Print Date: 6/8/2000

Trends and Patterns of WWW Browsing Behavior

by

Alan L. Montgomery and Christos Faloutsos

May 2000

Alan L. Montgomery is an Associate Professor at the Graduate School of Industrial Administration (e-mail: alan.montgomery@cmu.edu) and Christos Faloutsos is an Associate Professor at the School of Computer Science (e-mail: christos@cs.cmu.edu) at Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. The authors wish to thank Media Metrix for their generous contribution of data without which this research would not have been possible. Specifically, Tod Johnson, Steve Coffey, Oliver Ridley, and Paul Violino of Media Metrix have been instrumental to providing access. The work of Christos Faloutsos was supported by the National Science Foundation under Grants No. IRI-9625428, DMS-9873442, IIS-9817496, and IIS-9910606, and by the Defense Advanced Research Projects Agency under Contract No. N66001-97-C-8517. Additional funding was provided by donations from NEC and Intel. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, DARPA, or other funding parties.

Abstract:

Are there any trends in web browsing behavior by individual users? On the popularity of web sites? On the frequency of revisitation? These are questions that we address in this work. Our study uses an extensive clickstream data from U.S. households from July 1997 to December 1999. It is unique, for several reasons: (1) it uses an extensive clickstream data from U.S. households for a long interval (30 months), and (2) it uses web viewing patterns captured from the browser window, as opposed to relying upon hits recorded in server logs, and (3) these users comprise a representative sample of U.S. web users. We present some surprising empirical regularities that seem to hold consistently for the time-span of our dataset. Specifically, we found that (a) the growth of the internet usage seems exponential over time, with a growth rate of about 2.4% per month (b) the growth is mainly due to the new users, while the per-user behavior seems to remain stable over time (c) the popularity of web sites (visits per month per user) seems to follow a power law, with slope about -1.2 (d) the 'stickiness', that is, revisitation patterns (inter-arrival times) follows another power law, with slope -.6.

Keywords: Internet, World Wide Web, Clickstream Data, Zipf Distribution, Power Laws

Introduction

It is now more than six years since the World Wide Web (WWW) entered the popular culture, which provides an opportunity for thinking not only about current behavior but how changes in browsing patterns have changed through time. During this time the growth of the WWW has been phenomenal, growing from two million servers in 1994 to more than 70 million in 2000 according to estimates by Network Wizards. The number of U.S. home users during this time has also increased from 3 million to more than 62 million according to estimates by Media Metrix. One might suspect that this phenomenal growth has also resulted in fundamental changes in the way users browse through the net. Our study examines trends and patterns in web usage using clickstream data for the last two and a half years (July 1997-December 1999). While the number of web users has followed an exponential growth curve, we show the time that they browse online has increased at a slower rate. Many browsing patterns, such as the number of pages viewed during a session and the number of domains viewed within a session, have remained stable during this time. Also, the distribution of certain measures like the number of times a URL is viewed during a session and the number of clicks between a URL revisitation follows a Zipf-like pattern. Zipf distributions have been observed widely in computer science and are characterized by their long-tails.

A key element in this study is a large scale empirical analysis of clickstream data from a representative sample of home web users in the U.S. This research has been made possible by a clickstream data set collected by Media Metrix using their PC Meter with a nationwide panel of over 20,000 web users. The PC Meter is a Microsoft Windows or Apple Macintosh application program that records information about who is using the computer, what applications are running, and how long they are in use (Coffey 1999). When a WWW browser (Netscape's Navigator, Microsoft's Explorer, America Online, etc.) is active the PC Meter captures the Uniform Resource Locator (URL) being viewed. These households are recruited largely through telephone and mail and represent a random sample of U.S. PC owning households. To encourage households to participate they are given periodic gifts and prizes as incentives, although the nominal values of these gifts tend to be small. In addition, to the clickstream data collected by the PC Meter, extensive demographic information about the household is collected through conventional surveys. Our database is comprised of 290 million viewings by 74,000 unique individuals that occurred during 30 months.

There are two other potential sources for collecting clickstream data. The most popular source is

through the host server or the computer of the site being visited. As a user requests a page, identifying information (IP address, previous URL, and browser type) is recorded in the server logs. Another source is for a third party to capture information about web requests. For example, if a user connects to an Internet Service Provider (ISP) or Commercial Online Service (COS), such as AOL, any requests that the users makes can be recorded as they are passed on to the requested server. Therefore, ISPs and COSs can be a prime source for clickstream data.

For the purposes of this study, there are two important differences in the clickstream data collected by the PC Meter, as opposed to the clickstream data used in other published academic studies. First, our dataset contains a sample of representative U.S. home users, as opposed to all users who traffic a particular site. Therefore, it does not contain the self-selection biases present in most other datasets. Second, it contains viewing information as opposed to hits. A **viewing** means that a URL was displayed in the browser, as opposed to a **hit**, which occurs when a request to an HTML request was serviced by an web server. Many ISPs and COSs cache their users' page requests to speed user requests and reduce web traffic. Therefore many viewings will be cached either at the browser level or by ISPs or COSs, and will not result in a hit. This means that server logs contain only a subset of the viewings that occur.

Date and Time of Access	Seconds spent Viewing	URL
18JUL97:18:55:57	47	*http://www.voicenet.com/*
18JUL97:18:56:44	37	*http://www.weather.com/weather/us/cities/HI_Lahaina.html*
18JUL97:18:57:25	105	*http://www.weather.com/weather/us/cities/MI_Traverse_City.html*
18JUL97:19:03:00	7	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:03:56	2	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:03:58	6	http://www.weather.com/weather/us/cities/HI_Lahaina.html
18JUL97:19:04:58	2	http://www.weather.com/weather/us/cities/HI_Lahaina.html
18JUL97:19:05:00	1	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:15:24	39	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:17:00	7	http://www.weather.com/weather/us/cities/MI_Traverse_City.html
18JUL97:19:17:07	13	*http://www.realastrology.com/*
18JUL97:19:17:20	44	*http://www.realastrology.com/libra.html*

Table 1. Sample of clickstream data from one session, italicized URLs denote those hits that would be likely to occur in the logs of the host site.

