

The Laws of the Web  
Patterns in the Ecology  
of Information

Bernardo A. Huberman

The MIT Press  
Cambridge, Massachusetts  
London, England

3

Evolution and  
Structure

Figure 3.1 (and in color on the jacket of this book), a seemingly random set of lines and dots, depicts an actual snapshot of a piece of the World Wide Web, corresponding to the collection of Web sites and links in Finland that existed about a year ago. This picture can be made a bit more intelligible by noting that the circles denote individual Web sites and the lines correspond to the links between them. But in a way this clarification is of little help. Any reader trying to make sense of this rather abstract-looking painting will eventually conclude, and rightly so, that this messy diagram, however appealing it appears, conveys little useful information besides its dubious artistry. It actually looks as if it were drawn randomly by some computer program or video game.

However, this apparently random figure contains a hidden and general pattern, found not only in the sector of the Web corresponding to Finland but everywhere else in the World Wide Web, from the United States to Europe or Asia. This hidden pattern, which I will discuss in this chapter, throws much light on the evolution of the Web and its structure. Equally important, its explanation and emergence as the Web grows relies on a powerful methodology that has turned out to be

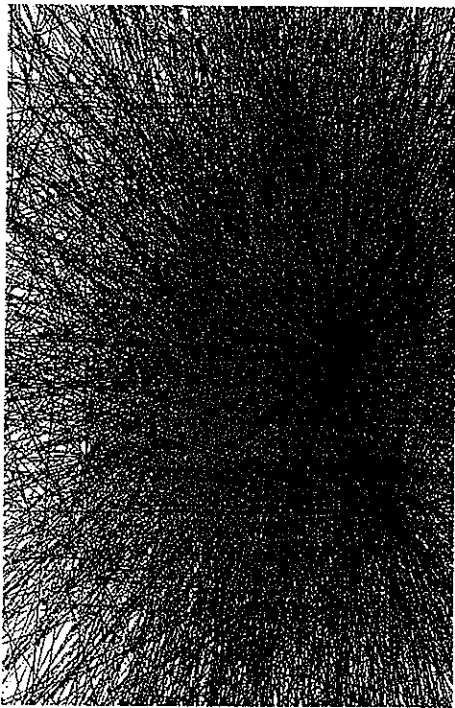


Figure 3.1  
Structure of the World Wide Web in Finland. The circles denote sites and the lines connecting them are links among them.

extremely useful in dealing with large distributed systems such as the Web.

This methodology aims to establish a tight relation between the local properties of large distributed systems and their global behaviors. Because of its analytical structure and its connection with observable properties of the Web, it has much predictive power. Examples of systems to which it can be applied are social organizations, markets, ecologies, and, most prominent, the Internet. In each case, the overall structure and dynamics of the system is determined by the collective interactions of its many autonomous parts, which include consumers, organisms, computer programs, or people. And because of the myriad interactions that are possible among the constituents of such large systems, the resulting behavior exhibits a panoply of complex and fascinating

results that range from the apparent stability of large ecosystems to the wild fluctuations of financial markets.

The connection between the actions of individuals and the global patterns one observes is not always an obvious one, the reason being that the system behavior cannot be explained by simply adding up all the actions and intentions of its individual parts. If that were the case, it would be an easy task to add all the possible actions in order to predict the resulting behavior.

Systems whose behaviors cannot be explained by just adding all the partial actions of their constituents are known as nonlinear, and while they are more difficult to study than linear ones, understanding them is worth the increased effort. This is because the dynamics of nonlinear systems can be fascinatingly complex. For example, in addition to a simple equilibrium situation where the overall system does not change over time, or the textbook situations where patterns sharply repeat in clocklike cycles, nonlinear systems can also display erratic behavior even when their mathematical description is totally deterministic. In this so-called chaotic situation, if one were to start the system with a given condition and follow its evolution over time, the end result would be vastly different from what would follow if one were to start all over again from an almost identical initial condition. As a result, the behavior of systems that have this extreme sensitivity to their initial conditions appears to be extremely erratic, and the only predictions that can be made about their behavior are probabilistic in nature.

