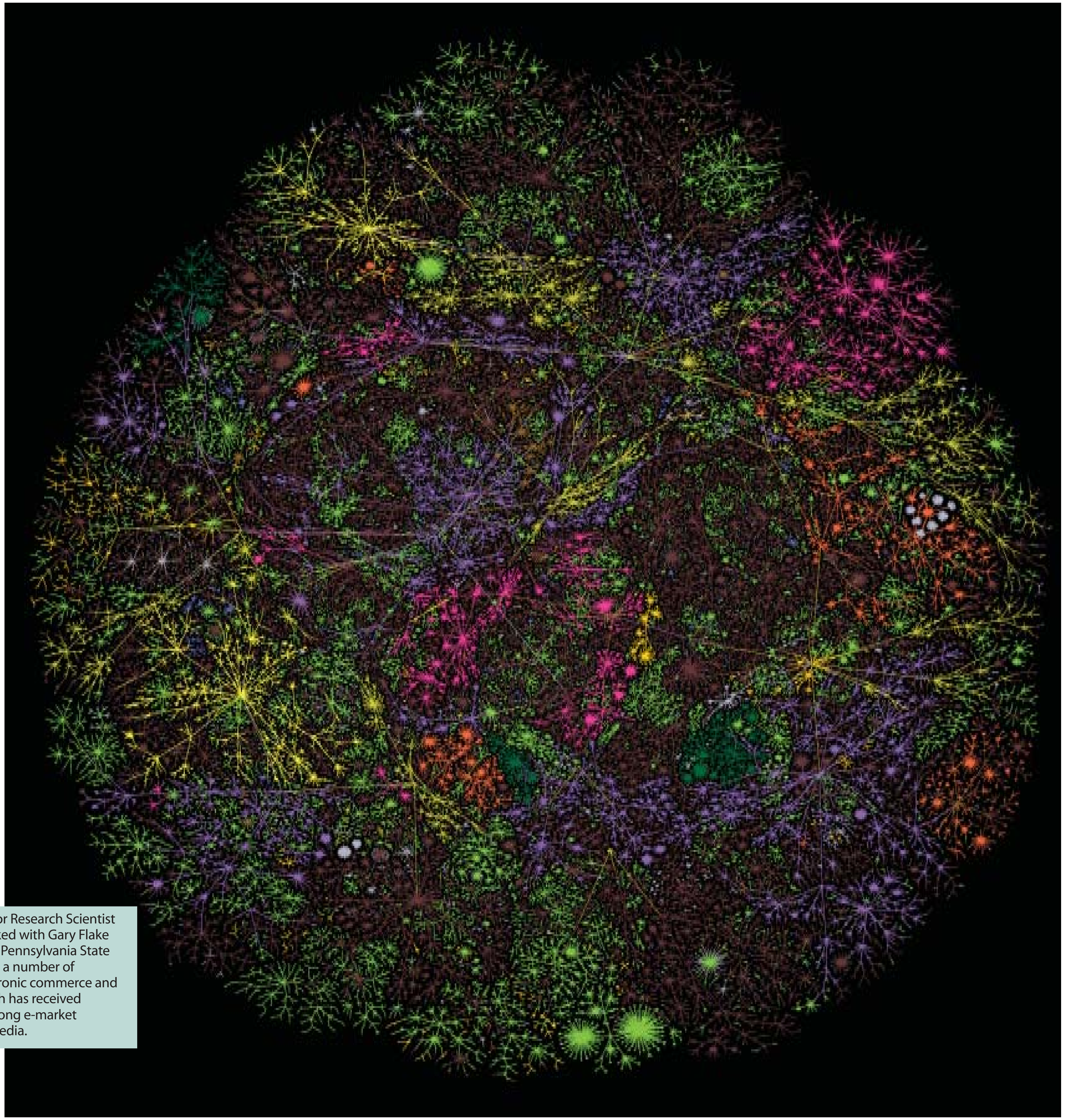
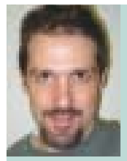
# 6 Self-organization, Self-regulation, and Self-similarity on the Fractal Web

**Gary William Flake and David M. Pennock**
**Yahoo! Research Labs**

The authors begin by modelling the World Wide Web as an ecosystem, a fractal, which reflects an intimate coupling of people, programs, and pages. Viewing the Web from a variety of scales and viewpoints, from macrocosmic to microcosmic, it is evident that users, authors, and search engines all influence one another to yield an amazing array of self-organizing, self-regulation, and self-similarity. Ultimately, the Web's organization is intimately related to the complexity of human culture and to the human mind, and it is this subtle relationship between humanity and the Web that is responsible for the Web's amazing properties.





Gary Flake is a leading researcher in Web analysis and modelling, and author of *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation* (MIT Press, 1998). He is presently head of Yahoo! Labs in Pasadena, California.

David Pennock is a Senior Research Scientist at Yahoo! Labs, and worked with Gary Flake at NEC. He has taught at Pennsylvania State University, has taken out a number of patents relating to electronic commerce and the Web, and his research has received significant attention among e-market companies and in the media.

# *The Web as an Ecosystem*

The World Wide Web is a digital entity like no other. Over the course of roughly fifteen years – and at an exponentially increasing rate – the Web has managed to capture, collect, organize, and connect a stunning amount of humankind's collective knowledge. It now reflects almost every aspect of our collective culture: from the peace prize to pornography; from academia to e-commerce; and from the mega-corporation to the personal home page. Though the Web is certainly a unique object in the history of the world, at its heart the Web is a social creation, and so perhaps it is not surprising that many of the Web's properties mimic those of nearly every other social and biological entity. The Web is in a very real sense an *ecosystem*, and as such can be viewed from a number of different perspectives spanning the microscopic to the macroscopic, with each vantage point showing an astonishing amount of complexity.

Natural ecosystems derive much of their complexity from a vast number of interdependencies: predators consume prey; individuals compete for the opportunity to reproduce; symbiotes cooperate with other species for improved viability; and the expired biomass from all organisms ultimately fuels the microbes at the lowest level of the food chain. In this way, an ecosystem is endlessly circular, with chains of dependencies streaming between individuals and species.

Fig. 6. 1 previous page: A map of part of the Internet's topology, updated March 2004, illustrating the macroscopic structure of the Web and the apparent fractal nature of link connectivity. Points correspond to distinct Internet addresses of computers on the Internet; lines correspond to the connections between computers.

Data and visualization courtesy Bill Cheswick and Hal Burch of Lumeta Corporation. Lumeta is a pioneer in analyzing and securing corporate networks, http://www.lumeta.com. Reprinted by permission.

We say that an ecosystem's state is *recursive* because of the circularity of the ecosystem's dependencies. The future of every creature is intimately coupled to the present state of every other member of the ecosystem. As a result, the life cycle of a single individual as well as the evolution of an entire ecosystem are both tremendously complex precisely because each is a function of the other.

The circular dependencies of the Web are rich as well. Web authors attempt to build pages that a target audience of users will value, and the authors add value by supplying a mixture of content and *hyperlinks* (or more simply, *links*) to other valuable pages. Hence, one instance of recursion on the Web is that valuable pages tend to accumulate incoming links, and pages can become more valuable by linking to other valuable pages. The subtlety of the Web's recursion partially hinges on the circular influences that authors and users have on one another, each taking actions that are influenced by the other. To complete the analogy between the Web and natural ecosystems: the behaviours of individual authors or users as well as the evolution of the entire Web are tremendously complex precisely because each is a function of the other.

Throughout this chapter, we will use the analogy between natural ecosystems and the Web to better explore the Web's fractal properties and from whence they come. We will focus on three different vantage points: the microscopic level of the individual author or user (single organism), the intermediate level of the Web community (the niche or species), and the macroscopic level of the entire Web (the entire ecosystem or biosphere). But first, we will step back from the Web completely to examine its origin and evolution.

## A Birds Eye View of People, Programs, and Pages

Before we dissect the Web in terms of scale, it is valuable to stand back and take a look at the Web's evolution in the broader context of human behaviour. Doing so will allow us to better understand and appreciate how the different scaling properties of the Web relate to one another and what external forces drive the dynamics of the Web. For this discussion, we will focus our attention on how users (people), search engines (programs), and Web sites (pages) impact one another.

At any moment in time, one can (in theory) measure the number of users that view a page over some period, the likelihood that a page will be the result of a typical query sent to a search engine, and the number of links that point to a particular page from other Web pages. Let's refer to these three properties more simply as the 'traffic,' 'rank,' and 'connectedness' of a page, respectively. Notice that each attribute superficially appears to be determined by only one type of thing: users determine traffic, search engines determine rank, and pages (and by implication authors) determine connectedness. However, in reality, all three properties are deeply intertwined; but it was not always this way.

In the beginning of the Web, there were no search engines, only links. As a result, users could visit a Web page only by directly typing in a URL (the part at the top of your browser that typically begins with 'http://'), or by clicking on a link. Relative to each other, a click is far easier for a user to do than it is to type in a URL. This leads us to the first observation on the relationship between traffic and connectedness:

*The greater a page's connectedness, the greater its traffic.*

After all, if users predominately arrive at pages via a link, then (all things being equal) the more pages that link to a certain page, the more clicks from different locations that it can generate.

Different stages of the Web also saw vastly different demographics between Web page users and Web page authors. Given the Web's academic origin, most early authors were scientists, as were most users. But as excitement for the Web spread, and being that it is far easier to be a user than an author, there was a brief period in time in which users and authors were very different groups of people.

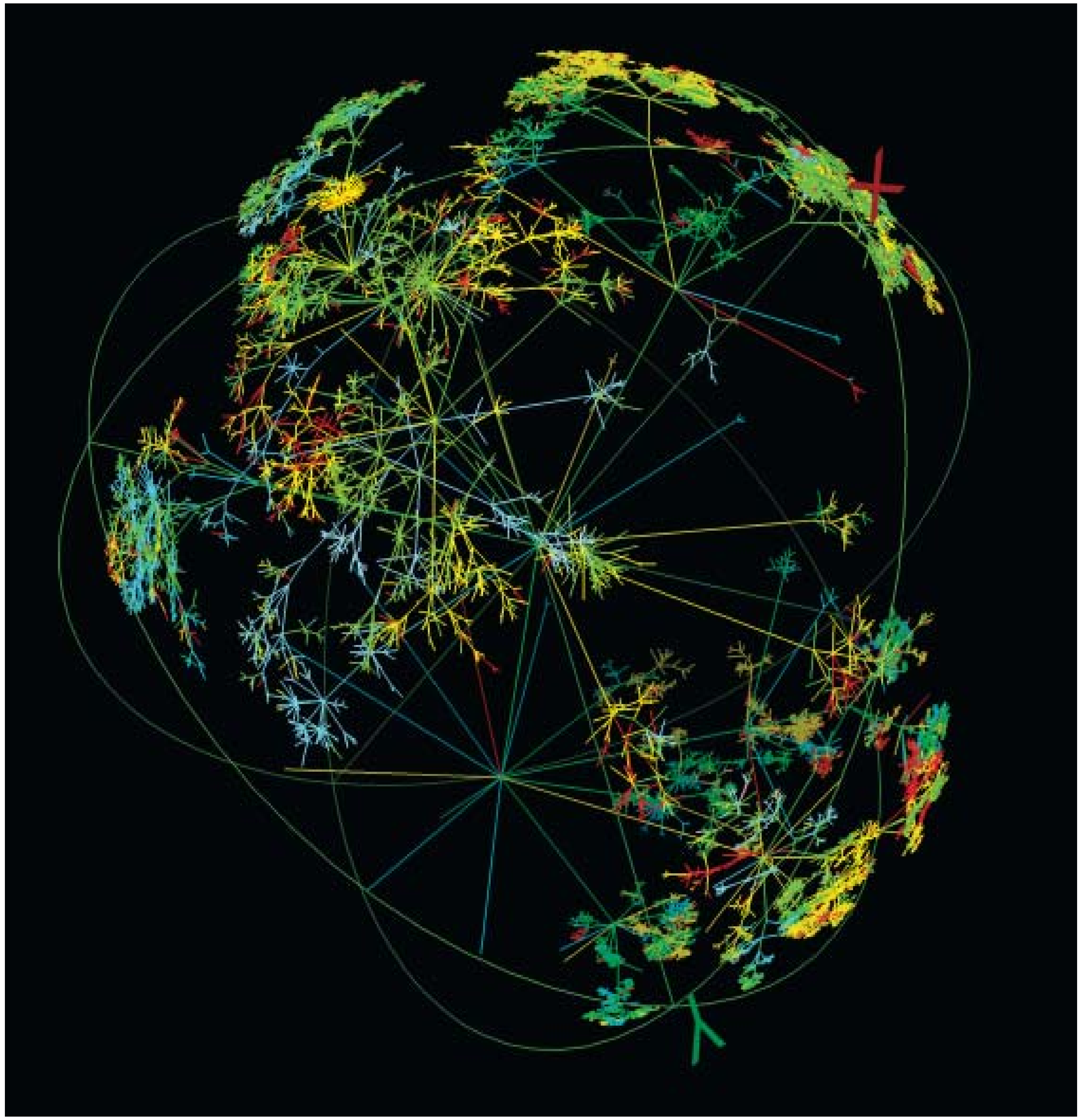
Over time, as Web-authoring tools became readily available and as Web resources became easier to attain, these two demographics gradually merged. Thus, in the current state of the world, many Web users are also Web authors. We will explore this fact more closely later when we discuss the phenomenon of Web loggers. However, for now, just consider the fact that when authors and users come from similar pools of people, a new relationship emerges:

*The greater a page's traffic, the greater its connectedness.*

This happens simply because people tend to link to pages that they themselves value.