To illustrate the clickstream collected by the PC Meter an actual user session is listed in Table 1. A user **session** includes all web browsing activity, with less than 120 minutes of inactivity between viewings[1]. Notice that the user frequently views the same page multiple times and may pause to do other tasks between page views (e.g., run other applications, watch television, etc.). Only five out of the 12 viewings, which are italicized in Table 1, would generate a potential request that would generate a hit on the servers logs. This illustrates an advantage of our clickstream since it is collected at a user's machine and not from the host site, which eliminates a potential source of bias.

The rest of the paper is organized as follows. First, we propose an exponential model for the increase in web usage over time. We show that web usage for the average user has been increasing steadily for all types of users. However, the dominate factor in understanding recent growth of the Internet has been the larger number of web users, and a secondary factor is the increased usage per individual. Second, we propose a power law for URL viewings. We demonstrate that the number of times a URL is revisited during a session can be well described by a Zipf distribution. Third, we consider the distribution of the number of domains visited by an individual during the month and find that is has remained stable through time. Fourth, we propose a power law for URL and domain revisitations. Finally, we conclude by summarizing our findings and consider some of their implications.

Exponential time trends of web usage

A straightforward measure of internet activity is the number of viewings a web user makes per month. An alternate measure is time actively viewing a page which tends to be highly correlated with viewings and yields similar results. The average number of viewings per month was 750 for December 1999 which is 50% more than the average of 500 viewings in July of 1997. Notice during this same time the number of users as estimated by Media Metrix has increased from 28 million to 62 million, or a 120% increase. So, not only are the number of users increasing, but each user is viewing more pages. However, the more important factor in explaining web growth during this period is the increase in number of users, while increased usage per individual is secondary.

The distribution of individual usage has very high variance. To better illustrate the distribution we plot

1. Several cutoff values were tried, but this choice does not alter the substantive findings of this research.

the 10th, 25th, 50th, 75th, and 90th percentiles of the distribution of monthly viewings per individual in Figure 1. For example, the mean number of viewings per month was 750 in December 1999 and the 50th percentile (the median) was 310. Since the mean exceeds the median we can characterize this as a distribution which is reverse-J shaped with quite a bit of mass in the tail of the distribution. Notice that the mean is very close to the 75th percentile, which again shows that the mean is not the best indicator for a typical user. This figure shows that growth in web viewings per user has been small and steady during this period. However, there is some fluctuation around this exponential trend, so that growth is not monotonic. Perhaps deviations from the trend can be explained by news events or seasonality.

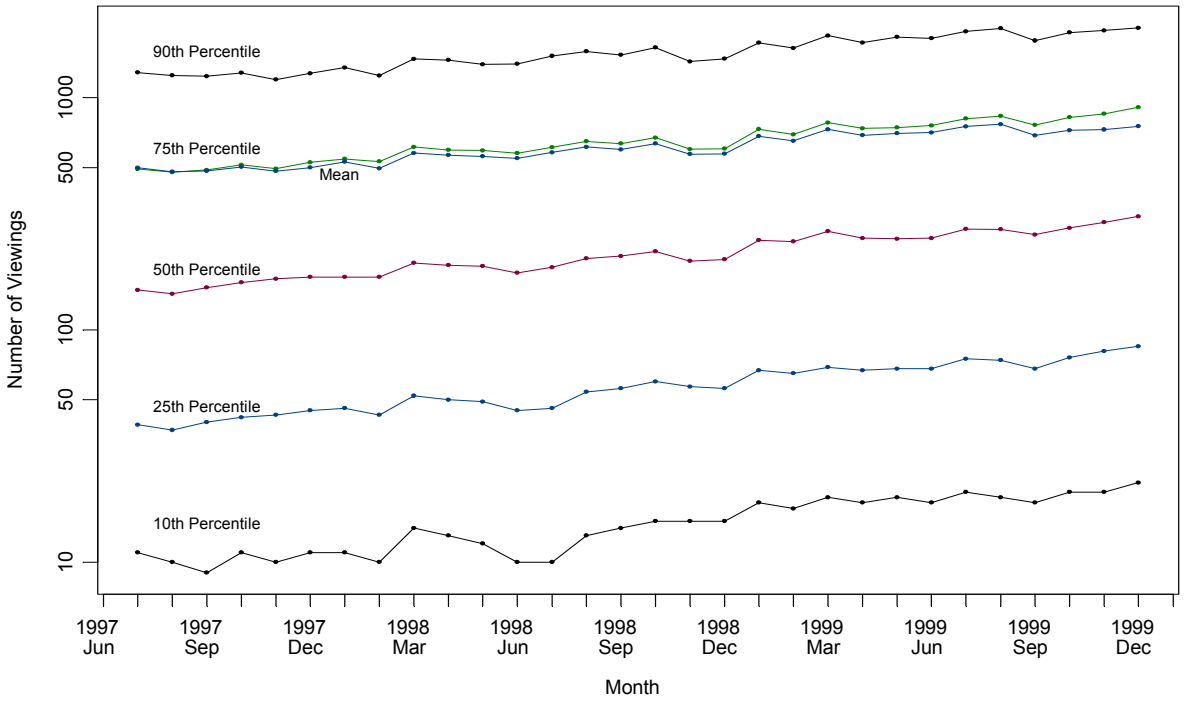

Figure 1. Percentiles from the distribution of the number of viewings per month.

The growth in the number of viewings per month is fairly linear on this logarithmic scale, which indicates exponential growth. If we fit a time trend we find that growth in the median is occuring at a rate of 2.4% per month. This means that every 30 months the number of viewings per individual will double, if current trends hold. For the average user, growth is occuring at a rate of 1.7% which would double in about 3½ years. Notice that there is growth in all percentiles, although the smaller percentiles are growing at a faster rate than the

larger percentiles. This implies that growth for less frequent users is occuring at a faster rate than more frequent users. For example, the 10th percentile is growing at a rate of 2.9% per month, while the 25th, 50th, 75th, and 90th percentiles are growing at 2.6%, 2.4%, 2.1%, and 1.8%, respectively. Hence, the users at the 10th percentile would double their usage in 24 months, and users at the 90th percentile will double their usage in 40 months. However, it would take more than 40 years for users at the 10th percentile to "catch up" to users at the 25th percentile. It would appear the present appearance of the distribution of viewings per user could stay in tact for sometime.

