The Nature of Markets in the World Wide Web

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

May 6, 1999

Abstract

We studied the statistics in the number of visitors to sites of the World Wide Web by examining usage logs covering one hundred and twenty thousand sites. We found out that both in the case of all sites and sites in specific categories, the distribution of visitors per site follows a universal power law, characteristic of winner-take-all markets. We developed a dynamical theory of site popularity which takes into account the stochastic nature of user decisions to visit given sites as well as the fact that newer sites are appearing at an ever increasing rate. The model accounts for the observed power law behavior and naturally provides the amplification factor responsible for the increased performance of the top performers.

The exponential growth of the World Wide Web (Web), depicted in Figure 1, has been accompanied by a global increase in the number of users of the Internet. Whereas in 1996 there were 61 million users, at the close of 1998 over 147 million people had Internet access worldwide. By the year 2000, the number of internet users is expected to double again to 320 million[1]. In addition to this remarkable growth, the Web has popularized electronic commerce, and as result an increasing segment of the world's population conducts commercial transactions in novel ways. Chiefly among them are transactions involving intangible goods such as entertainment, travel, information and banking services, as opposed to transactions in the traditional economy, which are made up of massive products such as computers, books and cars. Moreover, this massless electronic commerce is conducted on a global scale, in which many providers distributed over several countries can be accessed at the click of a button. Thus, consumers can profit from increased information about the products they purchase, lower transaction costs and prices, and a wider set of choices than those available in the traditional economy.

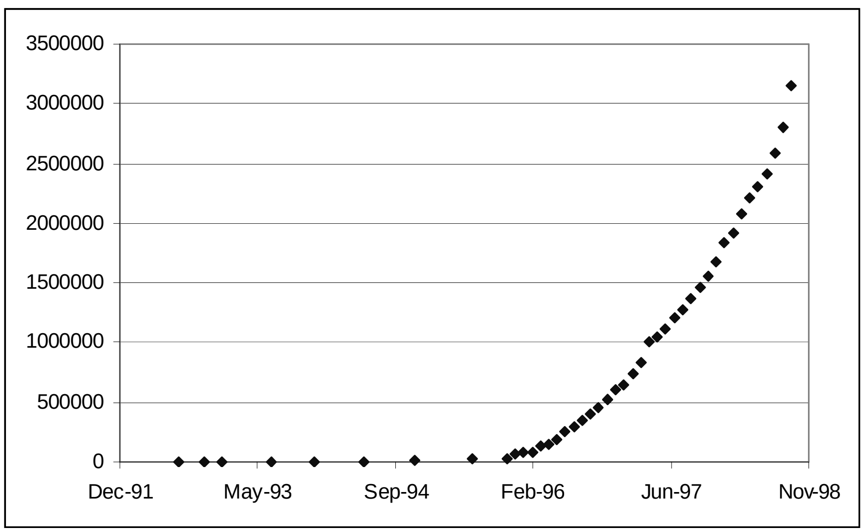

Fig. 1 Growth of the number of Web servers over time

On the supply side, the Web offers providers access to global markets without having to incur large entry costs or having to keep sizable inventories. This great opportunity is tempered by the increased competition that can result from newcomers continuing to offer novel combinations of products and ways of delivering them. The recent example of the free distribution of recorded music over the Internet by non-traditional recording firms provides an illustration of the speed with which an industry needs to adapt to novel mechanisms and technologies in order to survive[2].

Given these conditions of improved global access and diverse offerings, it is of interest to find out about the nature of the markets that are mediated by the Web. While in traditional economies transaction costs and geography conspire to segment the markets so that they cater to a local consumer base, one expects that the global reach of the Web could lead to different market share characteristics on the provider side. The distribution of the market share on the Web is interesting both to the economic theorist contemplating market efficiencies and the e-commerce provider contemplating the number of customers the business will attract.

In order to address the issue of market share on the Web, we studied the distribution of users among Web sites by examining usage logs from America Online (AOL)[3] covering 120,000 sites. Users in this context act as proxies for economic activity. We found out that both in the case of all sites and sites in specific categories, the distribution of visitors per site follows a universal power law similar to that found by Pareto in income distributions. This implies that a small number of sites command the traffic of a large segment of the Web population, a signature of winner-take-all markets[4]. In order to explain this data, we present a dynamical theory of site popularity which takes into account the stochastic nature of user decisions to visit given sites as well as the fact that newer sites are appearing at an ever increasing rate. The model accounts for the observed power law behavior and naturally provides the amplification factor responsible for the increased performance of the top performers. These results show that a newly established site will, with high probability, join the ranks of sites which attract a handful of visitors a day, while with an extremely low probability it will capture a significant number of users. Such a disproportionate distribution of user volume among sites is characteristic of winner-take-all markets, wherein the top few contenders capture a significant part of the market share.