Still in the dark ages of the Web, there suddenly emerged a new tool: the search engine. Now ubiquitous, the first general purpose search engine, AltaVista, represented a revolution in usability on the Web. Suddenly, pages could be found by *content* and not just by *location*. Instead of knowing where some piece of information was located on the Web, one could find it by supplying a rough sketch (say a few keywords) to describe the desired document. While there are many benefits to retrieving information in this reversed manner, there is an unfortunate side affect: a single query can have thousands or even millions of valid results. Worse yet, some results, while technically a valid match to a query, may actually be off topic to the intent of a user's query. For these cases, the 'right' result may be buried deep within a pile of 'wrong' results.

Search engines – back then and still today – try to



ease the burden on the user by ordering search results so that the high quality pages that are likely to satisfy a user's intent show up first. But the process of ranking results is both art and science and still far from perfect. In any event, with the emergence of search engines came a new relationship:

*The greater a page's rank, the greater its traffic.*

This new relationship holds simply because a search engine can introduce users to pages that they never knew about. Moreover, in the case where the user is an author as well, we also find the corollary:

*The greater a page's rank, the greater its connectedness.*

Hence, the programs behind the search engines have an impact on the traffic patterns of users and the linking patterns of pages (as search engines influence authors).

Over time, new search engines would come and go, offering different features with the goal of earning a dedicated user base. But the sticky feature – a feature that entices users to be repeat users – is a better ranking function, one that seems to anticipate the user intentions, and satisfies the user needs with relevant results better than the competition.

Two interesting breakthroughs in the search engine industry used an implicit form of intelligence embedded within the Web: traffic and connectedness. In the aggregate, traffic patterns on the Web reflect what users find valuable, while patterns in connectedness reflect what authors find valuable. Both represent something akin to a voting scheme for ordering pages by value. In the late 1990s each of these ideas were exploited by two new search engines, DirectHit and Google, which were able to use traffic and connectedness (respectively) to more effectively rank pages.

Today, virtually every major search engine uses traffic and connectedness as an ingredient to their ranking function, but at the time of their introduction, DirectHit and Google each represented another major step in search engine technology by using the collective wisdom of the Web to better satisfy users. However, these two innovations closed the loop, so to speak, on how people, programs, and pages influence one another:

*The greater a page's traffic, the greater its rank.*

*The greater a page's connectedness, the greater its rank.*

With these final two relationships, people, programs, and pages each have the ability to influence one another. We have seen throughout this book how circular relationships (i.e. positive feedback loops) are key to the creation of fractals and chaos, and so it is on the Web. Besides the benefits seen from an evolving Web, we can also see instances of spontaneous weirdness that are all a direct consequence of the Web's recursion:

A single link from an influential Web site can cause the linked Web site to collapse, due to a spontaneous increase in traffic. For example, the Web site Slashdot, http://slashdot.org/, is a daily compendium of links to interesting developments in technology, submitted by a vast and sometimes fanatical user base, and vetted by editors. When Slashdot adds a new link to an interesting Web page, the ensuing stampede of readers clicking on the link can bring an unprepared Web site to its knees under the weight of all its new audience. This phenomenon has been called the Slashdot effect (even if the originating site is not Slashdot itself), and affected sites are said to be slashdotted.

*We have seen throughout this book how circular relationships (i.e. positive feedback loops) are key to the creation of fractals and chaos, and so it is on the Web.*

Communities of Web loggers have colluded to form google bombs. By collectively linking to a page in an atypical manner, small groups of individuals have successfully tricked search engines into producing humorous results. For example, a search on 'more evil than evil itself'used to return Microsoft's Web site as the top-ranked result. This was accomplished by a loosely coordinated group of Web authors creating links to Microsoft, where the underlined text in the link (the so-called anchor text) said 'more evil than evil itself'. A similar phenomenon is known as link spam where individuals attempt to influence search engines to favor pages of their choosing.

All of these cases are a consequence of the Web reflecting an intricate coupling between people, programs, and pages. Throughout the rest of the chapter we will see how the Web's recursion yields a surprising degree of self-organization, self-regulation, and self-similarity on multiple levels.

## The Macroscopic Web

Assigning superlatives to the Web is easy: it's massive, it's dynamic, it's decentralized – it's unlike anything else in the world. But one of the Web's most amazing attributes is that it is arguably the largest self-organized artifact in existence. Every day millions of Web publishers add, delete, move, and change their pages and links, yet what results is far from random or haphazard. Rather, from these millions of uncoordinated decisions emerge a startling number of regularities. Figures 6.1 and 6.2 display two visualizations of the Internet's map, its complex flowering and branching structures tantalizingly fractal-like. Scientists have quantified that intuition, uncovering self-organizing fractal patterns in examining nearly every aspect of the Web, including the contents of pages, the hyperlinks between pages [Barabási 1999], the physical wires making up the Internet [Faloutsos 1999], the types of files found on the Web [Crovella 1998], the traffic patterns on the Internet [Leland 1993] [Crovella 1996], and the behavior of people as they surf the Web [Huberman 1998].

Consider traffic patterns. If you were to tap a particular wire on the Internet and listen as emails, Web page contents, and other data zipped back and forth, you would observe erratic rises and falls in the volume of traffic, marked with occasional bursts. Figure 6.3a shows a representative sample of traffic volume over the course of 100,000 seconds, or a little more than a day: you can see somewhat noisy fluctuations punctuated with large bursts. Figure 6.3b zooms in on a particular 10,000-second sub-period (about three hours) within the full series. The pattern of fluctuations and bursts looks roughly the same. Similarly, in Figures 6.3c through 6.3e, as we zoom in to shorter and shorter time scales, the same degree of fluctuations and bursts seems evident. The distribution of traffic is neither smoothing out nor getting choppier as we zoom in further and further. Here we have the classic appearance of self-similarity. We observe the same statistical behavior regardless of the resolution (time scale) of our plot. Scientific studies confirm mathematically what our eye suspects: statistical measurements of the variability of traffic on the Internet and on corporate networks do not differ substantially whether we are examining patterns across a month, a day, an hour, or a few seconds [Leland 1993] [Crovella 1996].

Why is Internet traffic self-similar? The answer is surprisingly simple. A particular wire on the Internet will carry a variety of data traffic, including email, Web pages, images, music, videos, and network control information. Each piece of data requires a different amount of information to encode: a single email usually requires little information, while a video clip of a movie trailer requires much more information. While the vast majority of pieces of data travelling around the Internet are quite small, a few pieces of data are many many times larger than average. The occasional video file or dissertation-length email punctuates a steadier stream of comparatively miniscule Web pages, emails, etc. This skewed distribution in the sizes of pieces of data is called a *power law* distribution or a *heavy-tailed* distribution, for reasons we will explain shortly. It turns

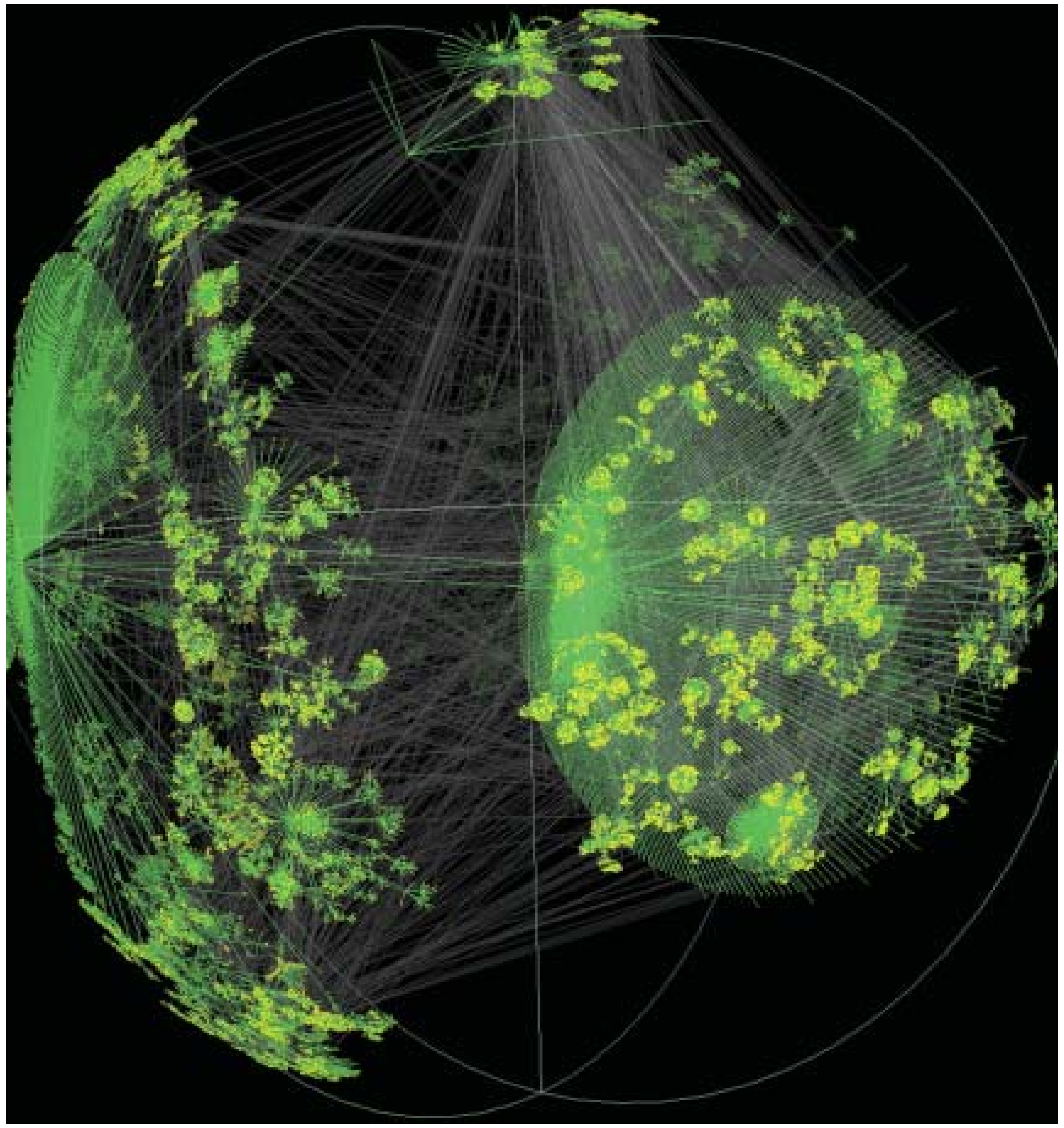


Fig. 6. 2 above: A second visualization of part of the Internet's topology, generated after a day of probing the Internet from a single source. The topological structure is rendered inside a sphere using hyperbolic geometry, which yields a fisheye-like display.

Data and visualization are courtesy scientists at the Cooperative Association for Internet Data Analysis (CAIDA), one of the leading academic centers for measuring, understanding, and supporting the Internet infrastructure, http://www.caida.org. Copyright © 2003 The Regents of the University of California. All Rights Reserved. Reprinted by permission.

out that, when the sizes of pieces of data in a stream of traffic are governed by a power law, that stream will be self-similar. That's all it takes for self-similarity to arise. The consistency of the series of plots in Figures 6.3a through 6.3e is a direct result of the fact that data traversing across the Internet is mainly a river of small and moderate bits of data littered intermittently with relatively monstrous chunks.