Nonlinearity is not the only complicating feature in the analysis of systems as complex as the Web. Its distributed nature also poses a formidable problem for its study. This is because the parts that make up the Web—sites, links added to them, or pages—can display complex nonlinear dynamics.

And when answers are found, their implications can be quite stunning, as they often imply an effective disconnection between the well-defined behavior of the components and the global outcome that one observes. What this means is that if one were looking at an economic system, for example, the precise knowledge about the plans and strategies of individuals in the market would not suffice to understand the behavior of a market. This important insight was first articulated by Frederick von Hayek (1937), when he stated that while economic outcomes are the result of actions by people, they do not necessarily reflect their intentions.

An example will help explain this important point. Consider an individual investor in the stock market of the day trader variety, and let's imagine that one can track all of his transactions in the market, as well as the messages that he exchanges with others. One might even imagine a rather implausible scenario in which one has access to his thoughts and deliberations on how and when to invest in particular stocks. If one were to do this for a very long time, one could end up learning whatever strategy this investor is using, and how successful it is for trading in the market. Even better, one might even get to know him so as to be able to predict with certainty his decisions in terms of buying or selling particular stocks. And while this imaginary scenario would teach one much about a given trader, it cannot predict the price of stocks in which he is about to make a decision. This is because the price of a stock depends on the behavior of many investors and their collective decisions—decisions that interact with each other in such complicated ways so as to drown out the effect of a single trader, who himself cannot anticipate the result of his own actions in determining the price of given stocks.

The example that I just described could be replaced with that of city traffic, where detailed knowledge of the driving

patterns and intentions of a single driver cannot be used to predict congestion at an intersection, or to forecast which streets will facilitate traffic flow at given times. And the same applies to the characteristics of the Web, where detailed knowledge of how a given site grows by the addition of new pages and links to others is not enough to understand a picture like the one in figure 3.1.

Since following a single individual in her surfing behavior on the Web will not predict much about surfing in general, or how congestion takes place on the Internet, or the commercial success of given businesses, we must abandon such individual knowledge and replace it with something more aggregate, the behavior of the system as a whole. In order to do so, we developed methods for treating large distributed systems that are largely inspired by the success that physics has had with explaining the behavior of matter in terms of its constituents, such as atoms and molecules. These methods are statistical in nature and resort to dynamical formulations that lead to precise predictions that can then be tested experimentally.

This aggregate way of looking at a large system is a powerful methodology for dealing with large distributed systems, from stock markets to computer networks and social organizations. It provides a bridge between the particular and the whole, a way of reasoning that resorts to the dynamics of averages and the behavioral departures from these averages, while keeping the essential ingredients of what the component pieces do. It is thus possible to link the growth pattern of a given Web site to the total number of pages in the whole Web, or to make a connection between the individual intentions of users and the number of people who visit a search portal like Yahoo! in a week or a month. And so on.

The insights gained from this increased understanding of distributed systems when using knowledge about the behavior

of their individual components has already led to improved methods for cooperative search algorithms, organizational design, and even distributed building controls. This gave me confidence that they would also be of use when applied to the Web, and indeed they did lead to the discovery of a number of strong regularities, such as the way the Web grows, how people surf it, the nature of markets in e-commerce, and how the act of downloading pages from a site contributes to the observed patterns of congestion.

How can we use this methodology to understand the growth patterns of the Web, which from its inception has demonstrated a tremendous variety in the size of its features? This variety is apparent to anyone who surfs the Web and anyone who notices the existence of large sites in terms of the number of pages they have, and also small ones consisting of one or two pages and few links to others. This is a natural reflection of the arbitrary way in which people design their own sites and decide what to link them to. A site belonging to a large firm, for example, might contain a lot of pages linked to each other and other sites, whereas that of an individual might have some biographical data, a picture, and one or two links to some of her friends.

