



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*

Whereas in 1996 there were 61 million Internet users worldwide, at the end of 1998, more than 147 million people were Internet users worldwide, and in 2000, the number of users more than doubled again to 400 million [1]. With its remarkable growth, the Web has popularized e-commerce, and as a result an increasing segment of the world's population conducts commercial transactions online.

From its onset in 1991, the Web has demonstrated notable variety in the size of its features, including connectivity and content. Surprisingly, we have found there is order to the apparent arbi-

★ THE PAST DECADE HAS SEEN THE BIRTH and explosive growth of the Web in terms of content and user population. Figure 1 represents the exponential growth of the number of Web servers; the number of users online

trariness of the increasing amount of Web content and its user population. One especially notable pattern we've observed is that there are many small elements within the Web but few large ones. At the

One especially notable pattern we've observed is that

THERE ARE MANY SMALL ELEMENTS WITHIN THE WEB BUT FEW LARGE ONES.

same time, a few sites consist of millions of pages, but millions of sites consist of only a handful of pages; few sites have millions of links, but many sites have only one or two. And millions of users flock to a few select sites, paying little attention to millions of others. This diversity can be 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 β , where β is greater than or equal to 1.

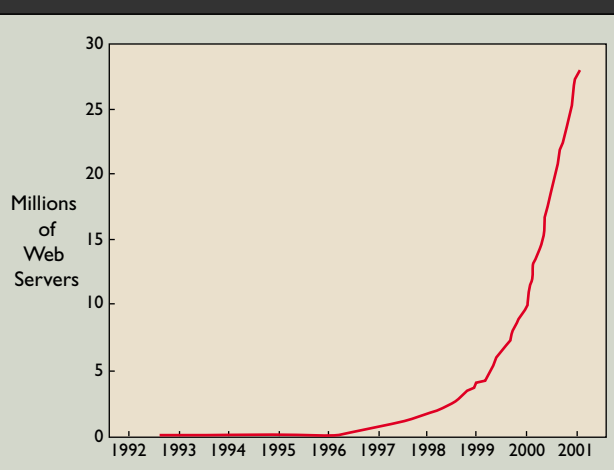
WHEN A DISTRIBUTION OF some property has a power law form, the system looks the same at all length scales. Therefore, if one were to look at the distribution of site sizes for one arbitrary range, say, just sites with from 10,000 to 20,000 pages, it would look the same as for a different range, say, 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 the distribution for many other ranges.

Equally interesting, power law distributions have very long tails, meaning there is a finite probability of finding sites extremely large compared to the

average. This finite probability of finding large sites is quite striking and can be illustrated by the example of the heights of individuals following the familiar normal distribution. It would be very surprising to find someone measuring two or three times the average human height of 5 feet 10 inches. On the other hand, a power law distribution makes it possible to find a site many times larger than average. Power laws also imply that the system's average behavior is not typical. A typical size is one that is encountered most frequently; the average is the sum of all the sizes divided by the number of sites. If one were to select a group of Web sites at random and count the number of pages in each of them, the majority would be smaller than average. This discrepancy between average and typical behavior is due to the skew of the distribution.

Figure 2 shows four power law distributions for number of site pages, visitors, inlinks, and outlinks. The distributions look almost identical, because all four site characteristics evolve according to the same growth process. In order to describe this process [2], consider first the addition of pages to a site, particularly when the site already has a million pages. Such an enormous site must be maintained either by a very prolific and dedicated author or by a team of

Figure 1. Growth in the number of Web servers 1992–2001.