To assess WWW usage in a more natural context this individual level browsing is considered on a session by session basis instead of a monthly basis. We define a session to be a period of sustained use of the WWW or computer, a new session is ended whenever the user has not accessed the web for more than two hours. The average number of viewings during December 1999 was 750, which were comprised of an average 8.1 sessions per month with 93 viewings per session.

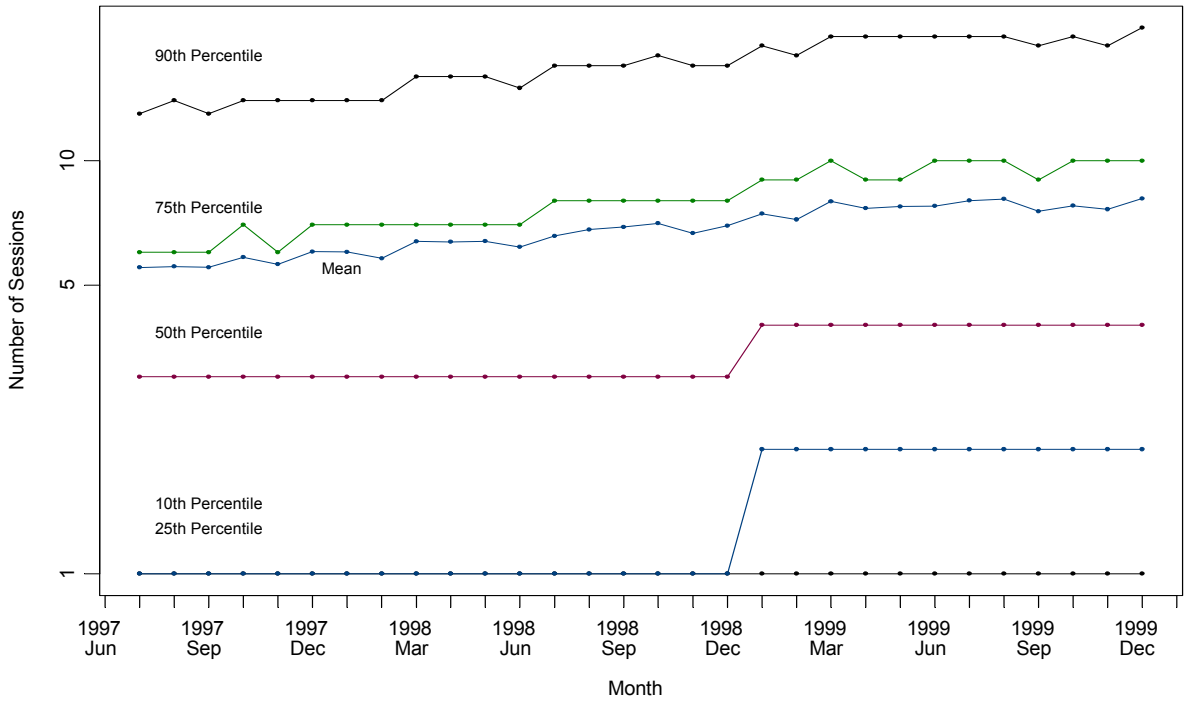

Figure 2. Percentiles from the distribution of the number of sessions per month across users.

The percentiles for the number sessions per month across users is given in Figure 2. Notice that the number of sessions has been increasing slowly–but steadily–through time. The median number of users have

four or fewer sessions per month, while 10% exceed 20. This leads to a very skewed distribution in which a small number of users are accounting for most of the sessions. The small numbers of sessions below the median gives the smaller percentiles their stepped appearance. However, there is still a consistent growth in sessions per month. The average number of sessions per user has been increasing at a rate of 1.4% per month, which means that the average number of sessions will double in a little more than four years if present growth trends continue.

Figure 3 provides the percentiles for the distribution of the number of viewings per session across individuals. For example, the median number of viewings per session is 48 in December 1999, which is an increase over 42 in July 1997. Notice that these distributions have been quite stable through time. In fact the median number of viewings has grown at .5% per month. By decomposing the number of viewings, into the number of sessions and number of viewings per session, we can determine that most of the growth of web activity is more frequent sessions, not longer sessions. The distribution of the number of viewings can be well approximated by a generalized gamma distribution to capture the thicker tails than is observed in the standard gamma distribution.

