeling out-degree statistics as other models, e.g., Refs. [11–13]. On the other hand, it can easily accommodate extensions that explicitly model the statistics of out-links (e.g., by implementing an additional probability, as in Refs. [11,12]). This does not affect the main result for the in-link distribution obtained here. The only parameter of the model that in such a case couples to the additional process assigning out-links to nodes is the ratio of node versus link creation rates $\alpha$.

Following Ref. [14], the above process is described by the evolution equations

$$f(k,t+1)-f(k,t)=K(t)[(k-1)f(k-1,t)-kf(k,t)] \tag{3}$$

for $k=2,\ldots,t+1$ and

$$f(1,t+1)-f(1,t)=\alpha-K(t)f(1,t), \tag{4}$$

where $f(k,t)$ is the expectation value of the number of nodes in class $[k]$ at iteration step $t$ and where $K(t)$ is a proportionality factor. In order to evaluate $K(t)$ one uses the fact that $K(t)kf(k,t)$ is the probability that the $(t+1)$st link is received by a node of class $[k]$. With the probability for (ii)

$$\sum_{k=1}^{t} K(t)kf(k,t)=1-\alpha \tag{5}$$

and the number of links at $t$,

$$\sum_{k=1}^{t} kf(k,t)=t, \tag{6}$$

one obtains

$$K(t)=\frac{1-\alpha}{t}. \tag{7}$$

The stationary solution of this process given by Simon is

$$P(k)=AB(k,\rho+1) \tag{8}$$

with the constants $A$, $\rho:=1/(1-\alpha)$ and the beta function

$$B(k,\rho+1)=\int_0^1 \lambda^{k-1}(1-\lambda)^{\rho}d\lambda=\frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)}. \tag{9}$$

This distribution approximates a power law $P(k)\propto k^{-\gamma}$ with exponent

$$\gamma=1+\frac{1}{1-\alpha}. \tag{10}$$

For finite iteration times $t$ the power law will hold for $1\ll k\ll t$ influenced by transient effects depending on initial conditions. The only free parameter of the model $\alpha$ reflects the relative excess growth of number of nodes versus number of links. In general small values of $\alpha$, therefore, predict scaling exponents near $\gamma\approx 2$.

Given that, during network growth, the number of nodes increases, the real time interval $\Delta\tau$ associated with each iteration step $t$ should also change,

$$\Delta\tau=\frac{1}{cn} \tag{11}$$

with the total number of nodes $n=\Sigma_{i=1}^{t}f(i,t)$ and a constant $c$ for time units. The mean increase in the number of nodes in each step is $\Delta n=\alpha$ and $\Delta t=1$ for the number of links, respectively, leading to exponential growth of both quantities

$$\frac{dn}{d\tau}\approx\frac{\Delta n}{\Delta\tau}=c\alpha n, \tag{12}$$

$$\frac{dt}{d\tau}\approx\frac{\Delta t}{\Delta\tau}=c(\alpha t+n_0), \tag{13}$$

where $n_0$ denotes the initial number of nodes. Let us apply this process to modeling the evolution of the WWW, identifying nodes with web pages. Then (i) describes the creation of a new web page, whereas in (ii) a new link to an old page is inserted in some other page. It is natural to assume in (i) that at the same time a new page is created there will be a new reference to it from an existing page (e.g., from a directory page or from another page of the same project).

Data from two recent comprehensive AltaVista crawls [3] provide an estimate for $\alpha$ in the present internet. These two measurements counted 203 million pages and 1466 million links in May 1999, and 271 million pages and 2130 million

links in October 1999. In the model, in each iteration step one link is created. Hence, the probability for adding a new web page is estimated from the ratio of the observed increase in page counts and link counts to

$$\alpha \simeq \frac{68}{664} \simeq 0.10. \tag{14}$$

The subsequent prediction of Simon's model for the exponent of the link distribution is $\gamma = 2.1$ comparing well to current experimental results $\gamma = 2.1 \pm 0.1$ [2] and $\gamma = 2.09$ [3]. Thus, we obtain an independent prediction of the scaling exponent $\gamma$, based on the measurement of the independent quantity $\alpha$, the ratio of the creation rates of pages and links.

To compare with recently proposed models it may be interesting to note that the model by Barabási and Albert [2] can be mapped to the subclass $\alpha = 1/2$ of Simon's model (leading to a scaling exponent $\gamma = 3$). Of the extended models noted above the closest connection is to the model of Dorogovtsev *et al.* [10], who consider a free parameter ''initial attractiveness of a node'' that can be related to $\alpha$. Another remarkable point is that these models use a more specifically defined preferential linking than (ii). They add one link from an arbitrary node to a node $j$ with a probability proportional to the connectivity of the receiving node

