Evolutionary Dynamics of the World Wide Web

Bernardo A. Huberman and Lada A. Adamic
Xerox Palo Alto Research Center
Palo Alto, CA 94304

February 25, 1999

Abstract

We present a theory for the growth dynamics of the World Wide Web that takes into account the wide range of stochastic growth rates in the number of pages per site, as well as the fact that new sites are created at different times. This leads to the prediction of a universal power law in the distribution of the number of pages per site which we confirm experimentally by analyzing data from large crawls made by the search engines Alexa and Infoseek. The existence of this power law not only implies the lack of any length scale for the Web, but also allows one to determine the expected number of sites of any given size without having to exhaustively crawl the Web.

The World Wide Web (Web) has become in a very short period one of the most useful sources of information for a large part of the world's population. Its exponential growth, as shown in Figure 1, from a few sites in 1994 to millions today, has transformed it into an ecology of knowledge in which highly diverse information is linked in extremely complex and arbitrary fashion (1). Moreover, several estimates of the total number of pages (2) indicate that due to the rapid growth of the Web, most search engines are only finding a fraction of all the available sites (2).

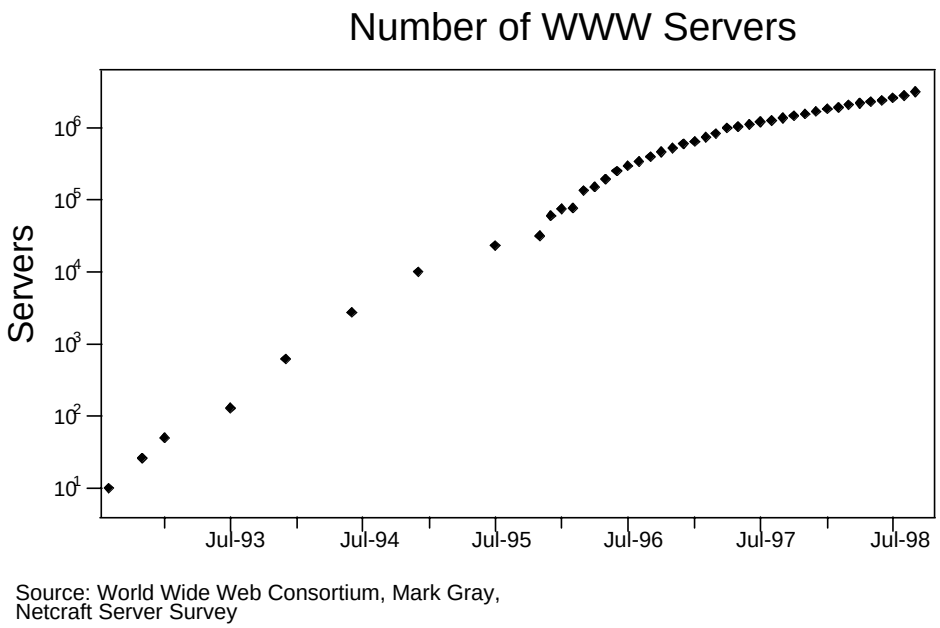

Figure1: The growth of the World Wide Web

While this growth at first sight may appear to be totally haphazard, it is characterized by a number of regular features which we uncover in this work. In particular, we show that a theory of stochastic growth dynamics for the Web leads to the prediction of a universal power law in the distribution of the number of pages per site. The prediction was confirmed experimentally by analyzing data from two large crawls by the search engines Alexa and Infoseek. The existence of this power law not only implies the lack of any length scale for the Web, but also allows one to determine the expected number of sites of any large size without having to exhaustively crawl the Web.

In order to develop an evolutionary theory of the growth of the Web, we first consider the number of pages belonging to a given site as a function of time. Since pages within sites are typically organized in hierarchical, tree-like, fashion, the number of pages added at any given time to a site will be proportional to

those already existing there. Thus, if $n_s(t)$ is the number of pages belonging to a site s at time t, the number at the next interval of time, $n_s(t+1)$, is determined by

$$n_s(t+1) = n_s(t) + g(t+1)n_s(t) \tag{1}$$

where $g(t)$is the growth rate. Given the unpredictable character of site growth, we assume that $g(t)$ fluctuates in an uncorrelated fashion from one time interval to the other about a positive mean value g_0. In other words

$$g(t) = g_0 + \xi(t) \tag{2}$$

with the fluctuations in growth, $\xi(t)$,behaving in such a way that $< \xi(t) >= 0$ and $< \xi(t)\xi(t+1) >= 2\sigma\delta_{t,t+1}$, i.e. they are delta correlated and with zero mean. This assumption was confirmed by a study of the growth of the Xerox Corp. Web site, whose fluctuations in growth are plotted in Figure 2. Pearson's correlation test accepts at the 0.05 level (with p-value 0.71) the hypothesis that the day to day fluctuations in the growth rate are uncorrelated.