In more detail, a power law states that, within a set of items, items of size *x* are a constant factor (say, two times) *more* frequent than items of size 2**x*. In turn, items of size 2**x* are twice as frequent as items of size 4**x*. Mathematically, the frequency of an item of size x is proportional to $x^{-\beta}$, where β is a constant. For example, suppose β=1. Then the frequency of an item of size 2 is $2^{-1}$ or $^1/_2$, while the frequency of an item of size 4 is $4^{-1}$ or $^1/_4$. The larger the item, the less frequently it occurs, in direct proportion to its size. The distribution is called a power law because of the constant power β used in the formula for frequency.

The distribution is said to be heavy-tailed because the *tail* of the distribution (the right-hand side of the distribution when plotted, or the part describing large values of *x*) actually contains a much larger proportion of items than would be predicted by the standard *bell-shaped* distribution (a.k.a. the Normal or Gaussian distribution) often used in statistics.

That is, as we move to the far right of a bell-shaped distribution – well past the center of the bell – the frequency of items approaches zero extremely quickly, much more quickly than in a power law distribution.

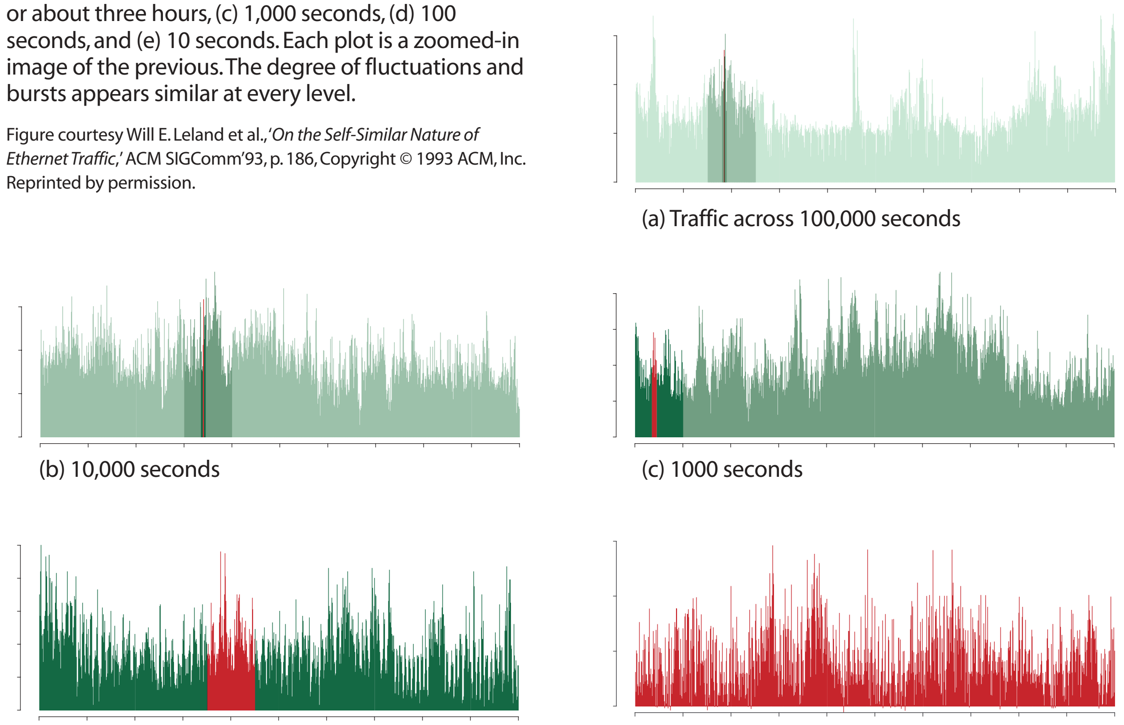
The power law is a fundamental indicator of fractal-ness [Schroeder 1995]. A power law is such that, no matter how much we zoom in or out, it looks the same. It doesn't matter if we draw a plot of the distribution over a huge range of sizes, say ranging from 1 to 100,000, or over a smaller range of sizes, say between 10 and 100, the shape of the distribution will be the same. Figure 6.4 illustrates the self-similar nature of the power law.

## Power Laws and the Log-Log Plot

The best way to understand the power law is by example. In Figure 6.5, we show a series of plots, all displaying the same information in different ways. All of the plots convey information about the number of inbound links to each of 100,000 randomly chosen Web pages. Each point on the graph can be read as follows: the point's x-value is a particular number of inbound links, while the point's y-value is the number of Web pages (among the

Fig. 6.3 below: Self-similarity of Internet traffic. Fluctuations and bursts in traffic over a period of (a) 100,000 seconds, or about one day (b) 10,000 seconds, or about three hours, (c) 1,000 seconds, (d) 100 seconds, and (e) 10 seconds. Each plot is a zoomed-in image of the previous. The degree of fluctuations and bursts appears similar at every level.

Figure courtesy Will E. Leland et al., '*On the Self-Similar Nature of Ethernet Traffic,*' ACM SIGComm'93, p. 186, Copyright © 1993 ACM, Inc. Reprinted by permission.



(a) Traffic across 100,000 seconds

(b) 10,000 seconds

(c) 1000 seconds

(d) 100 seconds

(e) 10 seconds

## Visualizing the Net

Creating a visual depiction of the Internet is no easy task. The difficulty is not only a matter of the Internet's size. Because the Internet is composed of independent computers distributed around the globe, no one person can hope to compile a specification of all the computers and connections involved. Visualization is also hampered by the fact that the overlapping connections in the Internet–and similarly the hyperlinks among Web pages–are impossible to flatten into two-dimensional or three-dimensional images suitable for human consumption.

Scientists have long examined the problem of visualizing high-dimensional data in two or three dimensions. Throughout this chapter, we report summary characterizations of statistical measures of the Internet that we can show using traditional two-dimensional plots. Figures 6.1 and 6.2 represent more direct attempts at capturing the structure of the Internet in images, using a variety of visualization techniques. The layout algorithm used for Figure 1a can take almost a day of computing time to optimize visual space. The method used for Figure 6.1b was developed by Young Hyun, and was based on the pioneering visualization techniques of Tamara Munzer. By plotting points within a three-dimensional sphere, the image is more comprehensible for viewers and allows a natural interactive mode where different points can be 'dragged' into the center of the sphere for closer inspection of that point and its neighborhood .(http://www.caida.org/tools/visualization/walrus/)

A number of other scientific efforts have focused on depicting the intricacies of the Internet using visual means. Many are cataloged in The Atlas of Cyberspace [Dodge 2002]. (http://www.cybergeography.org/atlas/) Ben Fry of the MIT Media Lab has created a real-time animation of Web traffic, growing and squirming like an anemone in immediate response to browsing behaviour across an MIT web site (http://acg.media.mit.edu/people/fry/anemone/). Beyond mapping, several teams have explored methods for presenting Web search results graphically, though none has yet supplanted today's standard text-based lists.

To many people, the inner workings of the Internet are a mystery: how do computers everywhere interact so that email and Web contents zip to and from the right places at the right times? An informative and entertaining computer-animated movie called The Warriors of the Net (http://www.warriorsofthe.net/) explains the Internet's mechanics by portraying its components (bits, wires, packets, routers, firewalls, etc.) as robotic creatures in a stark factory of the future.

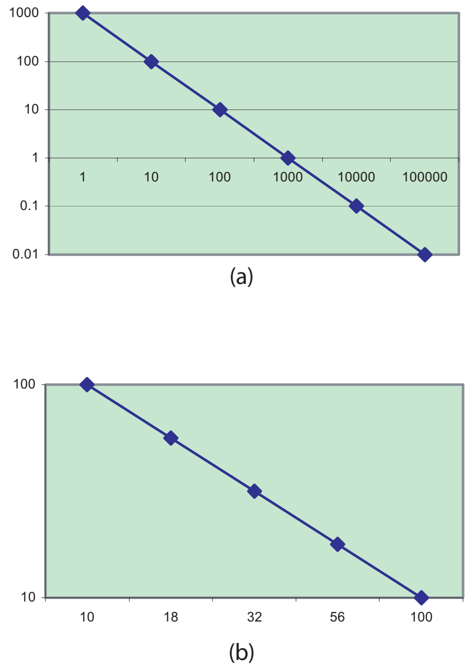


Fig. 6.4 above: Self-similarity of the power law distribution. Both plots show the same power law distribution with parameter *b*=1, so that frequency equals *x*-1. The top graph displays a large region from 1 to 100,000; the bottom graph displays a smaller region from 10 to 100. No matter what region is plotted at what resolution, the distribution will always appear as straight line (of the same slope) on a log-log plot.

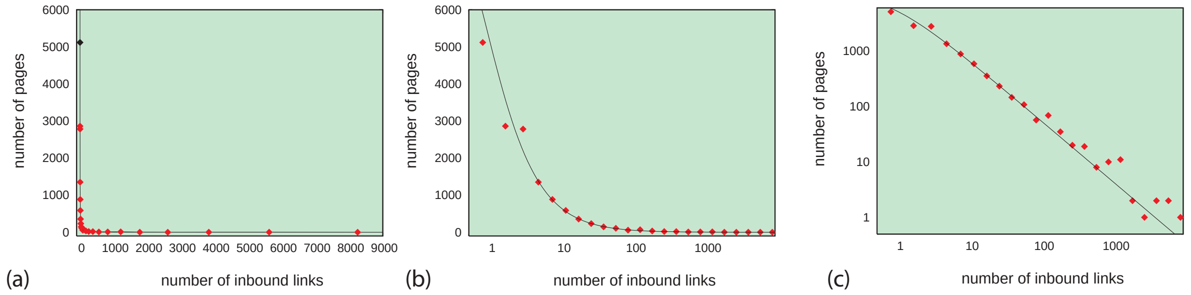


Fig. 6. 5: Different ways to visualize a power law distribution. All three graphs display the same data: a histogram of the number of Web pages (among a random subset of 100,000 pages) that have a specific number of inbound links pointing to them. (a) Shown with both axes on a linear scale; (b) The horizontal axis with a logarithimic scale; (c) both axes on a logarithmic scale.

100,000) that have the specified number of inbound links pointing to them. This type of plot, which displays the number of items that appear within specified ranges on the x-axis, is called a *distribution* or a *histogram*. Figure 6.5a shows the distribution with ordinary linear scales on each axis. The plot is an almost perfect L shape, revealing the extremely skewed distribution of links on the Web. Almost all Web pages have a very small number of inbound links, as seen by the points lying on the vertical portion of the L shape. On the other hand, a tiny handful of Web pages have a hugely disproportionate number of inbound links, as seen by the few points on the far right of the horizontal piece of the L.

Figure 6.5a is hard to read, since all the points are squashed onto the vertical and horizontal pieces of the L. Figure 6.5b displays *exactly the same information*: the only difference is that the x-axis is plotted on a log scale, where the distance between the x-values of one and ten is given as much visual space as the distance between ten and one hundred and the distance between one hundred and one thousand. The log scale stretches out the data points, making it easier to see the detail of the distribution.