We obtained the distribution of users among sites from access logs of a subset of AOL users active December 1, 1997. The subset comprises 60,000 users accessing 120,000 sites. A request for the document "http://www.a.b.com/c/d.html" was interpreted as a request for the site b.com. This definition of a site was chosen in absence of information about individual site content. "a1.b.com" and "a2.b.com" have a fair chance of containing related information and hence were considered a single site. Sites "b1.com" and "b2.com" might comprise a single conceptual site, but were counted separately. Further, visits to multiple documents within the same site or multiple visits to the same document were counted only once. Thus we relate the popularity of a site to the number of unique visitors it received in one day. Table 1. shows the percentage of volume (measured in unique visitors) accounted for by the top sites.

	% volume by user		
% sites	all sites	adult sites	educational sites
0.1	32.36	1.4	2.81
1	55.63	15.83	23.76
5	74.81	41.75	59.50
10	82.26	59.29	74.48
50	94.92	90.76	96.88

Table. 1 Distribution of user volume among sites

The top 119 (top 0.1%) sites, excluding AOL itself, capture a whopping 32.36% of user volume. The top 1% of sites capture more than half of the total volume. The top site was yahoo.com. Figure 2 shows a histogram of the number of sites with each particular number of unique visitors. Figure 3 is a distribution of the number of users per site, obtained by binning the histogram in Figure 2 and fitting it with a straight line. A straight line fit on a log-log scale implies a power law distribution in the number of users per site.

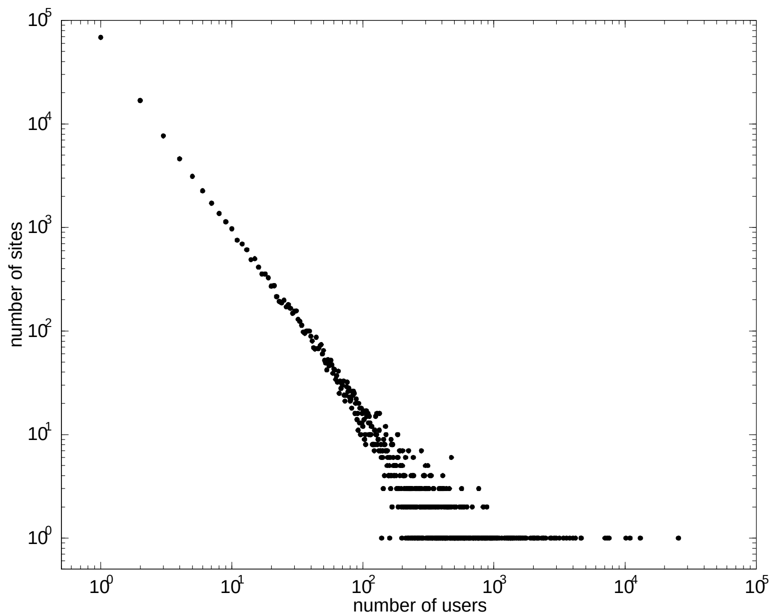

Fig 2. Occurence of sites by popularity

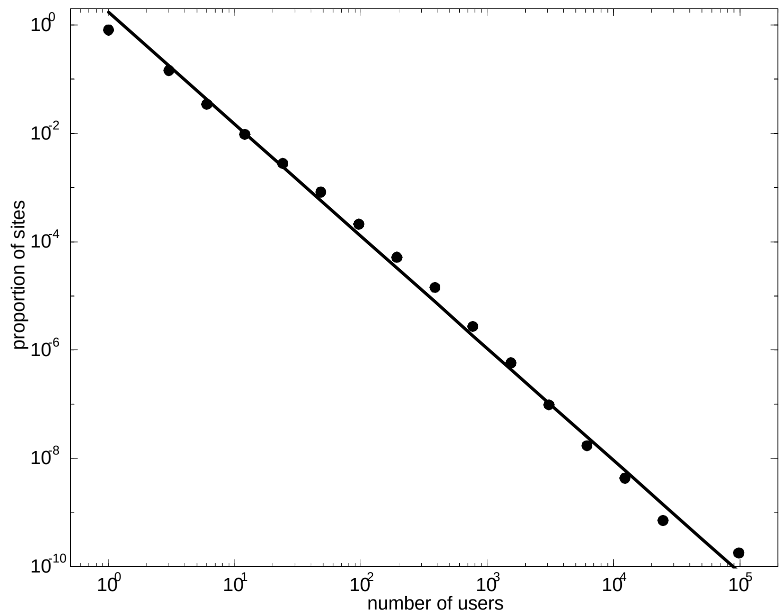

Fig. 3 The proportion of sites by popularity was obtained by binning the data shown in figure 2. The fit shown is a powerlaw distribution.