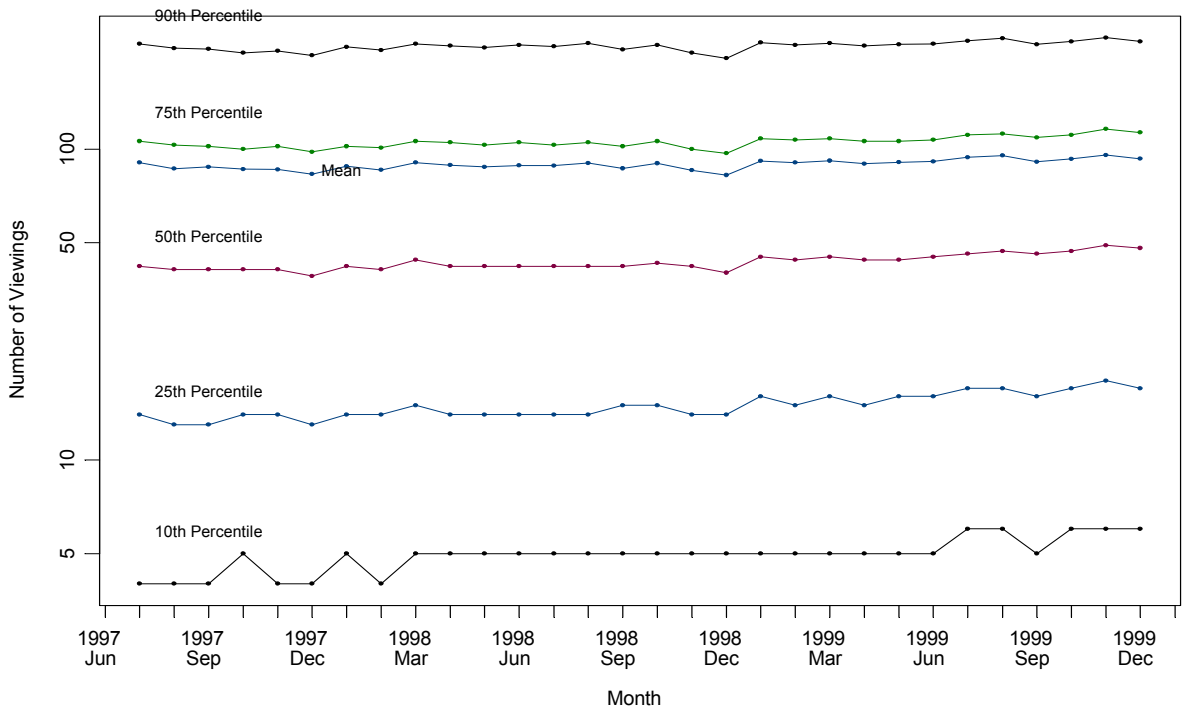

Figure 3. Percentiles of the distribution of the number of viewings per session across individuals.

A Power Law for URL Viewings

Skewed distributions and power laws appear often in practice. A power law is characterized by the relationships of the following form: $y=x^a$, where x and y are the variables of interest (eg., force and distance in Newton's Law of Gravity). A popular distribution with this property was discussed by Zipf (1949). He showed that the rank-frequency plots for many pieces of text follow a power law with slope close to -1. Power laws are intimately related to self-similarity properties and to fractals. For an intriguing discussion see the classic book by Mandelbrot (1977). Many other power laws have been discovered, in multiple settings: the distribution of income (Pareto's law), the distribution of publication counts (Lotka's law, Lotka 1926), the length of file transfers in web servers (Crovella and Bestavros 1996), the distribution of the areas of the islands in an archipelagos (Korcak's law, Schroeder 1991), and many more. In computer science, several power laws hold for the Internet topology (Faloutsos et al. 1999); Barabasi showed that power laws hold for the web if we consider it as a graph (web pages are nodes, and hypertext links are edges), with similar discoveries from the CLEVER group of IBM (Kumar et al. 1999). Huberman et al. (1998) showed that a power law holds for the number of pages that a user visits within a given web site. A more comprehensive review of empirical characteristics of web traffic is given by Pitkow (1998).

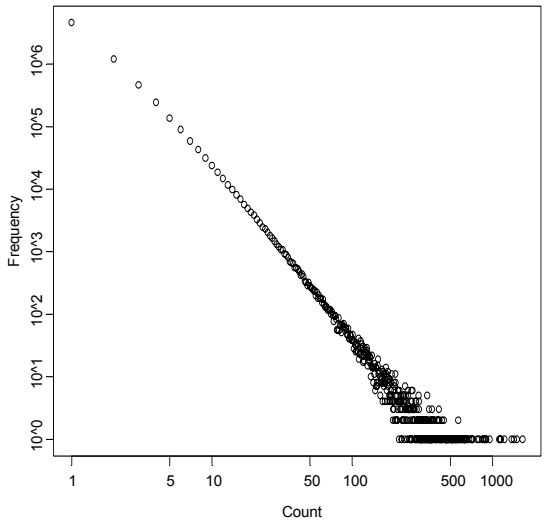

Figure 4. Plot of the count of the number of URL revisitations that occur during a session against the frequency of this count.

The distribution of the number of times a URL is revisited within a session is given in Figure 4, using data from December 1999. The linear relationship in the log of the frequency versus the log of counts is well described by a Zipf distribution. For example, there were 4,618,256 URLs that were viewed once, there were 1,205,333 URLs that were viewed twice, etc. The maximum likelihood estimate for the parameter of the Zipf distribution is 1.25. This is the first time viewings-not hits-of URLs has been considered. Analysis of the number of viewings of a URL per month yield a similar Zipf type pattern.

An explanation for the Zipf behavior of URL revisitations comes for an analogy with word usage. The frequency with a particular words is used follows a Zipf pattern. Certain words are repeated extensively (e.g., "the"), while other words are unique and occur infrequently, which is manifested by the long-tails of the Zipf distribution. To understand URL usage, we can think of the URLs as forming the user's vocabulary. Certain pages (i.e., "words") are constantly reused, perhaps as a navigational tool from which to move to other pages, while the remaining pages are somewhat unique and occur more infrequently. The huge diversity of web pages offers the user a large "vocabulary" from which to browse the web, and offers some insight into why the Zipf distribution may well describe the pattern observed in Figure 4.

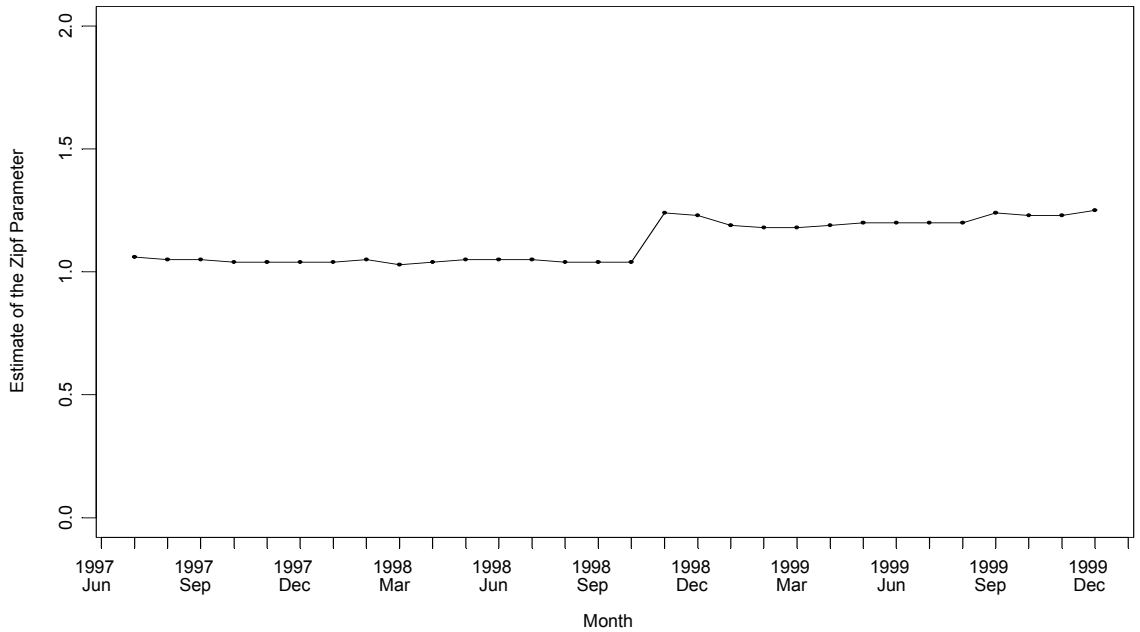

Figure 5. Estimates of the Zipf parameter for the distribution of the number of times a URL page is viewed during a session.