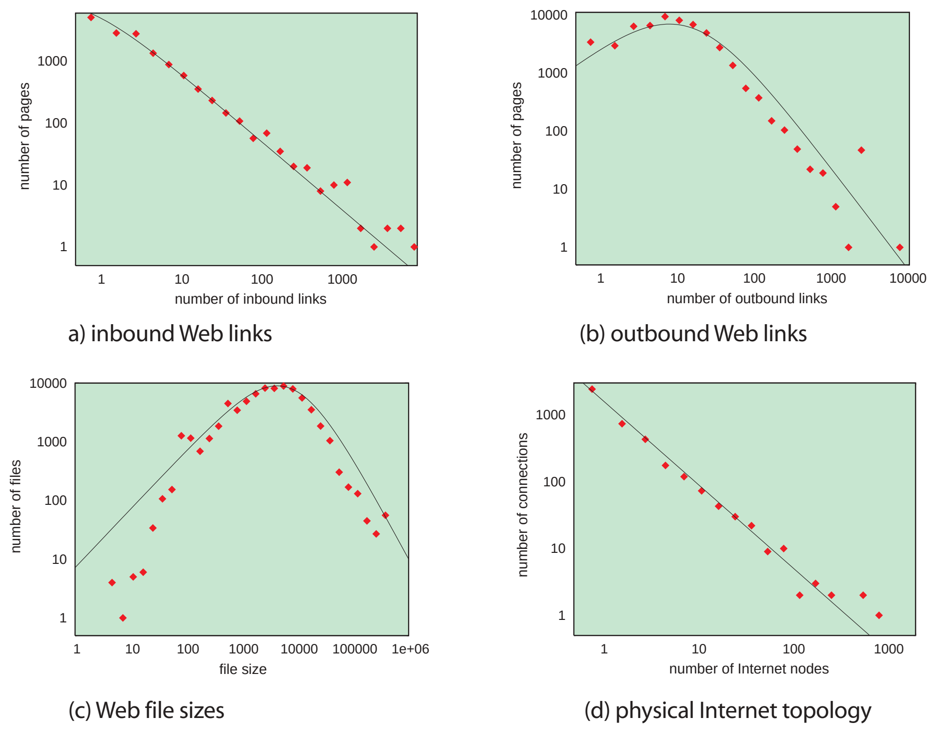
Figure 6.5c plots the same information using log scales on *both* the *x* and *y* axes. This is the so-called *log-log* plot. Notice that, once the data is drawn on a log-log plot, a striking regularity emerges that would be impossible to see using the linear scales of Figure 6.5a: the points follow an almost perfectly straight line. When a distribution drawn on a log-log plot follows a straight line, it is a power law distribution.

Power laws arise naturally. The amount of wealth spread among people follows a power law. The number of people spread across cities follows a power law. The number of connections in the metabolic network of a microorganism, the number of citations to academic papers, the number of connections in the electricity power grid, and the number of people seeing a particular movie are but a few of thousands of examples of naturally-occurring power laws.

Power laws also abound on the Web. As mentioned, the sizes of data pieces as they flow across the Internet are distributed according to a power law. The sizes of files themselves, residing on Web servers on the Internet, obey a power law. The number of queries submitted to search engines, the frequency of word usage on pages around the Web, the number of hyperlinks pointing to and from Web pages, the depth to which Web users surf, and the number of physical wires connecting to Internet hubs all follow power laws.

Let's examine more closely the pattern and formation of links on the Web. Figure 6.6a shows the distribution of inbound links on the Web plotted on a log-log plot. Notice that on a log-log plot a power law distribution appears as a straight line. We see that the distribution of inbound links on the Web is close to a pure power law, except for a very slight drop-off from a straight line at the top left of Figure 6.6a (the region of small values of x, or small numbers of inbound links).

Fig. 6. 6 right: A tour of the power-law Web. Distributions capturing nearly all aspects of the Web follow a power law, including (a) inbound links, (b) outbound links, (c) files sizes, and (d) the physical Internet itself (the wires connecting computers around the world). Power laws crop up elsewhere too, including people's behaviour as they surf the Web, and even the level of interest among advertisers to be showcased in conjunction with particular search queries.



a) inbound Web links

(b) outbound Web links

(c) Web file sizes

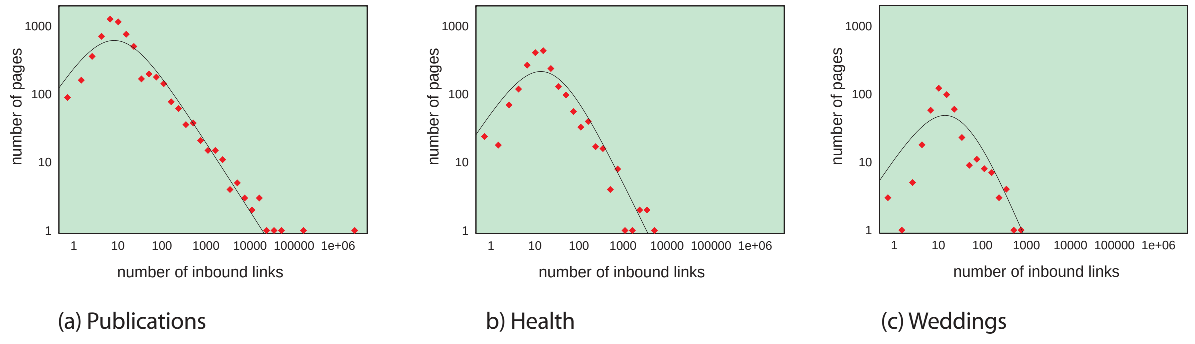
(d) physical Internet topology

Compare Figure 6.6a with Figure 6.6b. The latter shows the distribution of *outbound* links on the Web (links emanating *from* Web pages) instead of inbound links. Near the right end of the graph, the distribution looks very much the same as the inbound link graph: a straight line on a log-log plot. But on the left side, in the range of smaller values of *x*, the distribution deviates fairly strongly from the linear signature of a power law. There is a bump in the distribution of outbound links not seen in the graph of inbound links. It turns out that bumps like these are the rule rather than the exception (in this sense, the near perfectly straight line of Web inbound links is rare). For example, the graph of the distribution of file sizes pictured in Figure 6.6c has an even more pronounced bump before straightening out on the far right. Many of the power laws observed in nature are also marked with significant deviations in the region of small values of x.

Figures 6.7a, b, and c show inbound link distributions for specific e-commerce segments of the Web, comparing the communities of online booksellers, commercial health-related sites, and online wedding retailers, respectively. Here we see more examples of the modified power law: in each case, the plot displays a significant bump on the left side before converging toward the linear power law on the right-hand portion of the graph.

In the section that follows on the microscopic web, we will examine what low-level forces are at work in generating both the pure power law seen for inbound links and the modified 'bumpy' power law more common in other distributions. For now, simply note that the closer a community's distribution is to a linear power law, the more cutthroat the competition is to get noticed within that community, and the harder it is for new entrants to compete with the well-established players. The larger the bump on the left edge of the graph (the larger the divergence from a pure power law), the more egalitarian is the community, and the easier it is for new sites to rise to (or near) the top. From analyzing the data underlying Figures 6.7a, b, and c, one can infer that booksellers – led by Amazon.com with millions of inbound links – are extremely competitive, while wedding retailers are less so. Commercial health sites lie somewhere in between. Similarly, online sites for corporations and the entertainment industry are highly competitive, while Web sites for scientists, universities, and photographers are not.

There are multiple factors that can lead to the differences in competition that we see. For commercial wedding sites, one factor could be their local nature: many wedding-related retailers serve only a local area, and those serving different areas usually do not compete. Another factor may be that people looking for wedding services use methods other than the Web more often (e.g., referrals from friends). Perhaps because people use wedding providers rarely, they are less likely to create and share information among related sites on the Web.



(a) Publications

b) Health

(c) Weddings

Fig. 6.7 above: Inbound link distributions for specific e-commerce segments of the Web. (a) Online booksellers, (b) commercial health-related sites, and (c) online wedding retailers. Each plot shows increasing divergence from a pure power law, indicating decreasing competitiveness within that community.

Note that more difficulty competing with existing popular sites does not mean that substantially better newcomers cannot become popular quickly. For example, Google (a relative latecomer to the search business) has captured a huge fraction of the Web search business largely by providing better service and spreading through word of mouth.

### The Web is a Bow Tie

In 2000, a collaboration of scientists from AltaVista, IBM, and Compaq [Broder 2000] discovered a fascinating property of the Web: somehow, all of the billions of pages and links have organized themselves into an overall *bow tie* shape as pictured in Figure 6.8. The center of the bow tie is a core of *strongly connected* pages: every one of these pages can be reached from any other page within the core by clicking on a sequence of links (the sequence may need to traverse a number of intermediate pages, but some path exists between the two core pages). The left bow is connected to the core, but only through *outgoing links*. That is, there exist links *from* the left bow *to* the core, but not vice versa. Conversely, the right bow is connected from the core only via *inbound* links. One can traverse links *from* the core *to* the right bow, but not back again. Finally, disconnected pages that have no links either to or from the core surround the bow tie. The scientists measured the relative sizes of these four main components of the Web (the core, the left bow, the right bow, and the disconnected pages). To their surprise, all four components were roughly the same size.

## Zipf's Law

The Web is not alone in exhibiting power laws. Data gathered on the use of language, the population of cities, and the distribution of wealth all show clear power-law behavior. The frequency of words used in human language deserves special mention: it follows the famous Zipf's Law, named after George Kingsley Zipf, an early twentieth century scientist who revolutionized our understanding of power laws, and helped to reveal their astonishing prevalence throughout society and nature.

Zipf's Law states that the most common word used in language is a constant factor (say, two times) more common than the second most common word, and the second most common is twice as common as the third, etc. Remarkably, for almost any sizeable source of words you can think of–all New York Times articles, or all the works of Shakespeare, or all textbooks on molecular biology, or the Bible–Zipf's law holds.

In 1955, Herbert Simon sought to unify the observations of Zipf and others by formulating a single common explanatory model for many of the systems displaying power-law behaviour, including language, population, and wealth. Benoît Mandelbrot [1953,1959] proposed a fascinating alternative explanation for Zipf's Law as it relates to language. He showed that the distribution can be understood as the end result of centuries of adaptive maximization of the information content of language.

A year later, some of the same scientists [Dill 2001] showed that the bow tie property is a feature not only of the Web in its entirety, but also of various pieces of the Web. No matter how the Web is sliced – whether by content into topic-specific clusters, by geographic location into regions, or by organizational entity into groups of pages owned by the same person – the bow tie shape emerges, even retaining the rough equality of size among the four main compo-

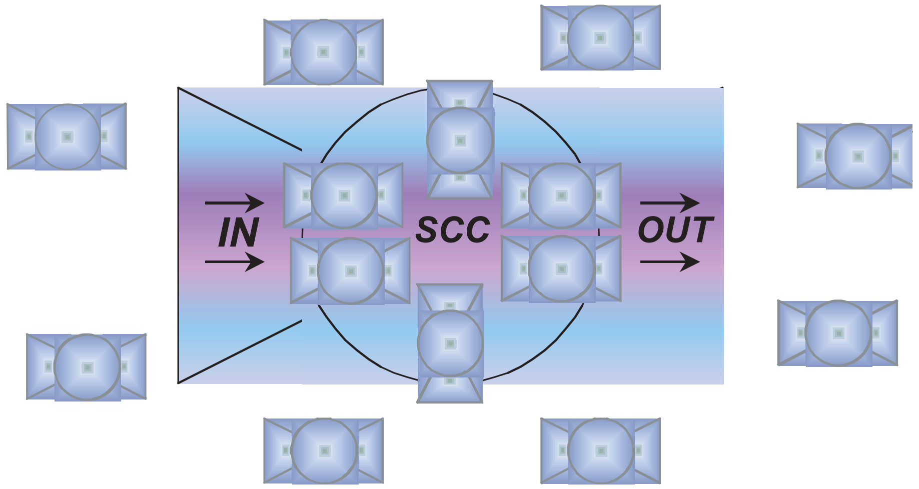


Fig. 6. 8: The Bow tie structure of the Web, consisting of a strongly connected component (SCC), a set of pages that follow into the SCC, a set of pages that pass out of the SCC, and a set of smaller disconnected islands that are themselves SCCs.

nents of the bow tie. This strange structure appears endemic to the Web and pervasive at all levels, revealing a beautiful new type of self-similarity not seen anywhere else.

### Search Engines: Tapping The Ebb and Flow of Ideas

In a way, search engines like Google and Yahoo! have a window into the mind of the masses. Search queries stream in by the second capturing people's thoughts, worries, and whims, whatever they happen to be looking for at that particular time. Web sites like Google's Zeitgeist / (http://www.google.com/press/zeitgeist.html), the Yahoo! Buzz Index (http://buzz.yahoo.com/), and the Lycos 50 (http://50.lycos.com/) report on fads and trends reflected in search traffic: the thoughts and ideas that people are searching for *en masse*, including what is hot and what is passé. It is fascinating to watch as memes appear, skyrocket, cycle, or decay, as the case may be.

In watching the top search terms from one week to the next, clearly some terms will stay perched among the top ten, while others will drop out. For example, as of Sunday August 3, 2003, 'Britney Spears' moved from third to second place, continuing a remarkable run of 123 straight weeks atop the Yahoo! Buzz Index charts. 'Tour de France' also remained in the top ten, though only for the second week running. Meanwhile, 'Beyoncé Knowles' and 'PlayStation 2' fell from their top-ten perch the prior week, supplanted by Kobe Bryant and Angelina Jolie, celebrities whose profiles rose during the week, fuelled by a criminal indictment and a new movie release, respectively. The percent of terms that disappear from the top ten from one week to the next – equivalent to the percent of *new* terms, and reciprocal to the percent of stationary terms – is called the *churn rate*. The churn rate of search terms captures the speed at which new memes rise and old memes fall.