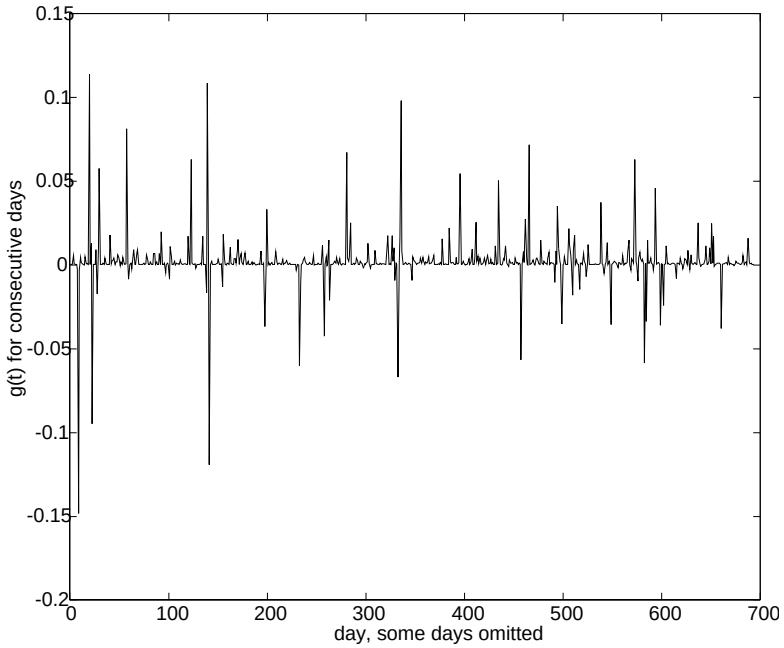

Figure 2: Rate of growth of the Web site for the Xerox Corporation.

In order to obtain the distribution of pages per site, we sum Eq. (1) to get

$$\sum_{t=0}^{T} \frac{n_s(t+1) - n_s(t)}{n_s(t)} = \sum_{t=0}^{T} g(t) \tag{3}$$

Changing the sum to an integral (which assumes that the differences in pages between two time steps is small) we obtain

$$\int_0^T \frac{dn_s}{n_s} = \ln \frac{n(T)}{n_s(0)} = \sum_{t=0}^{T} g(t) \tag{4}$$

Notice that the right hand side of Eq. (4) is a sum over discrete time steps, at each of which we assume the values of g to be normally distributed with mean g_0 and variance σ^2. This corresponds to a Brownian motion process with stationary and independent increments. By invoking the Central Limit Theorem we can assert that for every time step t, the logarithm of n_s is normally distributed with mean $g_0 t$ and variance $\sigma^2 t$(3,4). This means that the distribution of the number of pages for sites created at the same time and with the same average growth rate is log-normal (5), i.e, its density is given by

$$P(n_s) = \frac{1}{n_s \sqrt{t}\sqrt{2\pi\sigma^2}} \exp -[\frac{(\ln n_s - g_0 t)^2}{2\sigma^2 t}] \tag{5}$$

where the time dependent drift $g_0 t$ is the mean of $\ln n_s$, reflecting the fact that as time goes on there are more pages added on average than deleted. The variance of this distribution is related to the median $m = \exp(g_0 t)$ by $Var(n_s) = m^2 \exp\left(t\sigma^2\right) \left(\exp\left(t\sigma^2\right) - 1\right)$.

Some insight into the dynamics of this growth can be obtained by noticing that the stochastic differential equation associated with Eq. (1), which is given by

$$\frac{dn_s}{dt} = [g_0 + \xi(t)] n_s \tag{6}$$

can be solved exactly(6). The solution is the stochastic growth process

$$n_s(t) = n_s(0) \exp\left(g_0 t + w_t\right) \tag{7}$$

where w_t is a Wiener process such that $< w_t >= 0$ and $< w_t^2 >= \exp \sigma^2 t$. Equation (7) shows that typical fluctuations in the growth of the number of pages away from their mean rate g_0 relax exponentially to zero. On the other hand, the n^{th}moments of n_s , which are related to the probability of very unlikely events, grow in time as $< n_s(t)^n >= [n_s(0)]^n \exp[n(n - \sigma g_0 t]$, indicating that the evolutionary dynamics of the web is dominated by occasional bursts in which large number of pages suddenly appear at a given site. These bursts are responsible for the long tail of the probability distribution and make average behavior to depart from typical realizations(7).