Surprisingly, when one studies the structure of the Web on a large statistical basis, one finds out that in spite of the apparent arbitrariness of its growth clear patterns exist that reflect hidden regularities. 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 pages contain millions of links, but many pages have one or two.

This diversity can be expressed in mathematical fashion as a distribution, a mathematical entity that quantifies how many instances of a given size, say, appear in the system one

studies. The distributions describing the patterns observed on the Web have a particular form, called a power law. When we say that a distribution has a power law characteristic, we mean that the probability of finding a Web site with a given number of pages,  $n$ , is proportional to  $1/n^\beta$ , where  $\beta$  is a number greater than or equal to 1.

The interesting thing about a distribution with a power law form is that if a system obeys it, then it 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 that have between 10,000 and 20,000 pages, it would look the same as that for a different range, say from 10 to 100 pages. In other words, zooming in or out in the scale at which one studies the Web, 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 of pages.

This power law distribution describes the number of pages per site, and also the number of links emanating from a site or coming to it. It is a robust empirical regularity found in all studies of the Web.

Figure 3.2 shows two examples of such power laws for the Web, which appear as straight lines because the scales are linear in each decade (or mathematically, logarithmic).

In figure 3.2, two power law distributions are shown: the distributions of the number of pages per site and those of links from one site to other sites (outlinks). Both look almost identical because both are power laws. If they were not, different shaped curves would result when plotted on scales that are linear in each decade.

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

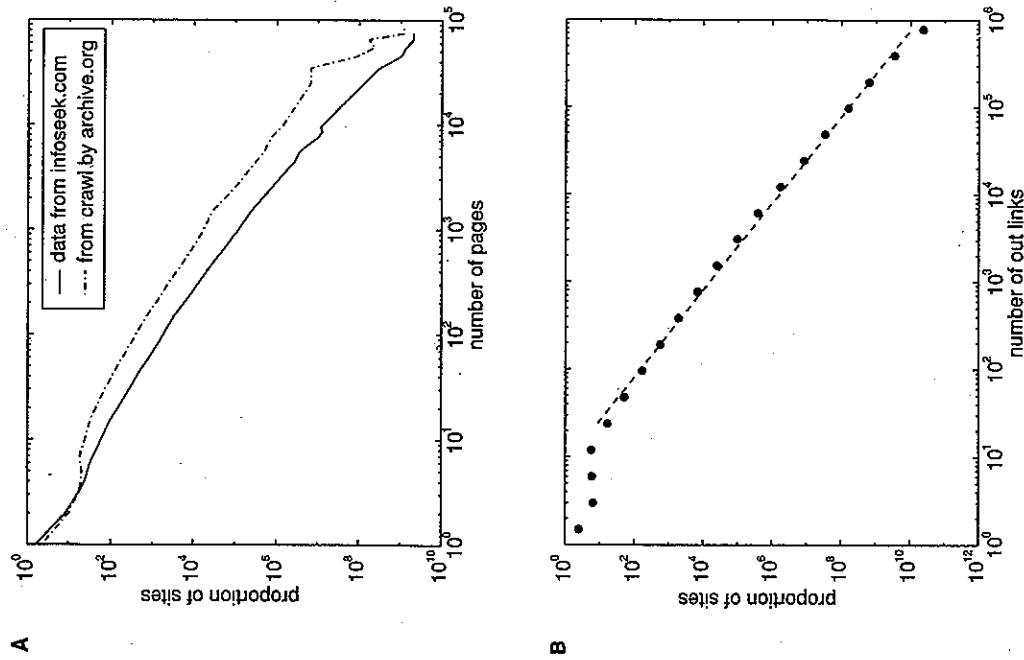


Figure 3.2  
The proportion of Web sites having a given number of pages (A) and links (B), plotted in logarithmic fashion.

average site. That this is quite striking can be illustrated by the heights of individuals, which we know follow the familiar bell-shaped normal distribution. The normal distribution determines, among other things, the average height of a person, which nowadays is about 5 feet 10 inches. Such a distribution is not a power law one, but rather it decays very fast for samples that depart from the average. Thus, one would find it very surprising to be walking in a city and to find someone measuring two or three times the average height of 5 feet 10 inches. On the other hand, when a distribution of some property, like the size of Web sites, has a power law distribution, it is quite likely to find a site many times larger (in terms of the number of pages or links) than the average size.