Churn rate can be computed for different numbers of top *N* terms. We can examine the proportion of terms lost from the top ten, or the proportion lost from the top 100, or the proportion lost from the top 50,000 terms. Note also that we can compute churn rate over any time frame: daily, weekly, monthly, etc.

You might hypothesize that churn rates would differ depending on whether you examine the top ten terms, or, say, the top 50,000 terms. For example, it

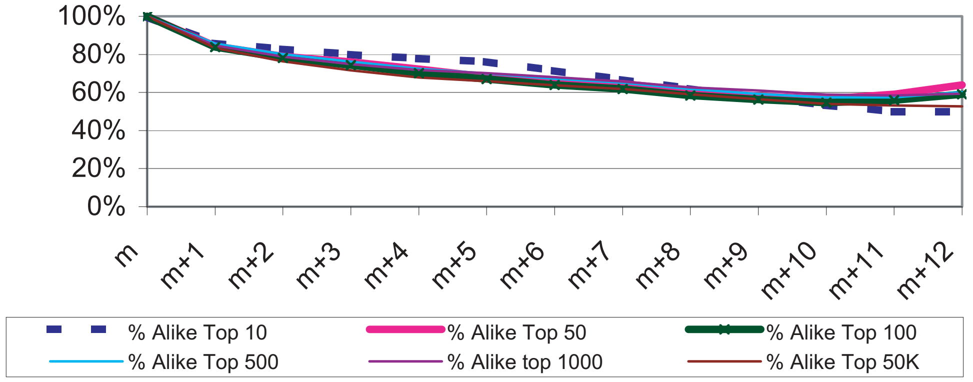


Fig. 6. 9 right: Month-to-month churn rates describing how popular search queries shift over time. Churn rates exhibit self-similarity, remaining the same as the number of terms considered ranges from 10 to 50,000.

seems reasonable that the status of the most popular terms could be so self-reinforcing as to render them more stable than the hordes of terms among the top 50,000. However, this hypothesis is not correct: in reality, the top ten is no more stable than the top 50,000. In fact, no matter what value for *N* is chosen – 10, 50, 100, 500, 1000, or 50,000 – churn rates are unaffected. Figure 6.9 shows churn rates after one month, two months, three months, etc., out to one year. As expected, the longer the time frame, the higher the churn rate, as a greater proportion of terms filter up and down. However, for any given time frame (say, seven months), churn rates are nearly identical for all values of *N*. Here we see a remarkable form of self-similarity: with no matter what granularity we look at search terms – whether we zoom in to examine the top ten, or zoom out to examine the top 50,000 – the percent of terms entering and leaving the identified set remains constant.

## The Middle Web

Having just seen how the Web contains some measure of order at the highest level, we now turn our attention to the next lower level, where groups of authors and users form patterns on the Web. The short version of this story is that the Web's content is effectively self-organized by the actions of individuals. Contrasting this self-organization to the more familiar phenomenon of centralized organization, we will see that the Web exhibits aggregate behaviour that begins to resemble a hive-like intelligence.

### Web Logs a.k.a. Blogs

One of the more recent additions to the Web site bestiary is the Web log or *blog*. Blogs began as something like online diaries with authors making regular postings that were topically focused on everything under the sun or nothing in particular. Journalists and pundits found the medium to be promising new ground for self-publishing. At its best, early blogs allowed for grass-roots journalism and an unbiased flow of ideas and information. At its worst, blogs were simply vanity sites.

The emergence of blogs is important for two reasons. First, blogs, more than any other phenomenon, blurred the line between author and user as most blog content was about the first hand experience of visiting other Web sites. Second, blog

*The essence of the self-organized nature of the Web is that authors – being somewhat independent of one another – can effectively do whatever they want.*

software – the programs that facilitate and automate the maintenance of a blog site – would evolve in sophistication, incorporating many new features including user accounts, discussions, postings by multiple individuals, rating systems (of users and posted stories), multimedia, and search. Today, sophisticated blog software is freely available, and modern blog sites come in many flavors including current event discussions, various grades of self-published journalism, community forums of differing degrees of speciality, and, yet still, the simple diary.

All told, blog sites represent a deliberate effort by individuals to cooperate towards a form of community publishing, with the authors, editors, and readers all coming from a similar pool of individuals. Blog sites also represent larger-scale communities, beyond a single site, because many individuals often contribute to the content of modern blog sites and the membership of related blog sites often overlap. Moreover, the content on one blog site often influences the content on other blogs.

Modern search engines, which use link structure for improving the relevance of served results, have had to co-evolve with the emergence of blogs for multiple reasons. The primary reason is that blogs, by and large, are quirky sites, yet they carry a disproportionate amount of influence in assessing the importance of Web sites because they contain so many links. When a quirky group of people link to pages in an atypical manner, their quirks are propagated to the mainstream if left unchecked.

This amplification property of blogs results in many interesting social phenomena on the Web that has no real-world analogue. Propagation of memes on the Web can start with a single blog site distributing a funny or unusual link. Other blog sites, exhibiting almost a flocking behavior, redistribute the meme, which impacts not only the content that people read but also the links that persist on the Web. In this way, ideas and information (both true, false, and otherwise) can circumnavigate the globe multiple times in a single day, making the circular influence of linking patterns all the more pronounced.

### Shared Taxonomies

Another form of deliberate cooperation by Web authors can be found in shared taxonomies, which is best exemplified by the Open Directory Project (ODP) located at http://dmoz.org/. The ODP consists of a topical taxonomy, not unlike the best-known taxonomy at Yahoo!. However, the ODP is a strictly volunteer effort, where individual editors assume ownership for different topics on the Web. The volunteer editors collect links to pages that are relevant to their particular speciality and incorporate them into their respective location within the taxonomy. All told, the ODP has thousands of editors that maintain links to millions of pages, which, in turn, are incorporated into the ranking algorithms of the most important search engines.

Clearly, the ODP is a distributed effort by individuals to bring order to the Web. However, as with blogs, the ODP represents a deliberate and intentional form of cooperation by individuals. There exists an unintentional form of cooperation by authors that is, perhaps, even more striking than the ODP and blogs because it represents the truest form of self-organization; namely, one in which the individuals cooperating do not even know that they are contributing to something larger.

### Hubs and Authorities

The essence of the self-organized nature of the Web is that authors – being somewhat independent of one another – can effectively do whatever they want. Some will post a flat collection of favorite bookmarks about nothing specific. Others contribute volumes of original material that is focused on a single topic. And still other authors produce nothing more than small collections of links that point to things that are all about the same thing. These last two examples of authors – those that create original material and those that point to focused material – are special in that they form two halves of a single relationship.

Web pages that contain compelling original material (without necessarily the emphasis on having many outgoing links) are often referred to as *authority* Web sites, or more simply as just *authorities*. Authorities have the property that they tend to accumulate incoming links because others interested in their content will create links that point to them. The name, *authority*, comes from the language of bibliographic studies where there is a notion of a work of literature as being authoritative if many authors cite it. As with literature, Web authorities are frequently cited but with links instead of proper citations.

A *hub* Web page (or more simply a *hub*) is the complement to the authority. Hubs are akin to a survey papers or focused reference books in that they contain links that point to many pages that are all about the same topic. Hubs are natural organizers of information because they group similar things together.

Together, hubs and authorities form a recursive relationship that reflects the dependencies between the two types of pages [Kleinberg 1999]. While authorities may earn links by having original content, they may also acquire links by the rich-get-richer process alluded to above (and which we will examine in greater detail in the next section), where highly-linked sites tend to obtain even more links due to their greater visibility. That is to say, hubs may have to link to very popular authorities if they are to retain their status as being a hub. (Not doing so would be like writing a survey article on evolution that fails to cite Darwin.) Similarly, authorities are only truly recognized as being authorities if important hubs link them. Together, these two facts yield a recursive definition for what it means to be a hub or authority.

Hubs are pages that link to authorities.

Authorities are pages that are linked by hubs.

Put simply, these two definitions are recursive because each entity in some sense defines the other. What is truly fascinating about this mutual dependence is that Web pages – in the wild, so to speak – seem to co-evolve via this recursive relationship.

### Community Signatures

In 1999, Ravi Kumar and his colleagues surmised that if the Web is, in fact, composed of many hubs and authorities, then one should be able to find a Web community core by looking for a group of hubs that all point to the same set of authorities. Mathematically speaking, these two groups of pages form what is known as a *bipartite core*. A bipartite structure is illustrated in Figure 6.10, and consists of two types of objects: those in the left set and those in the right set, with every object on the left pointing to every object on the right. Notice that this structure is identical to what you would expect to find if there existed some number of hubs that were all focused on the same collection of authorities.

Kumar et al. found that there were hundreds of thousands of community cores that contained this exact bipartite signature. When inspected by hand, these community cores were almost always focused on an extremely narrow topic such as Japanese elementary schools, Hotels in Costa Rica, or Turkish

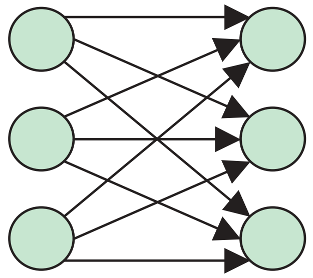


Fig. 6. 10 above: A bipartite core on the Web: every page on the left links to every page on the right

student associations. But most striking, the identified community cores are often so narrow and specific that they are not contained in any taxonomy like the ODP.

Because Web pages contain both regular content and links, there are multiple ways in which two pages can be said to be similar (or dissimilar) to each other. Ignoring text and focusing just on links, one can easily see that hubs within the same community core have outbound links that are similar or identical. Authorities within the same core have inbound links that are similar or identical. Thus, we can speak of two pages as being similar in content (they express similar words and concepts), in outbound links (they point to approximately the same pages), or in inbound links (they are pointed to by inbound links).

One remarkable attribute of the Web is that similarity in inbound or outbound links often implies similarity in page content. This relationship means that one can find new pages of interest by looking only at how pages link to one another within a local neighbourhood of a starting page. The connection between links and content also means that one can analyze link structure to find how topics on the Web relate to one another.

## Self-Organized Communities

The link structure of the Web is not unlike the social network of humans. We have reciprocal relationships with some people, and we know of people that don't know us, which are respectively akin to pages that mutually link to each other and pages where one links to the other only in one direction. Who we are is in some sense defined by the links we have in the human social network. Likewise, Web pages can also be better understood by examining the context in which pages exist within a Web community.

The notion of a Web community core, as defined above, is powerful in the sense that it gives an unambiguous signature from which to identify collections of related Web pages. However, this notion can be considered insufficient because most Web pages will not belong to a Web community core. How, then, can one identify the community in which a page belongs?

There are many different ways to define a Web community, and to be sure, there is no absolute correct definition. Nonetheless, some definitions for a Web community can be used to identify largish collections of pages that, in some sense, seem to belong with one another because they are all focused on a similar theme. We now turn to one particular definition for a Web community that is mathematically rigorous in that it is well defined, is surprisingly intuitive and simple to understand, and empirically corresponds well to real communities on the Web.

For reasons to be explained shortly, we will refer to this type of Web community as a *cut Web community*, or more simply as just a cut community. A cut community consists of a collection of pages that predominately link to one another (with links in either direction). That's the whole definition; it is simple, but yields several elegant properties.

First, note that it is a meta-definition in that it permits one to make more specific statements like 'the bicycle community consist of pages that predominately link to bicycle pages.' Also note that community membership is easy to test for and validate. Hence, if you know about the bulk of the bicycle community, you can look at the links coming in and out of a page in question. If more than half of the links refer back to the bicycle community, then the page in question is also a member of the bicycle community.

In 2000, Gary William Flake and colleagues [Flake 2000] discovered an effective method for identifying self-organized collections of Web pages that obeyed the cut community definition. The method works by recasting the community identification problem into what is known as the *s-t minimum cut network problem*. In this framework, one looks at a collection of pages and links and asks: for two pages, *s* and *t*, what is the smallest number of links that need to be 'cut' (i.e., removed) in order to completely separate *s* and *t* from one another, where *s* is a page that is indicative of the type of community that one is looking for and *t* is an artificial page that represents the whole of the Web. By looking for the smallest number of links to cut, the procedure effectively tries to find the smallest group of pages connected to *s* (our page of interest) that nicely separates from the rest of the Web.