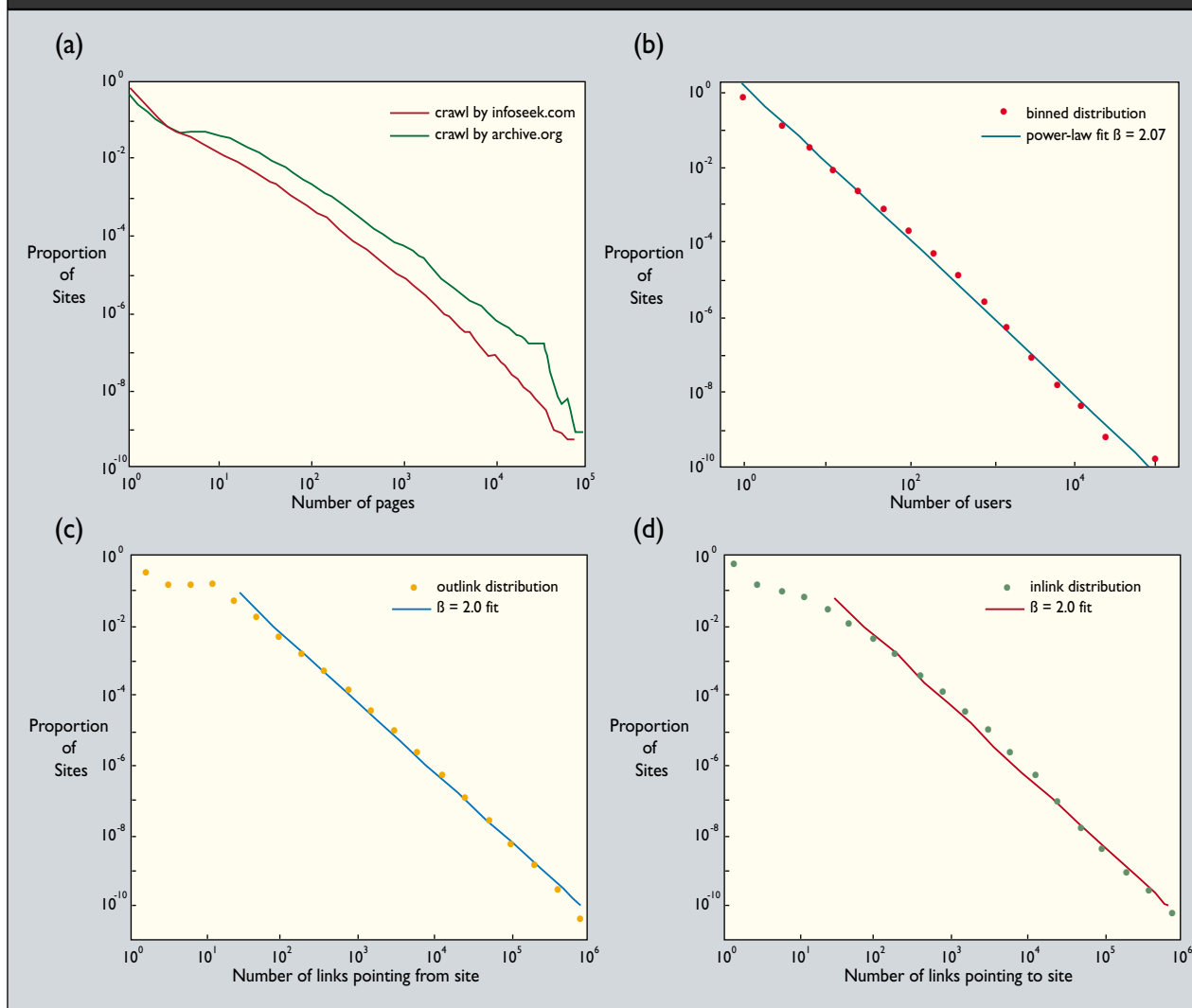


Source: Netcraft Survey

Web masters constantly modifying, deleting, and adding pages; some pages on the site might even be generated automatically. One would not be surprised to find that a large site of a million pages has lost or gained a few hundred pages on a given day.

Now consider a site with 10 pages that does not incorporate much content. Finding an additional hundred pages on the site within a day would be

Figure 2. Fitted power law distributions of the number of a site's (a) pages, (b) visitors, (c) outlinks, and (d) inlinks.



unusual, though not impossible. One could then safely say that the day-to-day fluctuations are proportional to the size of the site, that is, the growth process is multiplicative. 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, the size of the sites within each set of sites would 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 itself a power law distribution.

In order to explain the power law distribution of site sizes, one needs to consider two additional fac-

tors 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 the Web's inception, meaning there are many more younger sites than older sites. 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 period, one finds a distribution that can be evaluated analytically and which is power law in the number of pages per site. Younger sites, which haven't had much time to grow, contribute to the low end of the distribution. The older sites, which are far fewer in number, are more likely to have grown to large size 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,

Distribution of user volume among all sites, adult sites, and .edu domain sites, as determined by counting the number of unique AOL visitors Dec. 1, 1997.			
% of Volume by User	% of Sites		
	all sites	adult sites	.edu domain 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

regardless of how they are distributed among sites, result in a power law distribution of site sizes. The greater the difference in growth rates among sites, the lower the exponent β , meaning 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, provides an explanation supporting our contention of power law behavior on the Web.