In order to consider the evolutionary dynamics of the whole Web, it is important to notice that the distribution of the number of pages depends on the time that has elapsed since the site was created. Since the number of sites in the Web has doubled on average every six months, newer sites are more numerous than older one, and therefore the distribution of pages per site, for all sites

of a given growth rate regardless of age, is a mixture of lognormals given by Equation (5), whose age parameter t is weighted exponentially. Thus, in order to obtain the true distribution of pages per site that grow at the same growth rate, we need to compute the mixture given by

$$P(n_s) = \int \lambda \exp(\lambda t) \frac{1}{n_s \sqrt{2\pi t \sigma^2}} \exp -[\frac{(\ln n_s - g_0 t)^2}{2t\sigma^2}]dt \tag{8}$$

which can be calculated analytically to give

$$P(n_s) = C n_s^{-\beta} \tag{9}$$

where the constant C is given by $C = \lambda/\sigma(\sqrt{(g_0/\sigma)^2 + 2\lambda})$ and the exponent β is in the range $[1, \infty]$ and determined by $\beta = 1 - \frac{g_0}{\sigma^2} + \frac{\sqrt{g_0^2 + 2\lambda\sigma^2}}{\sigma^2}$.

Lastly we need to take into account different growth rates for sites of the Web, since the distribution given by Eq. (9) applies only to sites that have the same growth rate $g = g(g_0, \sigma)$. Since each growth rate occurs with a particular probability $P(g)$, and gives rise to a power law distribution in the number of pages per site with a specific exponent, the probability that a given site with an unknown growth rate has n_s pages is given by the sum, over all growth rates g, of the probability that the site has so many pages given g, multiplied by the probability that a site's growth rate is g, i.e.

$$P(n_s) = \sum_i P(n_s|g_i) P(g_i) \tag{10}$$

Since we have already shown that each particular growth rate gives rise to a power law distribution with a specific value of the exponent $\beta(g)$, this sum is of the form

$$P(n_s) = \frac{c_1}{n_s^{\beta_1}} + \frac{c_2}{n_s^{\beta_2}} + ... + \frac{c_n}{n_s^{\beta_n}} \tag{11}$$

which, for large values of n_s behaves like a power law with an exponent given by the smallest power present in the series.

We thus obtain the very general result that the evolutionary dynamics of the World Wide Web gives rise to an asymptotic self similar structure in which there is no natural scale, with the number of pages per site distributed according to a power law. This implies that on a log-log scale, the number of pages per site, for large n, should fall on a straight line.

In order to test this theory, we studied data generated by crawls of the World Wide Web made by two search engines, Alexa (8) and Infoseek(9), which covered 259,794 and 525,882 sites respectively. The plots in Figure 3 show the probability of drawing at random from the sites in the crawls a site with a given number of pages. Both data sets display a power law over several orders of magnitude, with a drop-off at approximately 10^5 pages, which is due to the fact that crawlers don't systematically collect more pages per site than this

bound because of server limitations. The power law, as well as the drop-off are illustrated in Figure 3.