In order to control for difference in kind and function of sites, we looked for unequal distribution of volume within categories of sites which we expected to have similar content. The two categories we chose were adult and .edu domain sites. Adult sites were assumed to offer a selection of images and optionally video and chat. Educational domain sites were assumed to contain information about academics and research as well as personal homepages of students, staff, and faculty, which could cover any range of human interest. Again, the distribution of visits among sites was unequal. 6,615 adult sites were sampled by keywords in their name. The top site captured 1.4% of the volume to adult sites, while the top 10% accounted for 60% of the volume. Similarly, of the .edu sites, the top site, umich.edu, held 2.81% of the volume, while the top 5% accounted for over 60 percent of the visitor traffic. Figure 4 shows a the binned and fitted distributions for both adult and .edu sites.

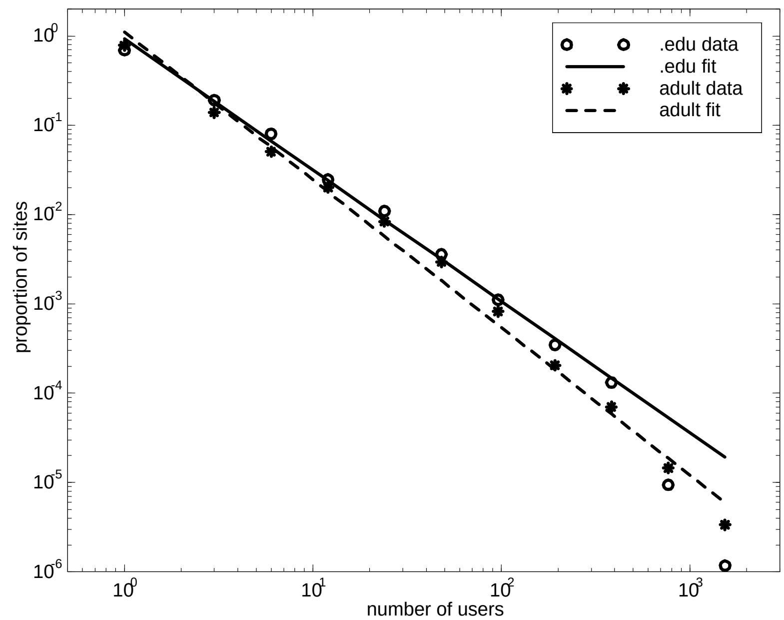

Fig. 4 Distributions of users among .edu and adult sites.

In the remainder of the paper we develop a theory of stochastic growth dynamics for Web usage which predicts of the power law behavior observed in the above data analysis. In developing an evolutionary theory of the growth of popularity of Web sites, we first consider the number of users frequenting a given site as a function of time. We claim that the difference in the number of visitors at a site in two successive time periods is proportional to the total number of visitors to that site. A large site with thousands of users might fluctuate by hundreds of visits on any given day, while a small site with dozens of users might experience a few visits more or less. Thus, if $n_s(t)$ is the number of visitors 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 user population 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 usage of the Xerox

Corp. Web site, whose fluctuations in growth are plotted in Figure 5. The weekly fluctuations in the growth rate were found to be uncorrelated at the 95% confidence level.

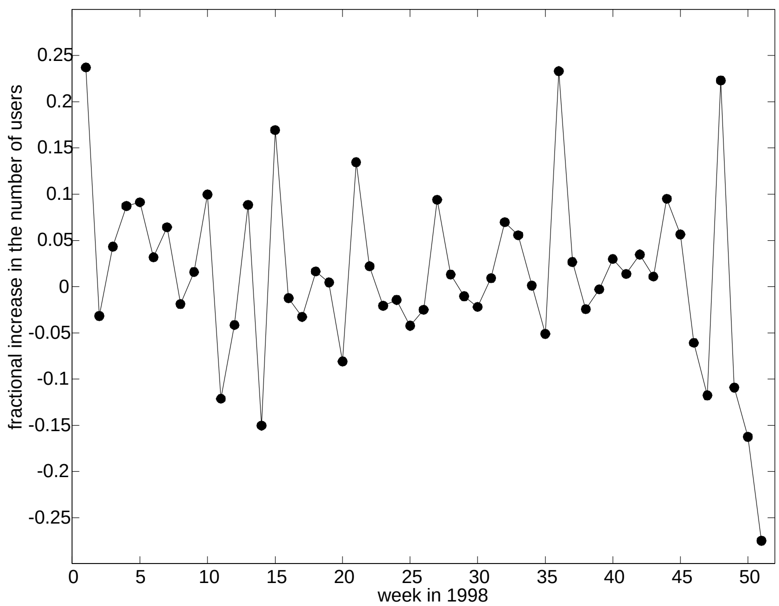

Fig. 5 Fractional fluctuations in the number of users at the Xerox web site.