Additionally, we find that this Zipf distribution has been very stable through time. Figure 5 plots the estimates of the Zipf parameter by month. For estimation purposes each month is treated independently. Notice that until October of 1998 the estimates of the distribution average 1.05, while after this period they average about 1.25. The jump in the estimates of the Zipf parameter coincides with a change in the definition of the sample of web users by Media Metrix. This change was incorporated after the merger with Relevant Knowledge.

Alternately, the stability of the distribution of the times a URL page is viewed can be viewed directly through the percentiles of the distribution as plotted in Figure 6. Again the highly skewed nature of the distribution can be viewed since the 10^{th}, 25^{th}, and 50^{th} percentiles all coincide at the value of 1. The mean, which is highly influenced by the outlying observations, decreases only slightly.

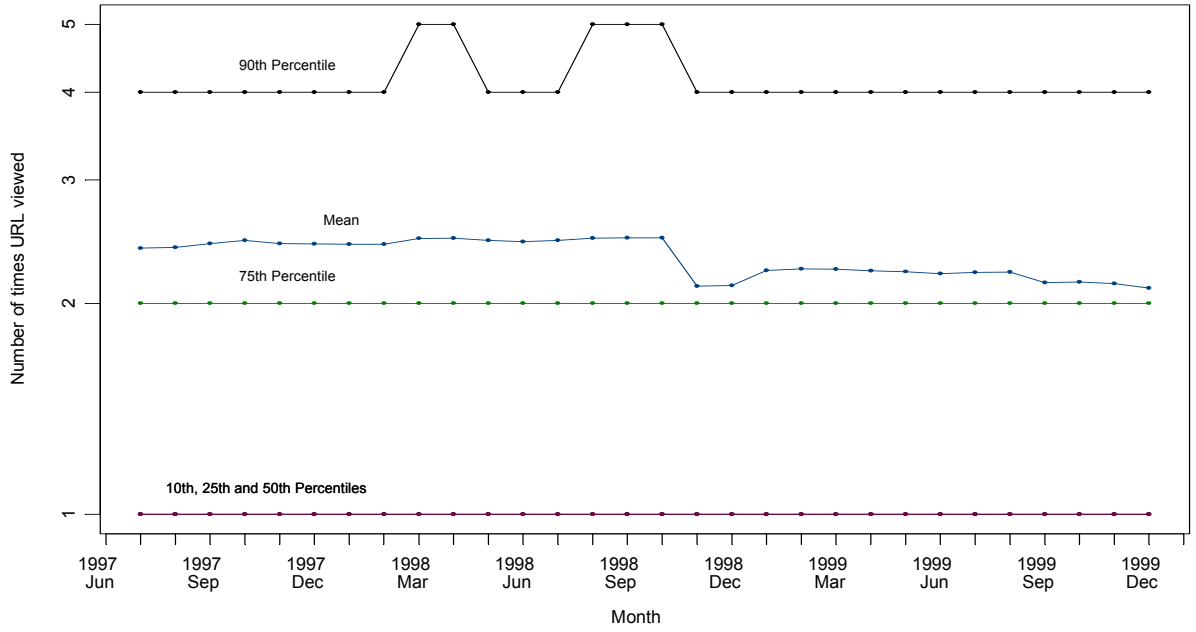

Figure 6. Percentiles of the distribution of the number of times a URL is viewed during a session.

Domain Viewership

The wealth of sources of information available might lead one to hypothesize that users are viewing diverse sets of information from a wide variety of sources. As the number of web sites has increased through time, it might be expected that this has encouraged users to look at a wider range of sources. However, to the

contrary usage data indicates users are looking at a fairly small number of sites relative to the number of viewings. The median number of hosts visited during a December 1999 is 25. This has been steadily rising over the past two and a half years, in July 1997 the median number was 14. Notice that the growth in the number of domains viewed parallels the growth of the number of URLs viewed. Again, apparently this ratio of viewings to the number of domains viewed has remained fairly steady through time, even as the number of viewings has increased. The percentiles of the distribution of the number of unique hosts viewed during a month is given in Figure 7. Notice that there has been a slow and steady increase as with the increase in the number of URLs viewed as shown in Figure 1.