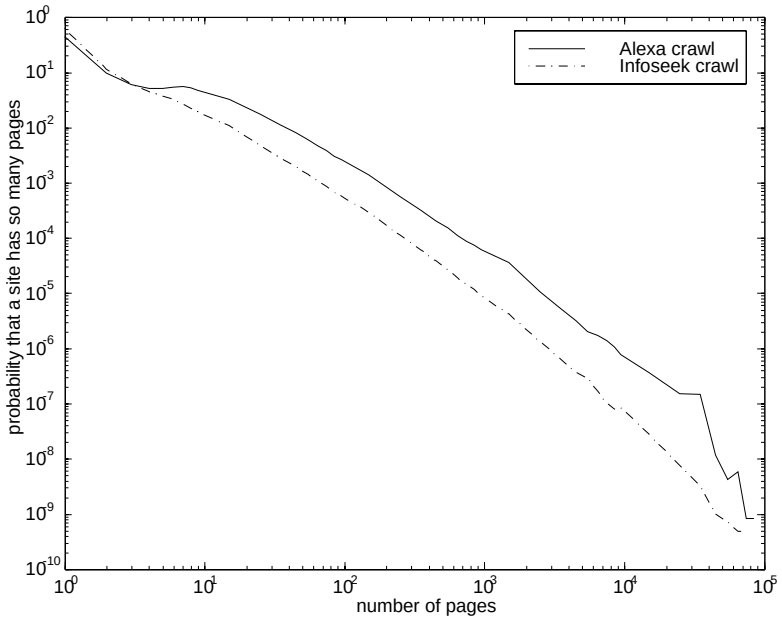

Figure 3: Distribution of pages per site.

A linear regression on the variables log(number of sites) and log(number of pages) yielded [1.647, 1.853] as the 95% confidence interval for the value ß in the Alexa crawl of 250,000 sites of the World Wide Web. For the Infoseek crawl, the 95% confidence interval for ß is [1.775, 1.909]. These estimates for the value of β are consistent across the two data sets and with the model, which predicts a linear dependence between the logarithm of the variables to be linear with slope $\beta > 1$.

The existence of this universal power law has practical consequences as well, since one can estimate the expected number of sites of any arbitrary size, even if a site of that size has not yet been observed. This can be achieved by extrapolating the power law given by Eq. 9, to any large n_s, e.g. $P(n_{s2}) = P(n_{s1})(n_{s1}/n_{s2})^{\beta}$. The expected number of sites of size n_{s2} in a crawl of N sites would be $NP(n_{s2})$. As an example, from the Alexa data we can infer that if one were to collect data on 250,000 sites the probability of finding a site with a million pages would be 10^{-4}. Notice that this information is not readily available from the crawl alone, since it stops at 10^5 pages per site.

Several points are worth making. First, since small values of n_s lie outside the scaling regime, our theory does not explain the data on sites with few pages. Secondly, as a consequence of the universality of our prediction, as more sites

will be created, the same power law behavior will be seen. This will once again allow for the determination of largest sites from data that will be limited in scope due to server limitations. Finally, since the process of ranking random variables stemming from any broad distribution always produces a narrow and monotically decreasing power law of the type originally discussed by Zipf (10), we expect that such ranking will lead to a Zipf-like law(11).

In summary, we presented a stochastic theory of the growth dynamics of the Web that takes into account the wide range of stochastic growth rates in the number of pages per site, as well as the fact that new sites are created at different times in the unfolding story of the Web. This leads to the prediction of a universal power law in the distribution of the number of pages per site, which we confirm experimentally by analyzing data from two large crawls by the search engines Alexa and Infoseek. The existence of this power law not only implies the lack of any length scale for the Web, but also allows to estimate the number of sites of any given size without having to exhaustively crawl the Web. This is yet a another example of the strong regularities(12) that are revealed in studies of the Web, and which become apparent because of its sheer size and reach.

Acknowledgement 1 *We thank Jim Pitkow and Eytan Adar for providing data for our analysis, and Rajan Lukose for many useful discussions. This work was partially supported by NSF grant IRI-961511.*

References

[1] *Scientific American*, Special issue on the Internet, 276 (1997).

[2] Lawrence, S. and Giles, L., *Science* **280**, 91-94 (1998).

[3] Ross, S. M. Stochastic Processes, John Wiley (1996).

[4] Athreya, K. B. and Ney, P. E. Branching Processes, (Springer-Verlag, 1972).

[5] Crow, E. L. and Shimizu, K. Lognormal Distributions: Theory and Applications, Marcel Dekker, (1988).

[6] Stratonovich, R. L. Topics in the Theory of Random Noise (Gordon and Breach, Newark, NJ, 1967).

[7] Lewontin, R. C. and Cohen, D. *Proc. Natl. Acad. Sci. U.S.A.* **62**, 1056 (1969).

[8] http://www.alexa.com

[9] http://www.infoseek.com

[10] Zipf, G. K. Human behavior and the principle of least effort (Addison-Wesley, Cambridge, MA, 1949.

[11] Gunther, R. Levitin, L. Schapiro, B. and Wagner, P. *Intern. J. of Theor. Phys*, **35**, 395-417 (1996).

[12] Huberman, B. A., Pirolli, P. Pitkow, J and Lukose, R. M. *Science* **280**, 95-97 (1998).