The Web's Hidden Order

Web site growth and popularity actually follow rules that can be explained mathematically and are useful for predicting the Web's future behavior.

LADA A. ADAMIC AND BERNARDO A. HUBERMAN

Hewlett-Packard Labs
Palo Alto, CA 94304
{ladamic,huberman}@hpl.hp.com

The past decade has witnessed the birth and explosive growth of the World Wide Web, both in terms of content and user population. Figure 1 shows the exponetial growth in the number of Web servers. The number of users online has been growing exponentially as well. Whereas in 1996 there were 61 million users, at the close of 1998 over 147 million people had internet access worldwide. In the year 2000, the number of internet users more than doubled again to 400 million[1]. With its remarkable growth, the Web has popularized electronic commerce, and as a result an increasing segment of the world's population conducts commercial transactions online.

From its very onset, the Web has demonstrated a tremendous variety in the size of its features. Surprisingly, we found out that there is order to the apparent arbitrariness of its growth. One observed pattern is that there are many small elements contained within the web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others. This diversity can expressed in mathematical fashion as a distribution of a particular form, called a power law, meaning that the probability of attaining a certain size x is proportional to $1/x$ to a power τ, where τ is greater than or equal to1.

When a distribution of some property has a power law form, the system looks the same at all length scales. What this means is that if one were to look at the distribution of site sizes for one arbitrary range, say just sites which have between 10,000 and 20,000 pages, it would look the same as for a different range, say 10 to 100 pages. In other words, zooming in or out in the distribution, one keeps obtaining the same result. It also means that if one can determine the distribution of pages per site for a range of pages, one can then predict what the distribution will be for another range.

Power laws also imply that the average behavior of the system is not typical. A typical size is one that is encountered most frequently, while the average is the sum of all the sizes, divided by the number of sites. If one were to select a group of sites at random and count the number of pages in each one, the majority of the sites would be smaller than average. This discrepancy between average and typical behavior is due to the skew of the distribution.

Equally interesting, power law distributions have very long tails, which means that there is a finite probability of finding sites extremely large compared to the average.

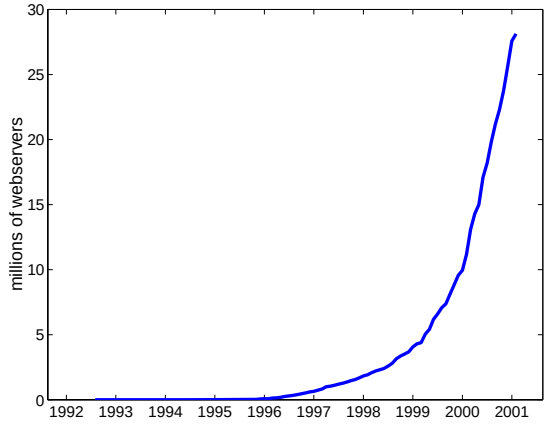

Figure 1: Growth in the number of web servers 1992-2001. *Source:* Netcraft Survey