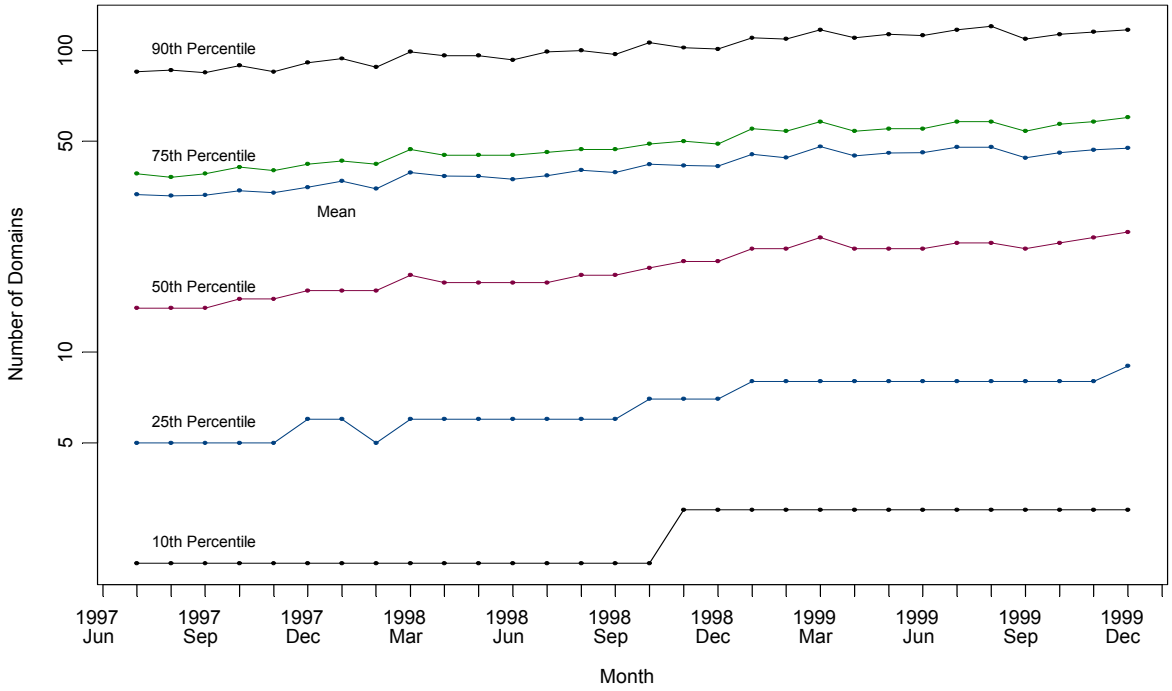

Figure 7. Percentiles of the distribution of the number of unique domains viewed during a month.

It is surprising that given the millions of sites available and a plethora of search engines users limit their attention to such a small number of sites. Perhaps users exhibit a certain persistence due for cognitive reasons or loyalty to a particular domain. One explanation is that it takes time and effort for a user to become familiar with a site, so users' focus their usage on a set of sites that they have more familiarity. If the power law of learning is in operation here with respect to a user's ability to navigate a web site, it make sense to return more

often to a site once its structure has been learned. The implication of this conjecture is that there are barriers to entry. For example, suppose a user wishes to search for a new book and has a great deal of experience with a particular web site. If the user wishes to find a new book, even if another web site offers a lower price, the user is more likely to return to the site with which he has high familiarity. Therefore, the experience previous users have with a web site represent a barrier to entry. Another explanation is that evaluating new sites is effortful and users retain those sites that satisfy their needs.

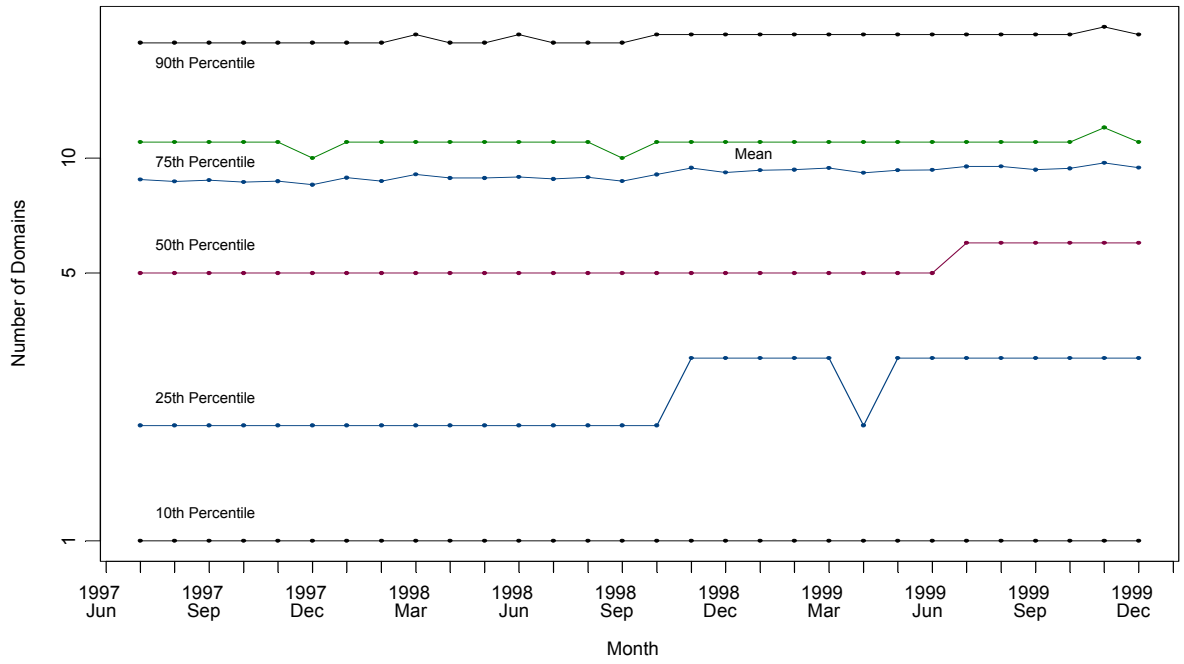

Figure 8. Percentiles of the distribution of the number of unique domains viewed during a session.

Figure 8 shows the percentiles of the distribution of the number of unique domains viewed during a session. Even though their has been growth in the number of unique domains viewed during a month, the number of domains viewed per session has been relatively stable. This is similar to the fact that total viewings is going up largely due to the increased number of sessions, and not more viewings per session.

Power laws for URL and Domain Revisitations

The previous discussion shows that users revisit the same page and site frequently, therefore it is natural to consider the number of viewings between the time that a viewer revisits a particular URL or domain. For

example, during December 1999 54% of URLs were revisited at least once during a session, and of those that were revisited 35% were viewed consecutively (i.e., two viewings of the same page occurred in a row), 22% had one viewing in between, perhaps the user returned to this page through the *back* button, and the remaining 43% had more than one viewing in between. Figure 9 shows the percentiles of the distribution of the number of viewings until a URL is revisited during a session conditional on the fact that the URL was revisited. The dotted line represents the percentage of URLs that are revisited at least once.