Site Popularity

The same theory of growth can be applied to the popularity of Web sites [9]. 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 an average day. Moreover, visitors are one of essentially two types:

- Those aware of the site and who may or may not return to it on a given day. A fraction of them does return, though the fraction varies from day to day.
- Those 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; individuals tell or email other individuals about a cool site they have just discovered or about one they use regularly. It can also be indirect; individuals discovering an interesting site might put a link to it on their own pages that in turn acts as a pointer for others to find it. A site with many users might get media coverage, bringing 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 generates; this revenue in turn depends on the number of visitors. Hence, the number of new visitors to the site is proportional to the number of visitors 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 one or more topics of interest to many people; others because they provide quality of service; and still others because they are linked-to from influential sites. Some sites may grow quickly because they bring in clientele from the physical world, while others start on the Internet but advertise heavily online and off. Some gather their entire user bases purely through customer loyalty and word-of-mouth advertising. Combining multiplicative growth at varying rates with differences in site age, one obtains a distribution in 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 visi-

The 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.

tors among sites from a portion of AOL logs obtained on December 1, 1997. The table here shows the top 0.1% of all sites capture a whopping 32.36% of total user volume. Moreover, the top 1% of sites capture more than 50% 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. We verified this conclusion by performing the same analysis for two categories of sites: adult sites and sites within the .edu domain. We assumed adult sites would offer a selection of images and optional video and chat. We

also assumed .edu-domain sites would contain information about and for academic research, as well as personal homepages of students, staff, and faculty, thus likely covering any range of human interest. Again, the distribution of visits among sites was unequal. We sampled 6,615 adult sites by keywords in their names. The top site captured 1.4% of the volume to all adult sites, while the top 10% accounted for 60% of visitor volume. Similarly, the top .edu site—umich.edu—claimed 2.81% of the volume, while the top 5% of all .edu sites accounted for over 60% of the visitor traffic.

The implication of this result is interesting both to the economist studying the efficiency of markets in e-commerce and to providers contemplating the number of customers a business will attract. We verified that 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. For example, a newly established site will, with high probability, join the ranks of sites attracting a handful of visitors per day, and, with extremely low probability, capture a significant number of users. Such a disproportionate distribution of user volume among sites is characteristic of winner-take-all markets [4] in which the top few contenders capture significant market share.

We also see power law behavior in the number of links per site, a phenomenon that can be viewed from two ends. On one end of the link is the originating site; on the other is the receiving site. Both incoming links, as 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 described here. The growth in the number 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 in the number of outgoing links is similar to the growth of a site in terms of the number of pages it contains. Outgoing links must be maintained constantly to record changes; some pages might be moved or deleted. Other links must be added to keep up with pages appearing at an exponential rate. Some sites add links rapidly, possibly directories and index pages, others more slowly, providing content directly rather than through pointers to other resources.

The fact that there is a large variance in the number of outgoing links sites have 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 with thousands of links. Consequently a random Web surfer must move on average through only four sites in surfing from one site to any other site [5]. On the page level, 18 links (at most) are required to move from any single page to any other page [6].

Conclusion

In spite of its seemingly random growth, many properties of the Web obey statistical laws describing its structure in a simple and nontrivial fashion. Equally important, and aesthetically pleasing to us, these laws can be derived from a dynamical organizing principle that helps reveal the structure's historical evolution and future behavior. The knowledge of strong regularities, such as the small-world phenomenon, or the law of surfing [7], can be used to design better Web services, including searches, and increase the time users spend at Web sites [8]. 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*; see www.c-ia.com/200103iu.htm.
2. Huberman, B. and Adamic, L. Growth dynamics of the World Wide Web. *Nature* 401, 131 (1999).
3. Crow, E. and Shimizu, K., Eds. *Lognormal Distributions: Theory and Applications*. Marcel Dekker, New York, 1988.
4. Frank, R. and Cook, P. *The Winner-Take-All Society*. The Free Press, New York, 1995.
5. Adamic, L. The small World Wide Web. In *Proceedings of the 3rd European Conference on Digital Libraries ECDL99* (Paris, 1999). *Lect. Notes Comput. Sci.* 1696, (1999), 443–452; see www.parc.xerox.com/iea/papers/smallworld/.
6. Albert, R., Jeong H., and Barabasi, A.-L. The diameter of the World Wide Web. *Nature* 401, 130 (1999).
7. Huberman, B., Pirolli, P., Pitkow, J., and Lukose, R. Strong regularities in World Wide Web surfing. *Science* 280 (Apr. 3, 1998), 95–97.
8. Adar, E. and Huberman, B. The economics of surfing. *Quart. J. Electron. Comm.* 1, 3 (2000), 203–209.
9. Adamic, L. and Huberman, B. The nature of markets in the World Wide Web. *Quart. J. Electron. Comm.* 1, 1 (2000), 5–12.

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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.