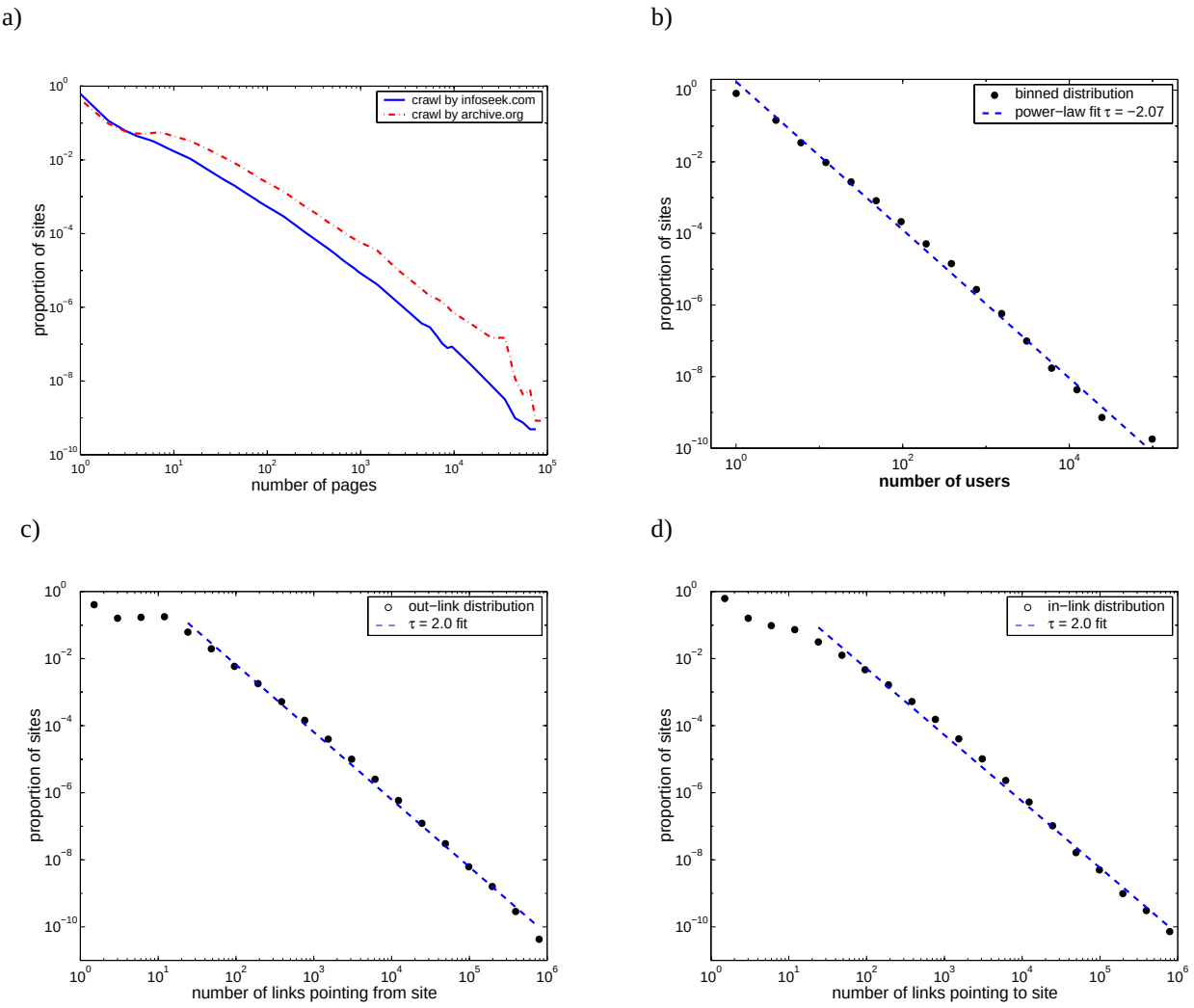

Figure 2. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links.

That this is quite striking can be illustrated by the example of heights of individuals, which follow the familiar normal distribution. It would be very surprising to find someone measuring 2 or 3 times the average height of 5'10". On the other hand, a power law distribution makes it quite possible to find a site many times larger than average.

In Figure 2, four power law distributions are shown: the distributions of the number of pages, visitors, and in and out links for a site. They all look almost identical because all four characteristics of a site evolve according to the same growth process. In order to describe this growth process[2], consider first the addition of pages to a site, in particular the addition of pages to site with a million pages. Such an enormous site must be maintained either by a very prolific author, or by a team of webmasters continuously modifying, deleting and adding pages. Some pages on the site might be automatically generated. One would not be surprised to find that the large site of a million pages has lost or gained a few hundred pages on any given day. Now consider a site with ten pages, a site that does not generate much content. Finding an additional hundred pages on this site within a day would be unusual - but not impossible. One could then safely state that the day-to-day fluctuations are proportional to the size of the site, i.e. the growth process is multiplicative. In other words, the number of pages on the site, *n*, on a given day, is equal to the number of pages on that site on the previous day plus/minus a random fraction of *n*.

If a set of sites is allowed to grow with the same average growth rate but with individual random daily fluctuations in the number of pages added, their sizes will be distributed lognormally after a sufficiently long period of time[3]. A lognormal distribution gives high probability to small sizes, and small - but significant - probability to very large sizes. But while skewed and

with a long tail, the lognormal distribution is not a power law one.

In order to explain the power law distribution of site sizes, one needs to consider two additional factors that determine the growth of the web: sites appear at different times, and some sites grow faster than others. First consider different start times. We know that the number of web sites has been growing exponentially since its inception, which means that there are many more young sites than older ones. Sites with the same growth rate appear at different times, only a few early on, but more and more as time goes on. After a sufficiently long time period, one finds a distribution that can be evaluated analytically and which is power law in the number of pages per site. The young sites, which haven't had much time to grow, are contributing to the low end of the distribution. The older sites, which are far fewer in number, are more likely have grown to large sizes, and contribute to the high end of the distribution.

In a second scenario, all sites appear at the same time, but their growth rates differ. We have demonstrated in simulations that different growth rates, regardless of how they are distributed among the sites, result in a power law distribution of site sizes. The greater the difference in growth rates among sites, the lower the exponent τ, which means that the inequality in site sizes increases. In summary, a very simple assumption of stochastic (random) multiplicative growth, combined with the fact that sites appear at different times and/or grow at different rates, leads to an explanation for the power law behavior so prevalent on the web.

This very same theory of growth can be applied to the popularity of web sites [4]. In this case, the day to day fluctuations in the number of visitors to a site is proportional to the number of visitors the site receives on average. Moreover, visitors belong to essentially two types. There are those who are aware of the site, and may or may not return to the site on this given day. A fraction of them does return, and this fraction varies from day to day. The second category of visitors are those who are visiting for the first time or rediscovering the site. Those belonging to the first type are familiar with the site and in turn influence the number of new visitors. The influence can be direct - one person telling or emailing another about a cool site they have just discovered, or one they use regularly. It can also be indirect, for a person that discovers an interesting site might put a link to it on his/her page, which in turn can act as a pointer for others to find it. A site with many users can get media coverage, which brings in even more traffic, with a consequent increase in the number of links from other sites. Finally, the amount of advertising a site can afford to pay to attract additional users depends on the amount of revenue it is generating,

% volume by user	% sites		
	all	adult	.edu domain
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 in general, adult sites, and .edu domain sites, as determined by counting the number of unique AOL visitors on Dec. 1, 1997.

and this revenue in turn depends on the number of visi tors. Hence the number of new visitors to the site is also proportional to the number of visitors on the previous day.