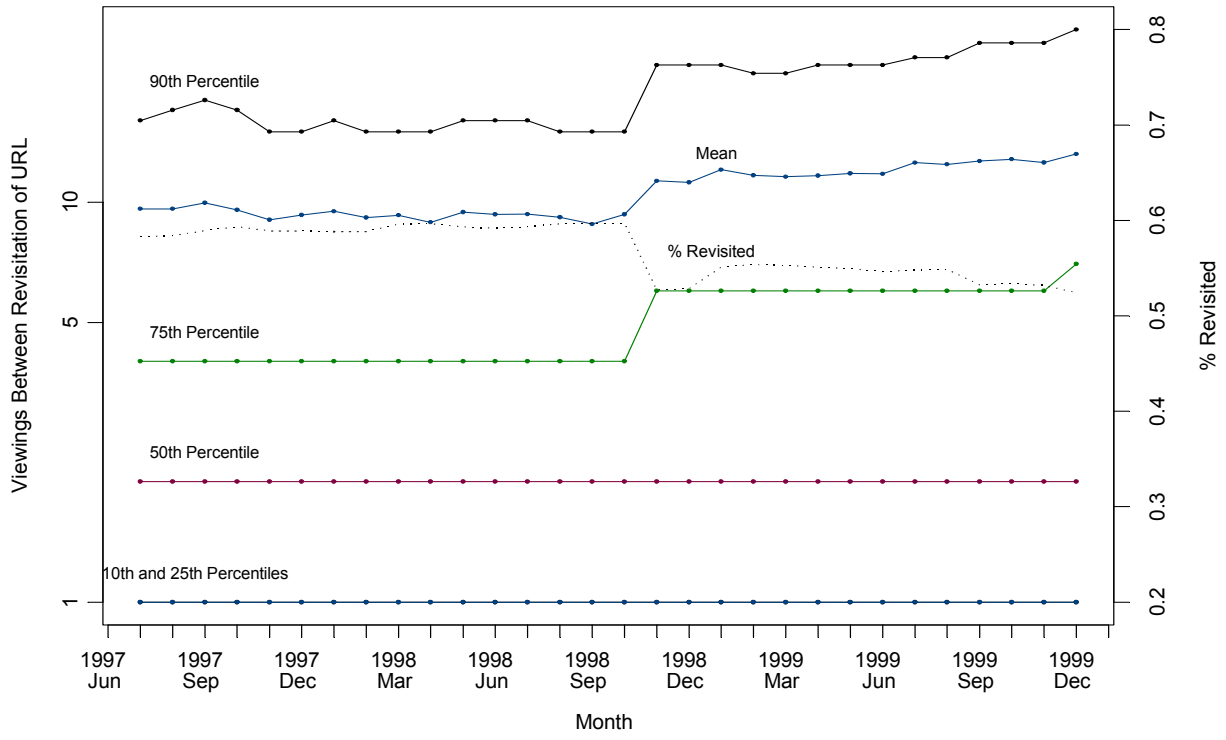

Figure 9. Percentiles of the distribution of the viewings between the revisitation of a URL during a session conditional on the fact that it was revisited. The percentage of URLs that were revisited is denoted with the dotted line.

The distribution of the number of viewings displays a long-tailed distribution, and may be approximated with a Zipf distribution. The count versus frequency of the revisitations is given in Figure 10 for the number of viewings until a URL is revisited for December 1999. An estimate of the Zipf parameter for this distribution is .58.

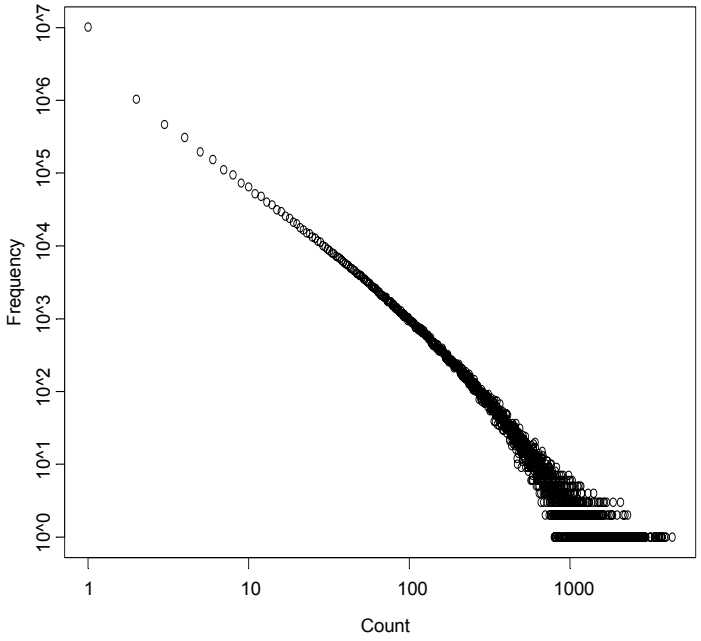

Figure 10. Plot of the frequency versus count of the number of viewings until a URL is revisited during a session for December 1999.

In contrast to the revisitation of a URL during a session, the revisitation of domains has an even greater probability of revisitation. In fact, if a revisitation to a domain occurs, the domain is revisited within one viewing 75% of the time. The probability that a domain will be revisited during a session has averaged 90%, and has stayed between 89% and 91% through the entire time period. The percentiles of the distribution of the number of viewings until a domain is revisited is given in Figure 11, again the dotted line represents the probability that a domain will be revisited at least once during a session. Notice the stability of this distribution through time. It too can be well approximated by a Zipf distribution with an estimated parameter of .6.