Flake's community algorithm also has the nice attribute that it is computationally efficient. Nonetheless, it is not at all clear that it should even produce collections of pages that are all focused on a single theme. However, in practice, the community algorithm is remarkably successful at finding large collections of related pages. When seeded with the personal home pages of famous scientists, the community algorithm will find hundreds or thousands of pages that are all focused on the specialty of the scientist in question [Flake 2002].

In fact, the community algorithm, and other link-based approaches, have been shown to be very effective in making sense of the Web. Notice the language-independent nature of link-based methods: since they ignore the textual content of pages, they work equally well for pages in English, Spanish, or Swahili, for that matter, or for pages composed nearly entirely of images and multimedia. But here's a secret of the power of the community algorithm and other methods like it: it's not the algorithm that's special, it's the Web.

### Topic Affinity

Consider a completely random Web surfer, who wanders about the Web clicking on randomly chosen links (we will have more to say about the properties and implication of the random surfer model in the next section). The surfer travels from page to page, each time moving forward by clicking on a random link on the current page. Assuming that the surfer starts at a random page, we can measure the relative bond between content and links by measuring how long it takes for the random surfer to visit pages that drift away from the topic of the starting page.

Soumen Chakrabarti and his colleagues [Chakrabarti 2002] found that on the whole, the correspondence between the topicality of a page, and the links that it contains is remarkably strong. In the example of our random surfer, Chakrabarti et al. found that for some subjects, a random surfer could remain on topic after following as many as 5 or 10 links. Interestingly, the degree of topic drift was strongly dependent on the starting topic. For example, 'soccer' pages would drift off-topic relatively fast, while 'photography' pages maintain topical focus for many more steps.

Related to all of this is the role of anchor text to content. Anchor text is the text that is contained in a link (usually underlined in most browsers). The author of a page that contains a link creates the anchor text, but anchor text is usually intended to be descriptive of the page that the link points to, not

### Small World Networks

Many people are familiar with the expression 'six degrees of separation' which suggests that for any two people in the world, there are at most six person-to-person relationships that separate those two people. Thus, you and I may not have any friends in common, but we will probably have a friend-of-a-friend-of-a-friend in common.

The remarkable feature of small world networks is that they contain few links relative to their number of members. Intuitively, small world networks have this dual property by having many members with mostly 'local' relationships (say, most of your friends and neighbours), and a very small number of members that have 'global' relationships (e.g., a celebrity that is known or knows thousands of people). Thus, the path that joins any two random people is likely to begin and end with some local relationships, but will pass through some global relationships in the middle.

Throughout this chapter, we have seen how the Web reflects the properties of our society – and so it does with the small world nature of human culture [Watts 1998]. Between any two Web pages in the Web's largest strongly-connected core, there are at most a few dozen links that connect those pages. E-mail and instant messaging relationships also form small world networks.

The good news about small world networks is that for those who know how to pick links to follow, a small number of clicks will lead one to a desirable location. The bad news is that it is also remarkably easy to spread problems (like viruses and misinformation) in a small world network as well.

necessarily the page that contains the link.

Computer scientists have long been working on the problem of how to recognize when a document is about a particular topic by analyzing a document's text. In this vein, scientists have used these tools to improve search engines and related technologies. Interestingly, many scientists studying the Web have found that the anchor text that points to a page is often a stronger indicator of the referred page's subject than its own text. This is truly a surprising result because it means that the links that point to a page are often a better descriptor of a page's content than its own title.

All of this goes to show that despite being decentralized, the Web seems to 'like' order instead of disorder. Authors don't have to link to pages that relate to their own; but they do. And authors don't have to use anchor text that is strongly relevant to the pages that it refers to; but they do that too. The bottom line is that links, instead of being unruly, are, in fact, self-organizing. In the aggregate, connections and content go hand-in-hand and they co-contribute to the Web's higher-level formation of patterns and structure.

Having seen the self-similarity evident in the top-level view of the Web, and the self-organizing niches and structures evident in the middle Web, we now turn to the underlying low-level processes and forces driving organization and structure on the Web.

## The Microscopic Web

The Web in its most fine-grained detail is the results of billions of individual decisions taken every day around the globe. CNN adds a breaking story; a job hunter updates her online résumé; a university department deletes the homepage of a graduated student; Amazon.com adds a new book title. All around the world, the content of the Web is modified in response to significant real-world events as well as trivial whims. Clearly, to understand the evolution of the Web, we must understand how individual pages are created and modified. However, it is simply impossible to factor into our understanding the details of the innumerable human motivations behind all pages in the Web.

Instead of focusing on minutiae, we need to abstract away as many inessential details as possible and look instead for the simplest rules that capture the most important aspects of the Web's behaviour. This modelling process will help us identify what are the essential ingredients responsible for the self-similar structures on the Web. But don't be fooled by the simplicity of the models that we will talk about. Despite their simplicity, they capture a considerable amount of the complexity that we've observed so far. As Ian Stewart eloquently states elsewhere in this book, simple explanations for complex observations lie at the heart of modelling nature, and fractals are a powerful tool in the mathematician's arsenal for doing so.

### Modelling Web Growth

As in the previous section, let's temporarily ignore Web page content and focus on the links that they contain. Moreover, let's also ignore the direction of all links and just focus on the fact that two pages can be linked to one another. To better understand how the Web evolves, we need to understand how individual pages contribute to the overall link structure. We will model the Web's evolution by iterating over the following five steps:

1. Create a new page, called $p$.
2. Randomly pick an existing link, $l$, not connected to $p$.
3. Randomly select one of $l$'s two adjacent pages, $q$.
4. Add a new link between $p$ to $q$.
5. Repeat steps 2-4 a total of $k$ times.

One pass through these steps adds one new page to the Web and $k$ new links. Obviously, one can repeat the entire process multiple times to add many new pages and even more links.

The recipe above specifies what is known as a *generative model* because it explicitly shows how one takes a current snapshot of the Web, and generates a successor to it that has grown a little bit. The model – introduced for the Web by Albert-László Barabási and Reka Albert [Barabási 1999] – is simple enough that with sufficient mathematical tools, one can effectively see how it would behave if iterated for an infinite number of steps.

The most interesting part of the recipe for growing the Web is in steps 2 and 3, where we pick a random page, $q$, in which to connect to our new page, $p$. If there were $n$ existing pages, and we were to select one of them purely at random, then we would find that each page has a $^{1}/_{n}$ chance of being selected. But that's not we are doing. Notice that we are picking a link, and then picking one of the two pages adjacent to it. This means that the more connected a page is, the more likely it is to be selected in step 2.

The selection process in step 2 can be reasoned as follows. Think of each link as owning two lottery tickets, and giving away one ticket each to the two pages connected to that link. Now you can verify that the probability that an existing page is selected in step 2 is equal exactly to the number of lottery tickets it has, divided by the total number of lottery tickets possessed by all pages. Thus, the more links (or lottery tickets) a page has, the more likely it is to 'win' by being selected in step 2. If a page has many links, it's bound to get more. But if a page has few links, it will probably not get many more. As a result of these facts, this pattern is often referred to as a 'rich get richer' phenomenon or as "preferential linking".

Clearly, Web page authors don't add links to their pages in precisely this manner. But, as we have seen, more links to a page implies many things, including more traffic, higher ranking, and more visibility, in general. Thus, it is not too outrageous to simplify things and just simply say that authors prefer to link to pages that are more connected.

While some may debate the fairness or desirability of such a state of affairs, the 'rich get richer' process is a common one that arises naturally in a large number of domains, including many social, biological, and physical systems ranging from the power grid network to the metabolic networks of microorganisms.

In an influential article published in the journal *Science* in 1999, Barabási and Albert revealed a fascinating discovery: their simple generative model for Web growth is sufficient to replicate many of the key features of the Web. Most notably, the structures generated using this simple model exhibit precisely that same power law distribution as observed on the real Web, and as we witnessed earlier in the section on the macroscopic Web.

## Power Laws and Communities

Barabási and Albert's first Web model touched off a wave of research aimed at capturing additional aspects of the real Web. While Barabási and Albert's model succeeded in capturing some of the highest-level properties of the Web (as well as showing that the Web is unambiguously fractal in construction), it was somewhat incomplete in that it did not account for how the Web operates when viewed on intermediate scales.

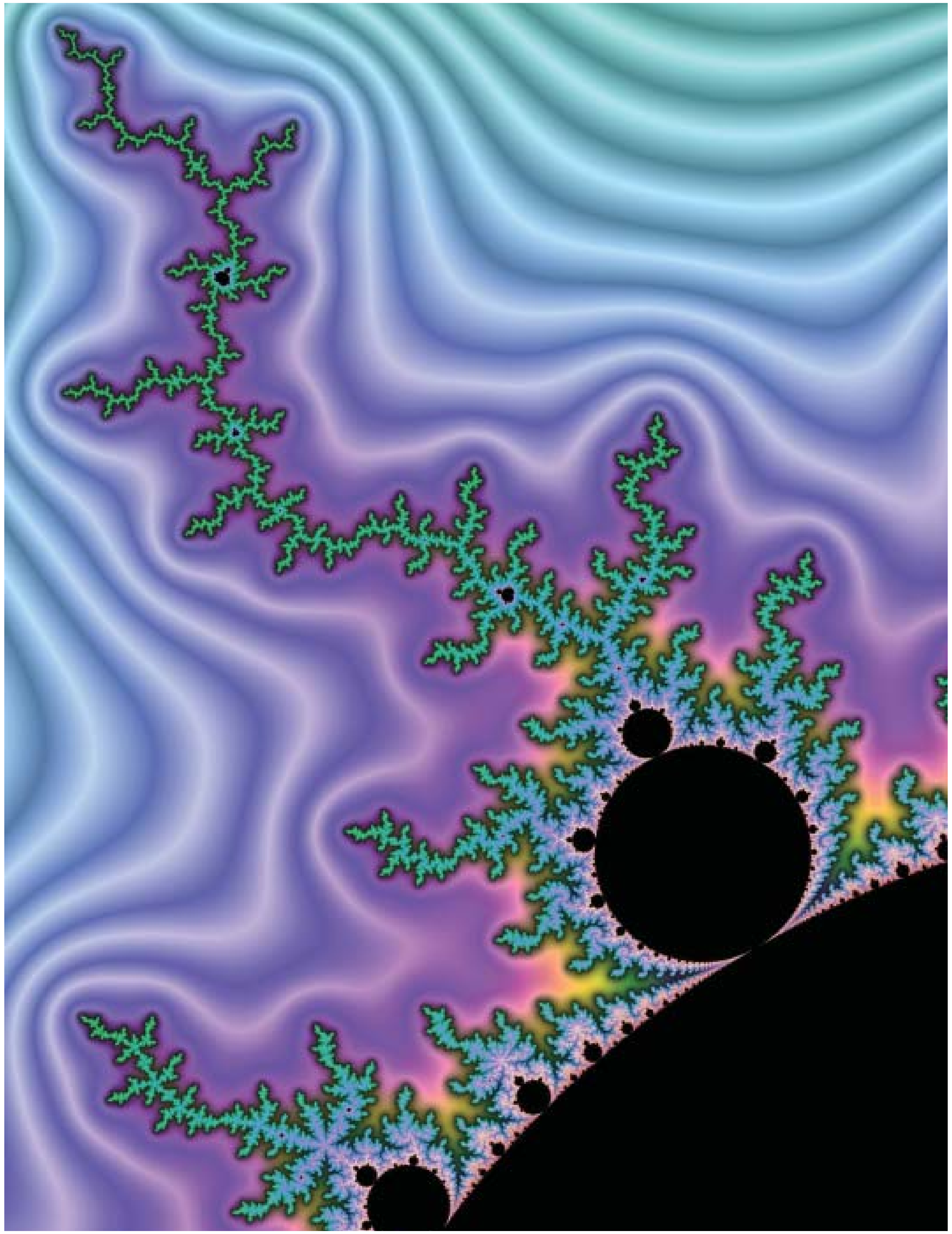
David Pennock and his colleagues [Pennock 2002] made a simple modification to the Barabási-Albert model that would account for some of the behaviours of Web communities. Before we get into the details, let's recap some of the intuition behind power laws and how they occur in nature.