Another peculiar consequence of a power law is 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. Thus, 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 have a smaller number of pages than the expected average. This discrepancy between average and typical behavior is due to the fact that a power law distribution, unlike the familiar bell-shaped one, is not symmetric around its maximum but skewed, and has a long tail.

The fact that the number of pages per site, and also the number of links per site, is distributed according to a power law is a universal feature of the Web. It holds throughout the World Wide Web, irrespective of the type of sites that one considers, from the smallest to the largest, and regardless of the nature of the site. The appearance of such a strong regularity out of a seemingly random process is quite striking, and points to some kind of universal mechanism that not only

underlies the growth of the Web but also produces a power law distribution in some of its characteristics.

In order to describe this growth mechanism, consider first how pages are added to a site, say, a big 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. Equally possible, 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 just 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 say that the day-to-day fluctuations are proportional to the size of the site or, as it is stated in the mathematical description of this process, that the growth 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 or 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, after a sufficiently long period of time their sizes will be distributed according to a distribution that is known as lognormal. 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, which is what one observes.

In order to explain the power law distribution of site sizes that one observes, one needs to consider two additional factors that determine the growth of the Web. The first one is that sites appear at different times, and the second is that some

sites grow faster than others. A first scenario takes into account different start times. One knows 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 that has a power law behavior 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 to 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. By using computer simulations, we demonstrated 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  $\beta$ , which means that the inequality in site sizes increases.

In summary, a simple assumption of random multiplicative growth, combined with the fact that sites appear at different times and/or grow at different rates, leads to an explanation of the power law behavior so prevalent on the Web.

The existence of this scale-free power law describing the number of pages per site for the whole Web is not only interesting but also useful. For example, a search engine that crawls the Web cannot know a priori how many pages per site it will encounter. If a program is written to stop downloading every page of a site beyond a certain number, this power law can tell, in a probabilistic sense, how many pages are left to crawl in that site. Another way of saying the same thing is to state that once a search engine crawls a large sample,

knowledge of that distribution is enough to predict the rest of the crawl.

The distribution of the number of pages per site is not the only hidden regularity in the structure of the Web. As shown in figure 3.2, power law behavior is also observed when one studies the number of links per page, which were obtained from a crawl of 260,000 sites. By "site," I mean an address where each site is a separate domain name. If one counts how many links sites receive from other sites, as one found out for pages, the distribution of links among sites is power law. Equally interesting is the discovery that no correlation exists between the age of a site and the number of links it has.

This absence of a correlation between age and the number of links is hardly surprising, for all sites are not created equal. A site with very popular content, which appeared in 1999, will soon have more links than a bland site created in 1993. It is likely that the rate of acquisition of new links is proportional to the number of links the site has already. After all, the more links a site has, the more visible it becomes and the more new links it will get. This means that there the growth rate varies from site to site.

The theory that accounts for the power law distribution in the number of pages per site can also be applied to explain the number of links a site receives. In this model, at each time step the number of new links a site receives is a random fraction of the number of links the site already has. New sites appear at an exponential rate and each has a different growth rate, and when analyzed mathematically, this model explains the data well.