$$P_j = \frac{k_j}{\sum_i k_i}. \tag{15}$$

Note that Eq. (15) implies Eq. (2), whereas the reverse is not true. Otherwise these models are based on the same two assumptions of growth and preferential linking as used here. From this viewpoint, it is interesting to reconsider a recent discussion of these models. Barabási and Albert mention that this type of preferential linking (15) implies a ''rich-get-richer'' behavior of individual nodes [2]. In other words, a node already receiving many links will grow much faster than a node with smaller connectivity. Adamic and Huberman point out that this ''rich-get-richer'' phenomenon correlates the age and connectivity of nodes [5]. This, however, is disproven by the data they present. One possible solution they suggest is to add individual growth rates to each node which could solve this correlation problem. In response, Barabási *et al.* [16] show how to introduce such intrinsic growth rates $\eta_i$ to each node $i$, thereby modifying preferential linking (15) to

$$P_j = \frac{\eta_j k_j}{\sum_i \eta_i k_i}. \tag{16}$$

While this solves the correlation problem, the price to pay is a large number of free parameters in the extended model. As shown above, a simple solution to this problem is already implicit in Simon's model: Linking is guided by Eq. (2) instead of Eqs. (15) or (16), considering not single nodes but classes of nodes with identical connectivities. This allows for different growth rates among class members, leaving just one free parameter.

With respect to the WWW, Simon's more general definition of ''preferential linking'' corresponds to separating the linking process into two natural parts: First, the process of getting to know a page, and second, the decision whether to link this page or not. In recently discussed models both steps occur at once, by directly linking a page with a probability proportional to its popularity. In reality, however, only the process of getting to know a page necessarily depends on popularity. Whether this page ends up being linked, in practice depends on many other variables as, e.g., contents, age, etc. This more general picture restricts the influence of popularity to the only necessary part in the dynamics of linking. This is what Simon formulates, resulting in a quite general scenario of linking dynamics. In particular, the specific criteria that we apply to decide whether to link a page, do not at all influence the exponent of the connectivity distribution. The recent models' preferential linking discussed above, is contained in this picture as a special case. To be specific, the first step in Simon's linking process is described by Eq. (2): the probability of getting to know a page depends on the number of links pointing to it. So, one encounters a page of class $[k]$ with a probability proportional to the number of links received by the whole class. Then, one decides whether to link or not. This means, one chooses a page of class $[k]$ with a probability influenced by ''preferences'' of the linking node, i.e., a probability $P(i \to j|[k])$ that node $i$ directs a link to node $j \in [k]$ given that $i$ will connect to a node of class $[k]$. A simple example may illustrate the generality of this approach. Consider the following modification of (ii): First, node $i$ will connect to a member of class $[k]$ with probability (2). Among the class members, let us then choose to link according to a completely different criterion, e.g., link to the youngest node (as this may be a trend setting page for example), i.e.,

$$P_{ij} = P_{[k_j]} P(i \to j|[k_j]) = \begin{cases} \dfrac{k_j f(k_j)}{\sum_\nu \nu f(\nu)} & \text{if } j \text{ youngest member of } [k_j], \\ 0 & \text{else.} \end{cases} \tag{17}$$

As the second step, the specific choice of a node from class $[k]$, is purposedly left unspecified in Simon's model, Eq. (17) leads to the same evolution equations (3, 4) and, therefore, to the same power law for the degree distribution $P(k)$. In other words, a node's ''bounded knowledge'' of the whole network, in terms of the probability of knowing a page (2) is sufficient for the emergence of a power-law connectivity distribution, independent of specific preferences in the actual linking process within the set of known pages.

In summary, Simon's classical scaling model has been reformulated to model an interesting aspect of WWW growth, predicting a characteristic power law for the distri-

bution of in-links. The one free parameter $\alpha$ of this model has been determined from recent measurements of web growth and the scaling exponent calculated to $\gamma = 2.1$. This estimate agrees well with direct experimental measurements and is robust in the case of small changes of $\alpha$. We find that recent models of WWW growth are closely related to Simon's original model. In contrast to these, Simon defines a more general process of linking that does not correlate age and connectivity, which presents a problem in other recent models as this effect is not observed in real data.

[1] D. Butler, Nature (London) **405**, 112 (2000).
[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[3] A. Broder *et al.*, Comput. Netw. **33**, 309 (2000).
[4] P. Erdös and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[5] L.A. Adamic and B.A. Huberman, Science **287**, 2115a (2000).
[6] B.A. Huberman and L.A. Adamic, Nature (London) **401**, 131 (1999).
[7] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[8] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).
[9] P.L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[10] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[11] B. Tadić, Physica A **293**, 273 (2001).
[12] M.E.J. Newman, S.H. Strogatzr, and D.J. Watts, Phys. Rev. E **64**, 026118 (2001).
[13] P.L. Krapivsky, G.J. Rodgers, and S. Redner, Phys. Rev. Lett. **86**, 5401 (2001).
[14] H.A. Simon, Biometrika **42**, 425 (1955).
[15] G.K. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).
[16] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, Science **287**, 2115a (2000).