Within the biosphere, we see far more small creatures than we do larger creatures: there are many more bacteria than there are insects; there are many more insects than medium sized animals; and there are still far fewer large animals, such as whales and elephants, than just about anything else. The distribution of sizes of creatures across all species is a power law.

On the other hand, within a species, we see a different pattern entirely. The size, weight, and height distributions of humans follow the more familiar bell-shaped or *Gaussian* distribution. This means that most individuals fall somewhere in the middle – that is, there are more average sized people than small people or large people. The trend of having more average individuals than big or small is found in just about all species, when a species is examined in isolation.

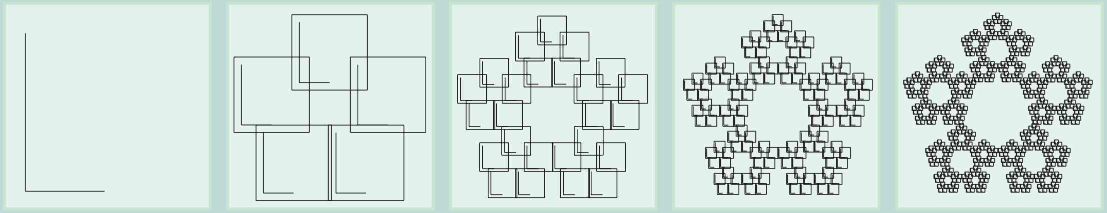
Returning to the Web, and thinking about Web communities and inbound links as being somewhat analogous to species and the size of animals, we find that if one looks at the number of inbound links to a Web page – but restricted to pages in the same community – the distribution is neither a strict power-law, nor is it a Gaussian. Instead, it is bump-shaped (like a Gaussian distribution) but on a logarithmic scale (like a power-law), as we saw in Figure 6.7. The important point in all of this is that at the intermediate level, there is something different going on than the strict rich-get-richer linking patterns that the Barabási-Albert model suggests.

As a people, we all know of celebrities such as famous actors, athletes, and politicians. But we also know many people based on our interests, where we live, and where we work. Likewise, Web authors create links not just to popular pages, but also to pages that are related to their own page in some manner. These non-popular links are akin to the people that we personally know, while the popular links (say to Yahoo!) simply represent an awareness of what the masses link to or know of in the aggregate (like a celebrity).

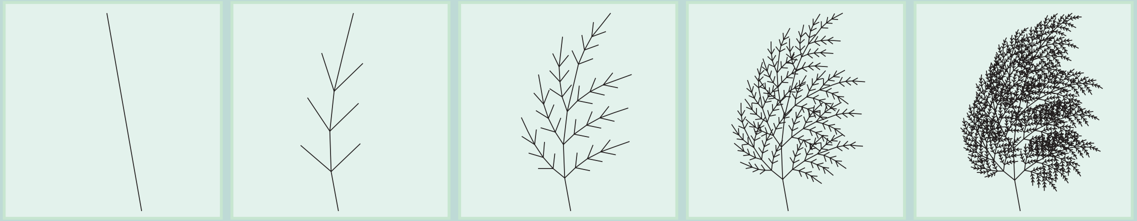


## Generative Fractals

The Web (and the output of the Barabási and Albert's model) may not look like a fractal to the casual observer, partly because it does not lend itself to visualization the way other fractals do. Nonetheless, the Web is just as much a fractal as the more familiar eye-pleasing fractals. It is just a little too much for the human eye to behold. However, we can see similarities between the Web and other fractals when we examine how each is produced.



L-systems, discovered by Aristid Lindenmayer [Lindenmayer 1968], simulate plant growth with only a small number of rules that specify how 'cells' grow into other cells. As can be seen, each iteration looks like how one would expect a plant to grow. Different 'seeds' and growth rules can be used to produce different types of plant-like structures.



MRCM fractals are produced by iteratively expanding parts of the fractal so that each part contains a smaller version of the whole. After a few iterations, the MRCM fractal will possess the signature look and feel of a fractal.

In both cases – as well as in the Barabási and Albert Web model – taking one stage, applying a simple rule to it, yields the next stage, and ultimately produces a fractal

In an article published in the *Proceedings of the National Academy of Sciences* in 2002, Pennock and his colleagues showed two important things. First, they showed that the distribution of inbound links for category-specific subsets of the Web, for example all University homepages or all movie homepages, follows the power-law but bump-like pattern seen in Figure 6.7. Second, they showed that a simple modification to the Barabási-Albert model predicts the observed data on the Web with remarkable accuracy. The new recipe looks like the following:

Create a new page, called *p*.

Flip a bias coin; if heads:

Randomly pick an existing link, *l*, not connected to *p*.

Randomly select one of *l*'s two adjacent pages, *q*.

Else, if tails

Randomly pick an existing page, *q*, not equal to *p*.

Add a new link between *p* to *q*.

Repeat steps 2-4 a total of *k* times.

The only new difference is that in step 2, we randomly switch between two types of new links, one that is preferentially based as before (in steps 2a and 2b) and one that is uniformly selected in 3a. In step 3a, the page we pick, *q*, is independent of how many links it has. The 'biasness' of the coin in step 2 influences to what degree the links tend to be preferentially based or non-preferentially based.

Looking back at our analogy between people, step 3a is akin having an association with a person that is not influenced by popularity, which is the main force behind the observed bump in the community link distributions. It turns out that this divergence from the pure power law, while most pronounced within topically coherent communities on the Web, also shows up to a lesser extent in a wide variety of distributions, including the distribution of *outbound* links on the Web and the distribution of movie actor collaborations.

### Web Surfing Patterns

Ultimately the Web is about people. Above our focus was on the behavioural patterns of Web authors. In this section, we focus on users. How do people typically surf the Web? Again, we won't get far by analyzing the intricacies of each and every surfer during each of their Web use sessions. Instead, we look instead for overall patterns of behavior and simplified rules of thumb that seem to capture the essence of observed aggregate behavior.

Along these lines, Bernardo Huberman and his colleagues [Huberman 1998] developed a simple and elegant model of surfing behaviour. In their model, a user continues to click deeper and deeper on a particular path of linked pages until he or she reaches a page of sufficiently low perceived quality; at which point the user abandons the current path and either gives up or begins anew, for example by typing in a URL directly, choosing a favorite bookmark, or initiating a web search.

Huberman and his colleagues showed that – assuming surfers on the whole obey the above tendencies – the depth to which the typical user surfs follows a type of power-law distribution called the inverse Gaussian distribution. In fact, data gathered from several different websites and user bases, over different time periods, match the conclusions of the model extremely well. Webmasters can even use the model to predict which pages will receive the most traffic on their site, and how to rearrange their site to maximize traffic to particular pages. Hence, users, in the aggregate, seem to surf Web pages in a fractal manner.

Another model of surfing behavior is called the 'random walk' model, and aptly so. Imagine a completely random surfer. Starting at a random page, this wandering surfer clicks on a random outgoing link, bringing him or her to a new page. From there, the surfer clicks another random link, moving to a third page, etc. The surfer continues like this ad infinitum, except that occasionally (with some small probability at each step) the surfer restarts, 'teleporting' from its current location to a completely random location.[1] Although the random walker model is by any measure an extreme simplification of reality, it turns out to be very powerful.

Because of the teleportation step, we know that the random walker can always move on to a new page. The key question is: which Web pages will the random walker visit most often if allowed to walk forever? It turns out that this question can be answered very elegantly with a remarkably straightforward calculation. The equations behind the calculation are very simple, but they must be performed for every page on the Web multiple times. Instead of diving into those mathematical details, we will instead try to capture the intuition of the random walker, which gets at the heart of what it means for a page to be important. If you think of a link as being an endorsement by an author that the page at the other end is high quality, then we get the following recursive rule:

*Web pages are important if other important pages predominately link to them.*

In the above rule, we use 'predominately' to mean that the page with the link has a relatively small number of outgoing links and, hence, is only 'voting' for a small number of other pages. (More outgoing links can be interpreted as an author is diluting his or her vote.)

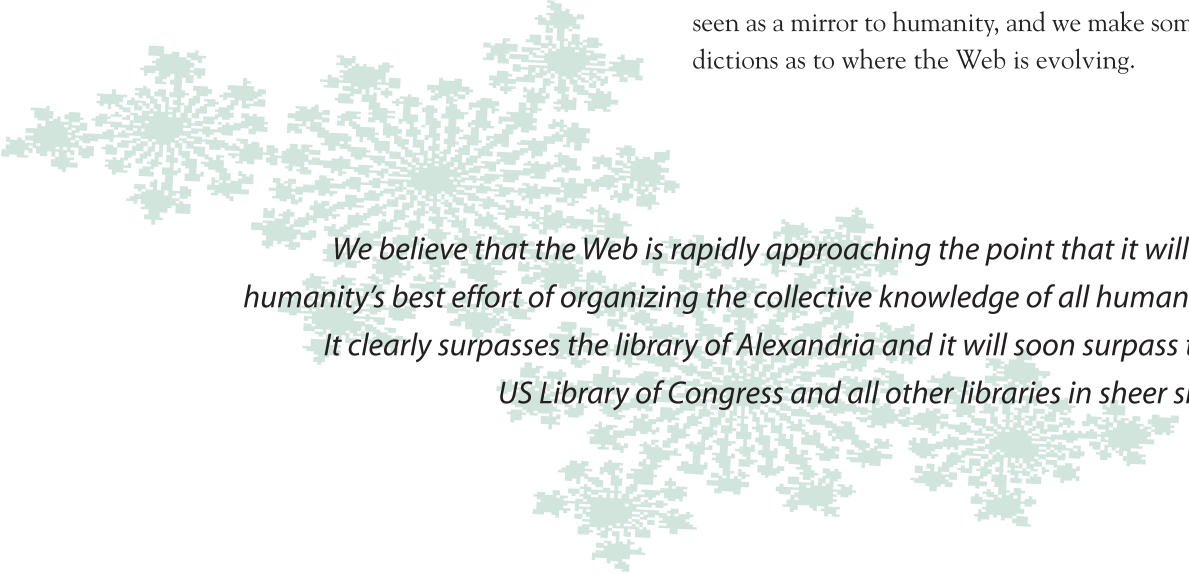
Larry Page and Sergey Brin, the founders of Google, discovered in 1998 that this calculation – which they dubbed PageRank after Larry Page – was very effective at separating quality pages from poor pages [Brin 1998]. In fact, when introduced, Google, with the help of PageRank, offered such a vastly better way of organizing the Web that Google came to lead the Web search industry.

The power of PageRank is that it uses the links of the Web (which are made by authors) and simulates how an infinite number of users given infinite time would visit those pages. The pages visited the most by the random walkers are deemed the best. Hence, Google makes explicit use of Web authors and implicit use of users to do a better job of finding quality content.

## The Web as a Mirror

We've now come full circle. Having examined the Web from a variety of scales and viewpoints, we have now seen how users, authors, and search engines all influence one another to yield an amazing array of self-organization, self-regulation, and self-similarity.

Ultimately, the Web's organization is intimately related to the complexity of human culture and to the human mind, and it is this subtle relationship between humanity and the Web that is responsible for the Web's amazing properties. In the remainder of this chapter, we will explore how the Web can be seen as a mirror to humanity, and we make some predictions as to where the Web is evolving.



*We believe that the Web is rapidly approaching the point that it will be humanity's best effort of organizing the collective knowledge of all humanity. It clearly surpasses the library of Alexandria and it will soon surpass the US Library of Congress and all other libraries in sheer size.*

### Search and the No Free Lunch Theorem

What computer scientists refer to as 'search' is perhaps the hardest mathematical problem in existence. By 'search' computer scientists mean all of the following:

Teach a computer to drive with only positive and negative reinforcement. That is, reward the computer when it gets to a destination scratch free, and punish it when it has an accident or goes to the wrong destination.

Find a model that accurately predicts the stock market both on historical data, and on future data, and make lots of money with it.

Beat anyone in the world at chess or the game of go.

Analyze the human genome and find all genes complicit in cancer.

Find the perfect document on the Web that satisfies the user's intent as indicated by a query.

Clearly, these are all hard problems. They all share in the fact that one is searching for a solution that is hidden among an infinite number of inferior solutions. Not only is what we are looking for hidden, but it may also be hard to recognize it when put right in front of your face.

Search is such an interesting problem precisely because it resembles learning, reasoning, evolution, and other forms of deep and profound adaptation. The topics of neural networks, artificial intelligence, and genetic algorithms are all subsets of the general search problem.