Before I continue, it is of interest to contrast these power laws with a similar kind of regularity that it is not only found on the Web, but in many other situations as well. This regularity goes under the name of Zipf's Law, in honor of George

Kingsley Zipf, a Harvard linguistics professor who sought to determine the frequency of use in English texts of the third or eighth or one-hundredth most common word. Zipf's Law states that the size of the  $r$ th largest occurrence of the event is inversely proportional to its rank, and so mathematically it looks very much like a power law. As a matter of fact, both power laws and Zipf's Law are used to describe phenomena where large events are rare, but small ones quite common. For example, few large earthquakes but many small ones occur. Few mega-cities but many small towns exist. A few words, such as "and" and "the," occur very frequently, but many occur rarely. And while there are a few multibillionaires, most people make only a modest income. Just as we found power law distributions in several properties of the Web, Zipf's Law also gives power laws but expresses them in terms of rank rather than numbers.

It is a familiar experience to meet someone for the first time at a party, business meeting, or conference, and to soon discover that one shares an acquaintance or perhaps a family relationship that traces back to some common ancestor or distant uncle. The commonality of this case is the manifestation of a surprising and interesting social phenomenon, sometimes called "six degrees of separation," which holds that between any two people on this planet is a path of no more than six acquaintances linking one person to the other. While the exact number might not be six, it is the case that most often a short chain of acquaintances is all one needs to connect any two people chosen at random. This remarkable fact was discovered by Harvard sociologist Stanley Milgram who, in the 1960s, asked a number of randomly chosen people in a small town in the Midwest to mail a postcard to a friend of his, who happened to be a stockbroker living in Boston.

What made this experiment unusual was that rather than asking these people to mail the postcard directly to his friend, Milgram instructed them to send the cards by passing them from person to person, with the proviso that the cards should be passed to someone the passer knew on a first-name basis. Since it was highly unlikely that the initial group was

depend on the existence of such short chains of acquaintances, in ways that have been documented by social network scientists for over four decades.

In the context of the Internet this phenomenon can be readily explored on the Web at a site called <http://www.starwars.com/6degrees/>. Inspired by the phenomenon of six degrees of separation, it offers its users the possibility of discovering all sorts of small world connections among the characters and actors of the *Star Wars* movies.

Milgram's experiment raised two complementary and interesting issues (Milgram 1967). The first one had to do with the properties that networks must have to become small worlds. If one were to draw a network consisting of nodes that would represent people, and links among those nodes that would represent who knows whom, it is by no means obvious that any two nodes would be separated by six links. There is something particular about a social network that is reflected in the link structure of the network.

The second issue concerns what the best strategies are for navigating such small-world graphs in a short number of steps. The people participating in Milgram's experiment did not have detailed knowledge of the social network in which they were embedded, and yet they managed to pass the messages in a fairly short number of hops. Even knowledge of being part of a small world network does not necessarily translate into a smart strategy to deliver a message to a target person.

The first issue was partially addressed by Watts and Strogatz (1998), two mathematicians at Cornell University, and by researchers at Notre Dame University (Albert, Jeong, and Barabasi 1999). Watts and Strogatz created a procedure whereby a regular lattice can generate a small-world graph. Basically, they showed that a regular lattice can be transformed

acquainted with the Boston stockbroker on a first-name basis, they had to pass it on to someone that they felt would be closer to Milgram's friend either geographically or professionally.

Some of the cards eventually made their way to the target, and when Milgram analyzed the route they took he discovered that the average number of steps taken to get from the town in the Midwest to Boston was only about six, which shows that most people in a large population are connected by only a short chain of acquaintances. His finding was eventually confirmed in a number of careful social studies that range from friendships in high school to some religious communities. The generality of this result found its way into popular culture as plays and newspaper articles eventually highlighted the six degrees of separation phenomenon.

The phenomenon of six degrees of separation also inspired a number of games, such as Six Degrees of Kevin Bacon, where one attempts to find the shortest path from any actor to Kevin Bacon. Among mathematicians, authors of papers that are even distantly related to the work of the great Hungarian mathematician Paul Erdos carry a so-called Erdos number, which describes the distance to having co-authored a paper with him. Thus, for a mathematician to have an Erdos number of 1 means that he or she wrote a paper with Paul Erdos, while having an Erdos number of 2 implies having published a paper with someone who was a co-author of Erdos. What started as a game turned into a truth: The smaller the Erdos number, the higher the "prestige" felt by those who advertise it.