We obtain the distribution of the number of users per site by considering that the distribution of users at a site to be completely determined by its age, α, its mean growth rate g_0, and the variance of its usage fluctuations σ^2. In our analysis we make use of mixtures of distributions. If the distribution of x, $p(x, \gamma)$, depends on the parameter γ, which in turn is distributed according to its own distribution $r(\gamma)$, then the distribution of x, is given by $p(x) = \int p(x, \gamma) r(\gamma) d\gamma$. First derive the distribution holding α, g_0, and σ^2 fixed. We then compute the mixture over the distribution of site ages α. Finally, we compute the mixture over g_0 and σ^2.

To obtain the distribution of users per site for a fixed α, g_0, and σ^2, 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 user populations 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 $\sigma^2 t$[6][7]. This means that the distribution of the number of visitors to sites created at the same time and with the same average growth rate is log-normal[8], 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, over time, site usage increases on average. 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 usage 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[9]. 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 users 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 market dynamics of the web is dominated by occasional bursts in which large number of users are suddenly drawn to given site. These bursts are responsible for the long tail of the probability distribution and make average behavior depart from typical realizations[10].

Next we factor in the fact that the distribution of the number of users of a site depends on the time that has elapsed since that 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 ones. Therefore the distribution of users 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 α is weighted exponentially. Thus, in order to obtain the true distribution of users per site that grow at the same growth rate, we need to compute the mixture given by

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

which can be calculated analytically to give

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

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 $\beta = 1 - \frac{g_0}{\sigma^2} + \frac{\sqrt{g_0^2 + 2\lambda\sigma^2}}{\sigma^2}$. We find that this range of values for β is consistent with our data analysis in which we found that $\beta_{edu} = 1.45$, $\beta_{adult} = 1.65$, $\beta_{overall} = 2.07$.

Given these results, one would expect that the oldest sites are the most dominant ones. But the measurements show that site popularity and age are only slightly correlated. For example the seventh most popular adult site was only created four months prior to our measurements - fairly young, even by Web standards. Clearly , assuming the same growth rate for all sites is insufficient and hence we need to take into account differences in the mean and variance of the increase in popularity of Web sites, since so far the distribution is valid for a single growth rate $g = g(g_0, \sigma)$.

Factors affecting a site's growth rate might include name recognition, advertizing budgets and strategies, usefulness and entertainment value of the site, and ease with which new users can discover the site, to name a few. Since each growth rate occurs with a particular probability $P(g)$, and gives rise to a power law distribution in the number of users per site with a specific exponent, the probability that a given site with an unknown growth rate has n_s users is given by the sum, over all growth rates g, of the probability that the site has so many users 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 market 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 users per site distributed according to a power law. This implies that on a log-log scale, the number of users per site, for large n, should fall on a straight line. We point out that since small values of n_s lie outside the scaling regime, our theory does not explain the data on sites with few users. A consequence of the universality of our prediction, is that the same power law behavior will be seen as more sites are created.

In summary, we presented empirical data on the nature of markets in the Web that shows that the winner-take all- phenomenon is pervasive over thousands of sites and particular domains. This implies that sites are rewarded by relative performance rather than absolute performance. As a result, the rewards tend to be concentrated into a few of the sites providing the same service, thus leading to very large difference in user visits to sites with similar offerings. We also developed a stochastic theory of the usage dynamics of the Web that takes into account the wide range of growth rates in the number of users per site, as well as the fact that new sites are created at different times in the unfolding story of the Web. This led to the prediction of a universal power law in the distribution of the number of users per site, which agrees with data from a portion of AOL logs, both for sites overall, and adult and educational sites separately. The existence of this power law implies the lack of any length scale for user populations on the Web. This is yet a another example of the strong regularities[11] 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 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] www.c-i-a.com/199902iu.htm

[2] Grossman, W. M, Cyber view, *Scientific American*, **280**, 38 (1999).

[3] www.aol.com

[4] Frank, Robert H. and Cook, Philip J. The Winner-take-all Society, Free Press, New York, NY 1995.

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

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

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

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

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

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

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