There is a mathematical result, published in 1997, due to David Wolpert and William Macready [Wolpert 1997], referred to as the 'No Free Lunch' theorem (or NFL for short) that is often misunderstood. The theorem deals with algorithms for solving the search problem. The theorem has both pessimistic and an optimistic interpretations, hence the confusion surrounding it. The pessimistic interpretation can be summarized as:

All search algorithms are equally bad.

Put in this way, it should be clear why so many people are unhappy with it. In fact, if you are a scientist that has invested years showing that your type of search algorithm is better than most, than the NFL theorem is outright slanderous.

A more complete characterization of the NFL theorem would be:

Averaged over all possible spaces, even crazy ones that never occur in nature, all search algorithms are equally bad.

This alternative view clarifies the major caveat with the NFL theorem, namely, that it is making a statement about all search algorithms if they were applied to all possible search problems *even ones that could never exist in our universe*. There is still another way to characterize the NFL theorem, which we believe is both optimistic and realistic:

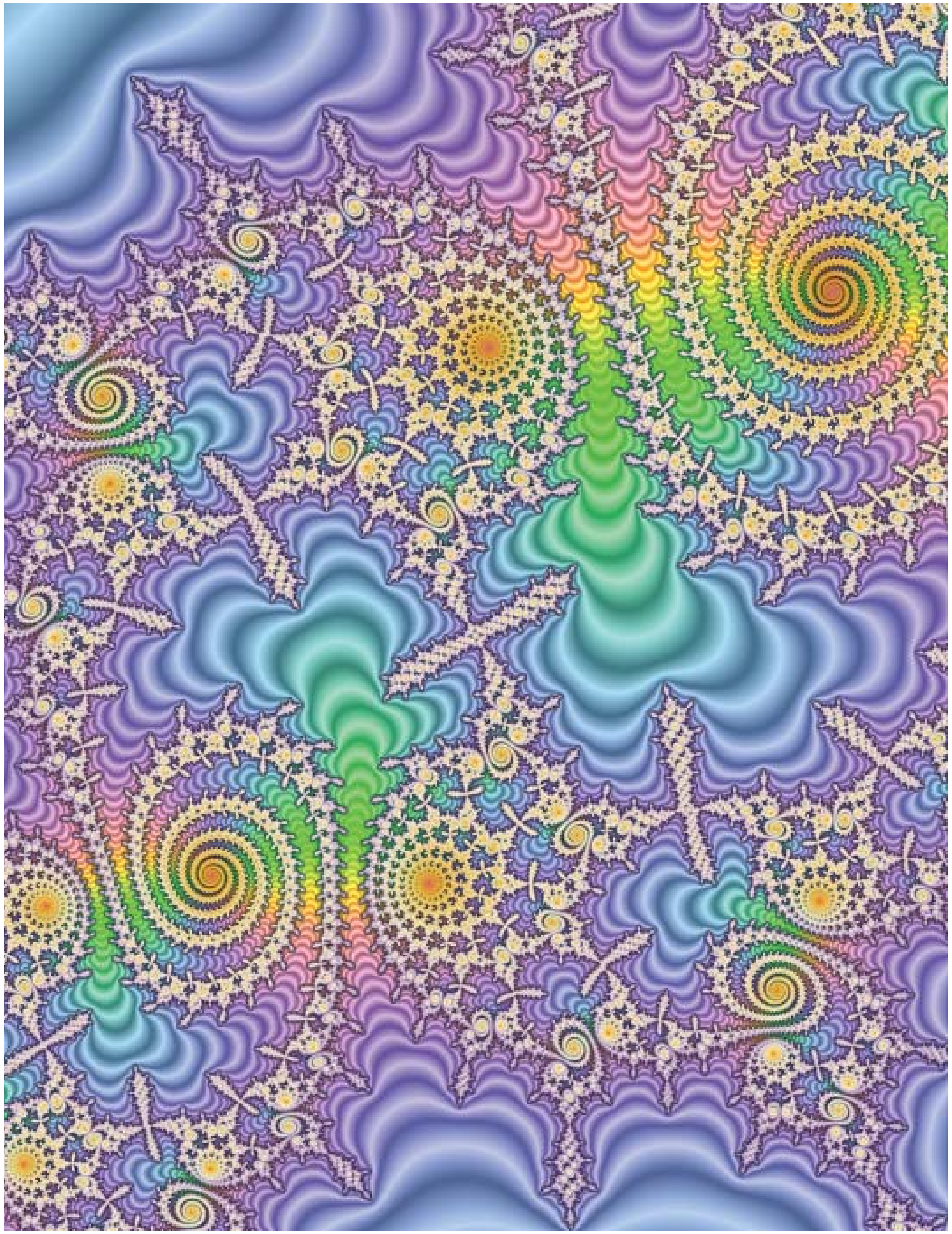
If your search algorithm is moulded to a particular problem space, it can work better than most other search algorithms.

The remaining caveat to this more gentle interpretation is that the penalty for being optimized to a particular problem domain is that the same solution that works well in one domain may prove horrible in every other domain. So it goes, we say.

All of this may seem to be completely unrelated to this chapter; however, we believe that the NFL theorem is key to understanding the current state of the Web and how it will evolve over time. In a nutshell, our claim is that the Web has co-evolved with humanity, and it will continue to do so. Moreover, we believe that the Web will approach a level of complexity that is on par with all human culture and with the human mind.

### Simplicity, Complexity, and Search

Much of this chapter has focused on how the Web possesses an amazing array of properties that smack of both simplicity and complexity. To better appreci-



ate this point consider how complicated a miniature version of the Web could be with only ten pages.

If each page is permitted to link to any of the ten, including itself, then there are $2^{100}$ different ways in which to connect up ten pages. This number is larger than the number of electrons in the universe. Now, instead of ten pages, think about billions and consider how complicated the Web could be if authors, pages, and users were not so regular in their collective behaviours. A billion pages with links pointing everywhere would truly be intractable and effectively unimaginable because no one would be able to make any sense out of it.

The point of this exercise is simple: the Web could have been tremendously complex, but it is not. In fact, the Web is exceedingly regular given its size and the lack of central authority. Moreover, this regularity can be exploited to make more effective algorithms for finding information on the Web.

Recall from the previous section our discussion of the PageRank algorithm. PageRank is mathematically very well understood. As an algorithm, it performs an iterative calculation that must be repeated multiple times, and the required number of iterations is easily known in terms of the error rate (associated with not running it for an infinite number of steps) and the properties of the link structure of the Web.

If the Web was not self-organized and if the link structure did not follow a power law, in all likelihood PageRank would not be a practical algorithm because its required number of iterations would be close to infinite. Instead, we know that the Web is a forgiving domain for PageRank, in the sense that its power law properties all but guarantee that PageRank will quickly converge to valuable results.

This is an extremely subtle but important point: the Web's self-organized and fractal properties make it easier for algorithms to make sense of it. Moreover, these self-organized and fractal properties are a direct consequence of our (humanity's) own self-organization and fractal nature.

## The Future

We believe that the Web is rapidly approaching the point that it will be humanity's best effort of organizing the collective knowledge of all humanity. It clearly surpasses the library of Alexandria and it will soon surpass the US library of Congress and all other libraries in sheer size.

The Web will continue to become an integral part of society, nearly blending into the background, as much of our society transitions into a dual nature that includes both a physical and a virtual existence.

We also believe that the generalized search problem – and the problem of building a nearly perfect search engine, in particular – will increase in importance as the need to find information on the Web becomes more ubiquitous and necessary to our day-to-day lives. In the future, Web search engines will radically change, ultimately possessing enough intelligence to simultaneously recognize the needs of the users that use it while making sense of the plethora of available information.

In short, we believe that the Web will become a mirror to humanity in the aggregate, and that the search engine will become a mirror to the human mind, and it is the self-organized and fractal nature of the Web that is both a symptom and a cause for this co-evolution.

[1] The teleportation step is required to deal with the sort of dead ends shown in the Bow-Tie model of the Web. Without the teleportation step, a random walker could effectively walk into a dead end and never be able to return. With the teleportation step, we are guaranteed that the random walker could walk forever and eventually visit each page on the Web if given enough time.

## Acknowledgements

We thank Bill Cheswick, KC Claffy, Roger Corn, William Decker, Dan Fain, Steve Lawrence, Will Leland, Margaret Murray, and William Pryor.

## References

[Barabási 1999] Albert-Lászlo Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, **286:** 509–512, 1999.

[Brin 1998] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th World Wide Web Conference (WWW7)*, 1998.

[Broder 2000] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: Experiments and models. *Proceedings of The 9th World Wide Web Conference (WWW9)*, 2000.

[Chakrabarti 2002] Soumen Chakrabarti, Mukul M. Joshi, Kunal Punera, and David M. Pennock. The structure of broad topics on the Web. *Proceedings of The 11th World Wide Web Conference*, 2002.

[Crovella 1996] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *Proceedings of ACM SIGMETRICS*, 1996.

[Crovella 1998] Mark E. Crovella, Murad S. Taqqu, and Azer Bestavros. Heavy-tailed probability distributions in the World Wide Web. In *A Practical Guide To Heavy Tails*, chapter 1, pp. 3–26, Chapman & Hall, New York, 1998.

[Dodge 2002] Martin Dodge and Rob Kitchin. *Atlas of Cyberspace*. Addison-Wesley, 2002.

[Dill 2001] Stephen Dill, Ravi Kumar, Kevin McCurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. Self-similarity in the Web. *Proceedings of International Conference on Very Large Data Bases*, pages 69–78, 2001.

[Faloutsos 1999] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. *Proceedings of ACM SIGCOMM*, 1999.

[Flake 2000] Gary William Flake, Steve Lawrence, and C. Lee Giles. Efficient Identification of Web Communities. Proc. ACM SIG KDD 2000, 2000.

[Flake 2002] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization of the Web and identification of communities. *IEEE Computer*, **35**(3):66–71, 2002.

[Gary 1979] M. R. Gary and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1979.

[Gribble 1998] Steven D. Gribble, Gurmeet Singh Manku, Drew Roselli, Eric A. Brewer, Timothy J. Gibson, and Ethan L. Miller. Self-Similarity in File-Systems, *Proceedings of ACM SIGMETRICS*, pp. 141–150, 1998.

[Huberman 1998] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, **280**(5360): 95–97, 1998.

[Kleinberg 1999] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**(5):604–632, 1999.

[Kleinberg 1999b] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a graph: Measurements, models, and methods. *Proceedings of the 5th International Conference on Computing and Combinatorics*, pp. 1–18, 1999.

[Kleinberg 2001] Jon Kleinberg and Steve Lawrence. The structure of the Web. *Science*, **294:** 1849–1850, 2001.

[Kumar 1999] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Proceedings of the 8th World Wide Web Conference (WWW8)*, 1999.

[Leland 1993] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Internet traffic. *Proceedings of ACM SIGComm*, 1993.

[Lindenmayer 1968] Aristid Lindenmayer, Mathematical Models for Cellular Interaction in Development, parts I and II, *Journal of theoretical biology*, **18**, 1968.

[Mandelbrot 1953] Benoît Mandelbrot. An informational theory of the statistical structure of language. In *Communications Theory* (Willis Jackson, ed.), Academic Press, New York, 1953.

[Mandelbrot 1959] Benoît Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon. *Information and Control*, **2**:90–99, 1959.

[May 1988] R. M. May. How many species are there on earth? *Science*, **214:** 1441–1449, 1988.

[Menascé 2002] Daniel Menascé, Bruno Abrahão, Daniel Barbará, Virgílio Almeida, and Flávia Ribeiro. Fractal Characterization of Web Workloads. *Proceedings of the 11th World Wide Web Conference, Web-Engineering Track*, 2002

[Pennock 2002] David M. Pennock, Gary William Flake, Steve Lawrence, Eric J. Glover, C. Lee Giles, Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, **99**(8):5207–5211, 2002.

[Pitkow 1998] James E. Pitkow. Summary of WWW characterizations. *Computer Networks and ISDN Systems*, **30:**(1–7):551–558, 1998.

[Schroeder 1995] Manfred Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*, W H Freeman & Co., 1995.

[Simon 1955] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, **42:** 425–440, 1955.

[Watts 1998] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, **393:** 440–442, 1998.

[Wolpert 1997] David H. Wolpert, and William G. Macready. No Free Lunch Theorems for Search. *IEEE Transactions on Evolutionary Computation*, 1, 1997.

[Zipf 1949] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Massachusetts, 1949.