Once again, in order to understand the dynamics of site visits, we need to incorporate the fact that sites appear at different times and have different growth rates. Some grow quickly because they deal with a topic that is of interest to many people, others because they provide quality of service, still others because they are linked to from influential sites. Some sites may grow quickly because they bring in their clientele from the physical world, while others start on the internet but advertise heavily both on and off-line. Some gather their entire user base purely through customer loyalty and word of mouth advertising. Combining multiplicative growth at different rates with differences in site ages, one obtains a distribution in the number of visitors per site which once again is power law, with important consequences for the nature of e-commerce.

Figure 2(b) shows the distribution of unique visitors among sites from a portion of AOL logs obtained on Dec. 1, 1997. As can be seen from Table 1, the top 0.1% of all sites capture a whopping 32.36% of user volume. Moreover the top 1% of sites capture more than half of the total volume. This concentration of visitors into a few sites cannot be due solely to the fact that people find some types of sites more interesting than others. This we verified by performing the same analysis for two categories of sites: adult sites and sites within the .edu domain. 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.

The implication of this result is interesting both to the economist studying the efficiency of markets in electronic commerce, and to providers contemplating the number of customers the business will attract. As we verified, the distribution of visitors per site follows a universal power law, implying that a small number of sites command the traffic of a large segment of the Web population. 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[5], wherein the top few contenders capture a significant part of the market share.

We also see power law behavior in the number of links per site, a phenomenon which can be looked at from two ends. On the one end of the link there is the originating site, and on the other there is the receiving site. Both incoming links, as seen in Figure 2(c) and outgoing links, as in Figure 2(d), are distributed among sites according to a power law and follow the same growth process (see [6] for a similar treatment). The growth of links to a site can be equated with the growth of the site's popularity. The more a site is linked to, the more users are aware of the site, and the more additional links it receives. The growth of outgoing links is similar to the growth of a site in terms of the number of pages it contains. Outgoing links must constantly be maintained to record changes - some pages are moved or deleted. Other links must be added to keep up with pages which are appearing at an exponential rate. Some sites add links rapidly - they might be directories or index pages, others more slowly - they provide content rather than pointers to other resources.

The fact that there is a large discrepancy in the number of outgoing links a site has leads to the so-called small world phenomenon. While sites predominantly link to only a few sites of similar content, vast numbers of sites are linked together by directory and index sites which have thousands of links. Consequently one must move on average through only 4 sites in surfing from one site to any other[7]. On the page level, at most 19 links are required to move from any single page to any other[8].

This short review shows that in spite of its seemingly random growth, many properties of the web obey statistical laws that describe its structure in simple and non-trivial fashion. Equally important, and aesthetically pleasing to us, these laws can be derived from a dynamical organizing principle, which throws light into its evolution and future behavior. The knowledge of such strong regularities, such as the small world phenomenon, or the law of surfing[7], can be used to design better web services such as searches, or to increase the time spent by users at web sites[9]. As reflected in our study of online markets, these patterns apply not only to the virtual space of the Web but to interactions and transactions in the real world as well. As the information made available and captured online becomes richer, these methods will provide further insights into the dynamics of information and how people interact with one another.

References:

1. Computer Industry Almanac Inc., Internet Report, http://www.c-i-a.com/200103iu.htm

2. Huberman, B. and Adamic, L. Growth Dynamics of the World Wide Web, Nature, 401:131, 1999.

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

4. Adamic, L. and Huberman, B. The nature of markets in the world wide web, *QJEC*, 1:5-12, 2000.

5. Frank, R.H. and Cook, P. J. The Winner-take-all Society, Free Press, New York, NY 1995.

6. Barabasi, A.-L. and Albert, R., Emergence of scaling in random networks, *Science*, 286:509, 1999.

7. Adamic, L. The Small World Wide Web, *Proceedings of ECDL99, pp 443-452*.

8. Albert, R., Jeong H., and A-L. Barabasi, The Diameter of the World-Wide Web, *Nature*, 401:130, 1999.

9. Huberman, B.A., Pirolli, P., Pitkow, J. and Lukose, R., Strong Regularities in World Wide Web Surfing, *Science* 280:95-97, 1998.

10. Adar, E. and Huberman, B.A. The Economics of Surfing, *QJEC*, 1:203-214, 2000.

LADA A. ADAMIC (ladamic@hpl.hp.com) is a member of the research staff at Hewlett-Packard Laboratories, Palo Alto, CA.

BERNARDO A. HUBERMAN (huberman@hpl.hp.com) is an HP fellow at Hewlett-Packard Laboratories, Palo Alto, CA.