The existence of these so-called small worlds is not just a curiosity with possibly interesting mathematics behind it. It has practical and important consequences as well. Political influence, searching for a job, even the spread of diseases and other forms of social contagion, such as rumors or news,

into a small-world network by making a small fraction of the connections random. But small-world networks are not just random graphs, since they have the property that they exhibit a high degree of clustering. This means that nodes are interconnected in such tight fashion that it would be unlikely to find such clustering in a random graph. By comparison, random graphs are not clustered and have short distances, while regular lattices tend to be clustered and have long distances.

One problem with this construction, however, is that these small world graphs lack an important property of many networks, namely, their approximate power law distribution in the number of links. This distribution amounts to stating that a few nodes or people or sites in the Web have many links whereas most have a few. Whereas some small-world graphs (for example, the electric power grid) are not power law-like, many are, such as the graph of who telephones whom. Such data can be obtained from the main telephone carriers.

Albert, Jeong, and Barabasi (1999), on the other hand, described a procedure for producing random graphs with a power law distribution in the number of links per node while failing to produce graphs that also have the clustering property of small worlds. While this work seems to explain the backbone structure of the Internet, it fails to account for the clustering property known to exist in the link structure of the Web. So, we are still in search of a proper understanding of small-world graphs that have a power law distribution in their link structure.

The issue of navigation in small-world graphs was also partially addressed by a computer scientist at Cornell University, John Kleinberg, who, starting with a perfect lattice, was able to construct a random graph with the right clustering properties. Unfortunately once again, the distribution of the number of links per node is not a power law, a fact that

makes it irrelevant to those real-world problems where power laws are observed. The particular navigational algorithm that Kleinberg (2000) proposed has the property that a node has no knowledge of where the links from his neighbors go. While not leading to an optimal solution to the traversing of the shortest path between any two nodes, it does show the existence of fairly short paths for a particular value of the model parameters.

If the underlying mechanism for such small chains of connections among people reflects some mathematical property of random networks, and the Web is a good example of such a network, one may wonder if the same small world phenomenon exists among sites and pages, as opposed to people. Recently, Lada Adamic of the Xerox Palo Alto Research Center undertook a study of the average number of links that one has to traverse in order to go from one site of the Web to the other. Adamic (1999) found out that just as in the social sphere, one could pick two Web sites at random and get from one to the other within about four clicks. That this phenomenon constitutes a strong regularity of the Web was confirmed by looking at the link structure of a Web repository containing over 50 million pages and 260 thousand sites. Figure 4.1 exhibits this remarkable regularity in graphical fashion.

This phenomenon was also shown to exist for the number of links between any two pages, as opposed to sites of the Web, by researchers at Notre Dame University (Albert, Jeong, and Barabasi 1999). In this case the number is nineteen, as opposed to the four one encounters between sites.

The small-world phenomenon on the Web is not only interesting in itself but also useful, since it can be exploited in the design of better search engines and for the marketing of specific products in the world of electronic commerce. The reason for its usefulness lies in the common observation that a

across sites besides the most common ones. What she noticed was that connected clusters spanning several sites tend to contain the main relevant pages and are rich in "hubs," or pages that contain links to many other good pages. One can then find the center of the cluster by computing the number of links among all the members of the cluster.

What this means is that rather than presenting a list of documents that contains many sequential entries from the same site, a search engine using the phenomenon of the small world can present just the center from each cluster, and then users can explore the rest of the cluster on their own.

The phenomenon of the small world on the Web has implications that go beyond the improvement of search engines. This is because the link structure of the Web implies the existence of communities that share some common affinities. The Web represents a wide range of human interests, and as a result some sites are devoted entirely to a single interest or cause. Others, such as Yahoo!, have clubs or chat rooms where people can meet and share their ideas on particular topics. Since many people document their interests and affiliations on their personal homepages, and link them to people with whom they have some common interests, the exploration of the link structure of documents on the Web can reveal the underlying relationship between people and organizations.