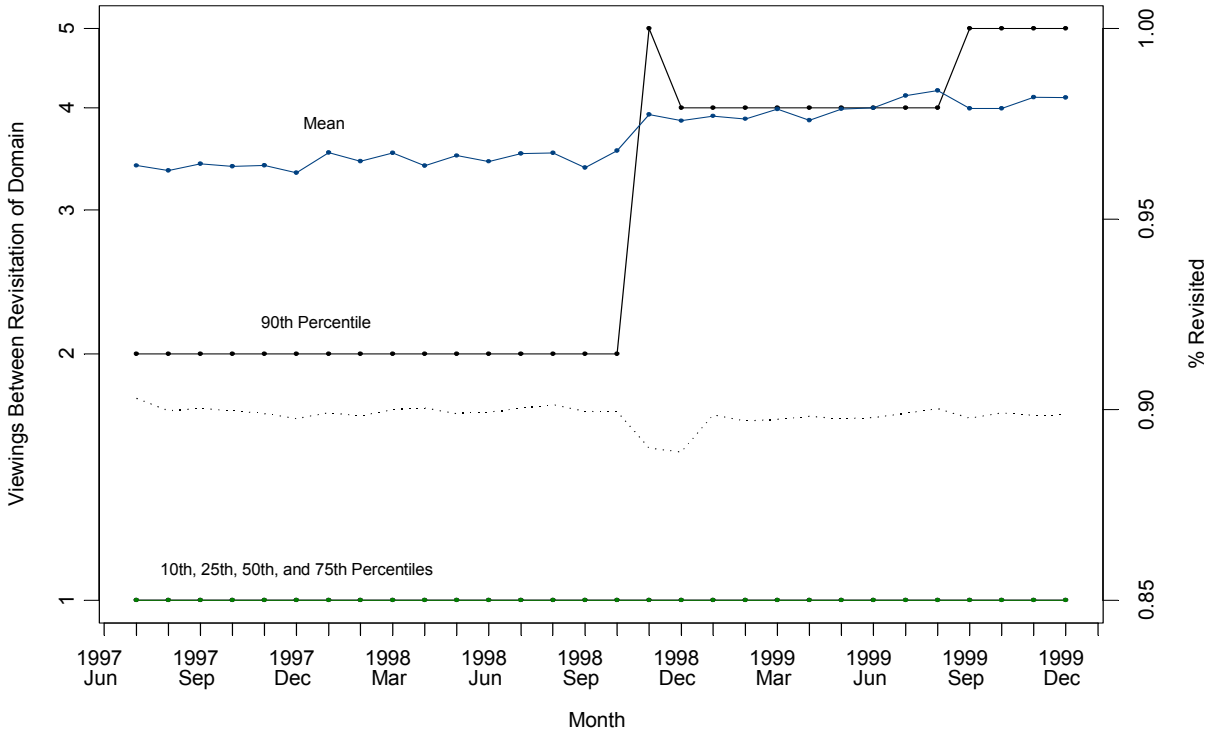

Figure 11. Percentiles of the distribution of the viewings between the revisitation of a domain within a session conditional on that domain being revisited. The dotted line represents the percentage of domains revisited during a session.

Summary of Findings

This research has provided an academic study of a sample of WWW home users that is unique in its representativeness of U.S. households and the fact that it has been collected over a two and a half year. This time period represents a time of explosive growth of the Internet, both in terms of content and users. Our results helps shed light on WWW usage and demonstrates that even through this period of explosive–and seemingly unruly–growth, there are some characteristics of browsing behavior that have pronounced trends and others that have remained very stable through time. We summarize our findings about trends and patterns in web browsing in the following empirical 'laws':

1. *The number of monthly viewings per user is growing exponentially.* The growth rate has remained steady at about a growth rate of 2.4% per month. This growth rate implies that the usage of the median user will double every 30 months. This bears some resemblance to Moore's Law, which states transistor density on integrated

circuits will double every 18 months.

2. *The number of sessions per user has been growing exponentially.* The mean number of sessions across users per been growing exponentially at a rate of 1.4% per month.

3. *The number of viewings per session is stable through time.* Currently the median number of viewings per session is 48.

4. *The number of domains viewed per session is stable through time.* Currently the median number of domains viewed per session is 6.

5. *Inter-arrival times follows a power law.* The probability that a URL is revisited at least once during a session is about 54%. Given that a page is revisited, the number of viewings before returning to a page follows a Zipf distribution with a slope of -.6.

Some implications of these laws are:

a. The dominant growth factor of the WWW has been the increase in the number of users. A secondary contributing factor is an increase in the usage per individual.

b. The number of sessions has increased, yet the number of viewings per session has remained stable. This means that the increase in usage per individual is coming through more sessions, while the length of a session has remained fairly stable. This implies that for all the change that has occurred with web content and access, how an individual users the web during a session has remained fairly stable during the past two and a half years.

c. Using hits to measure viewership, undercounts the times a page is viewed by an order of magnitude. Every time an individual requests a URL, that page will be viewed in the browser window 2.1 times. Therefore, using server logs to estimate viewership can result in a substantial undercount of how often a user looks at that page.

d. Users are loyal to hosts. It is surprising to the authors that given the millions of domains that can be accessed, users sample only a handful. Moreover, the number of domains per viewing has remained fairly constant. This supports the notion of "stickiness", i.e., that users are persistent in their viewing habits.

e. Viewings and revisitations follow power laws, i.e., Zipf-like distributions. We found some surprising power laws, whose slope remains fairly constant through time. An important facet of these distributions are their long tails. A byproduct of our study is that the average is a poor estimate of central tendencies, indicating that it is much better to consider a particular percentile of the distribution, or even better, the slope of the 'power law'.

References

Barabasi, Albert-Laszlo and Reka Albert (1999), "Emergence of scaling in random networks", *Science*, 286: 509-512.

Coffey, Steven (1999), "Media Metrix Methodology", Media Metrix Working Paper, http://www.mediametrix.com/Methodology/Convergence.html.

Crovella, M. and A. Bestavros (1996), "Self-Similarity in World Wide Web Traffic, Evidence and Possible Causes", *Sigmetrics*, 160-169.

Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos (1999), "On Power-Law Relationships of the Internet Topology", SIGCOMM.

Huberman, Bernardo A., Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose (1998), "Strong Regularities in World Wide Web Surfing", *Science*, 280, 95-97.

Kumar, S. Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins (1999), "Extracting Large-Scale Knowledge Bases from the Web", VLDB, 639-650.

Lokta, Alfred J., (1926), "The Frequency Distribution of Scientific Productivity", *Journal of the Washington Academy of Sciences*, June 19, 1926: 317-323.

Mandelbrot, B. (1977), *Fractal Geometry of Nature*, W.H. Freeman, New York.

Pitkow, James E. (1998), "Summary of WWW characterizations", *Computer and ISDN Systems*, 30, 551-558.

Schroeder, Manfred (1991), *Fractal, Chaos, Power Laws: Minutes from an Infinite Paradise*, W.H. Freeman and Company, New York.

Zipf, G. K. (1949), *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*, Addison Wesley, Cambridge, Massachusetts.