Recently, Lada Adamic and Eytan Adar realized that if both social networks and the Web were small-world graphs, then one would expect that networks of personal homepages would be small-world graphs as well. They confirmed this conjecture by studying the networks of personal homepages at two main universities in the United States.

By looking at listings of friends on the Stanford University and MIT homepages, they found out that users typically link

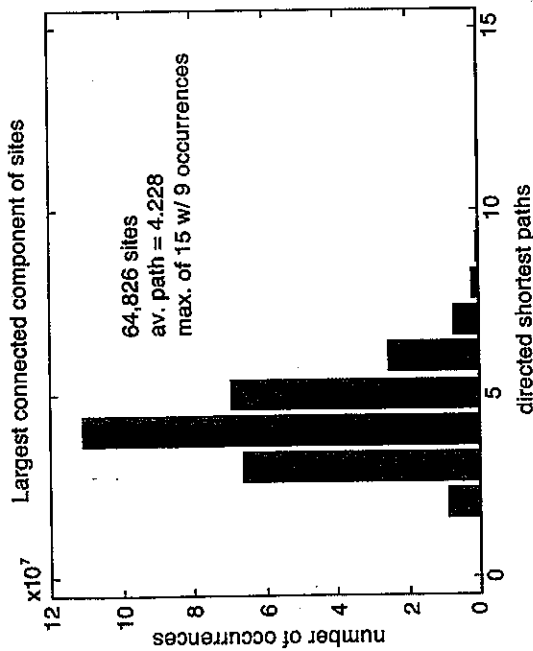


Figure 4.1 Histogram depicting the number of occurrences among sites having given directed shortest paths. From Adamic (1999).

good quality Web document tends to link to other good documents of similar content. Thus, one expects that within the Web there are groups of pages of similar content and perhaps quality, which refer to one another. The quality of these pages is guaranteed by the recommendations implicit in the links among them.

Adamic built an application of these ideas around a repository of Web pages crawled by the search engine Google in the first half on 1998. For any given search word, her application returned queries according to their page rank and their text match, while also providing link information for each page. It then identified all the connected clusters and selected the largest one, since most likely it is the one that contains links

to only one or two other users, with a very small but still significant fraction linking to dozens of users. This is yet another manifestation of the power law-like distributions discussed in chapter 3. In this case such power law implies that one finds some users with lots of links to others while most users have a few links to others. Some users are very popular, attracting lots of links, while most get only one or two. The more startling result that they uncovered is that users linking to only 2.5 other people on average create a virtual connected social network of 1,265 people and a few smaller networks. Furthermore, exploiting the notion that people who link to each other usually have something in common, they could predict who could be friends with whom by analyzing text, links, and mailing lists. Their methods show a lot of promise for discovering small-world communities of people by studying the way they link their pages.

## 5 As We Surf

### 5

Bookstores, particularly those offering a warm and inviting atmosphere, seem to have a special hold on many people. It is a common experience to drift into one's favorite bookstore and to browse books and magazines, drifting from one table or shelf to the other, connecting with topics close to one's interests, noticing new titles, sensing trends, or looking for news beyond that available at home. Whether one ends up buying a book or magazine is not that important; what matters is having spent a quiet time browsing and learning about the new offerings in print.

On reflection, this leisure browsing activity can appear quite mysterious and arbitrary to a curious observer, for if there is a pattern to it is by no means apparent. What makes us look at some books and not others? How long do we finger a book before moving to the next one? What makes us shift from one subject to the other, or what holds our attention to a particular page? Is it purely idiosyncratic, or is there a pattern that while different in its details applies to most people?

Move the focus away from bookstores onto the Web, and the browsing that I just described becomes replaced by the activity of clicking on page or site links that take the user from page to page within a site, or to new sites offering a