Glottometrics 3, 2002,143-150 *To honor G.K. Zipf*

Zipf's law and the Internet

Lada A. Adamic[1]
Bernardo A. Huberman

Abstract. Zipf's law governs many features of the Internet. Observations of Zipf distributions, while interesting in and of themselves, have strong implications for the design and function of the Internet. The connectivity of Internet routers influences the robustness of the network while the distribution in the number of email contacts affects the spread of email viruses. Even web caching strategies are formulated to account for a Zipf distribution in the number of requests for webpages.

Keywords: Zipf's law, caching, networks

Introduction

The wide adoption of the Internet has fundamentally altered the ways in which we communicate, gather information, conduct businesses and make purchases. As the use of the World Wide Web and email skyrocketed, computer scientists and physicists rushed to characterize this new phenomenon. While initially they were surprised by the tremendous variety the Internet demonstrated in the size of its features, they soon discovered a widespread pattern in their measurements: there are many small elements contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others.

This pattern has of course long been familiar to those studying distributions in income (Pareto 1896), word frequencies in text (Zipf 1932), and city sizes (Zipf 1949). It can be expressed in mathematical fashion as a power law, meaning that the probability of attaining a certain size x is proportional to $x^{-\tau}$, where τ is greater than or equal to 1. Unlike the more familiar Gaussian distribution, a power law distribution has no 'typical' scale and is hence frequently called 'scale-free'. A power law also gives a finite probability to very large elements, whereas the exponential tail in a Gaussian distribution makes elements much larger than the mean extremely unlikely. For example, city sizes, which are governed by a power law distribution, include a few mega cities that are orders of magnitude larger than the mean city size. On the other hand, a Gaussian, which describes for example the distribution of heights in humans, does not allow for a person who is several times taller than the average. Figure 1 shows a series of scale free distributions in the sizes of websites in terms of the number of pages they include, the number of links given to or received from other sites and the number of unique users visiting the site.

[1] Address correspondence to: Lada A. Adamic, HP Laboratories, 1501 Page Mill Road, ms 1139, Palo Alto, CA 94304, USA. E-mail: ladamic@exch.hpl.hp.com.

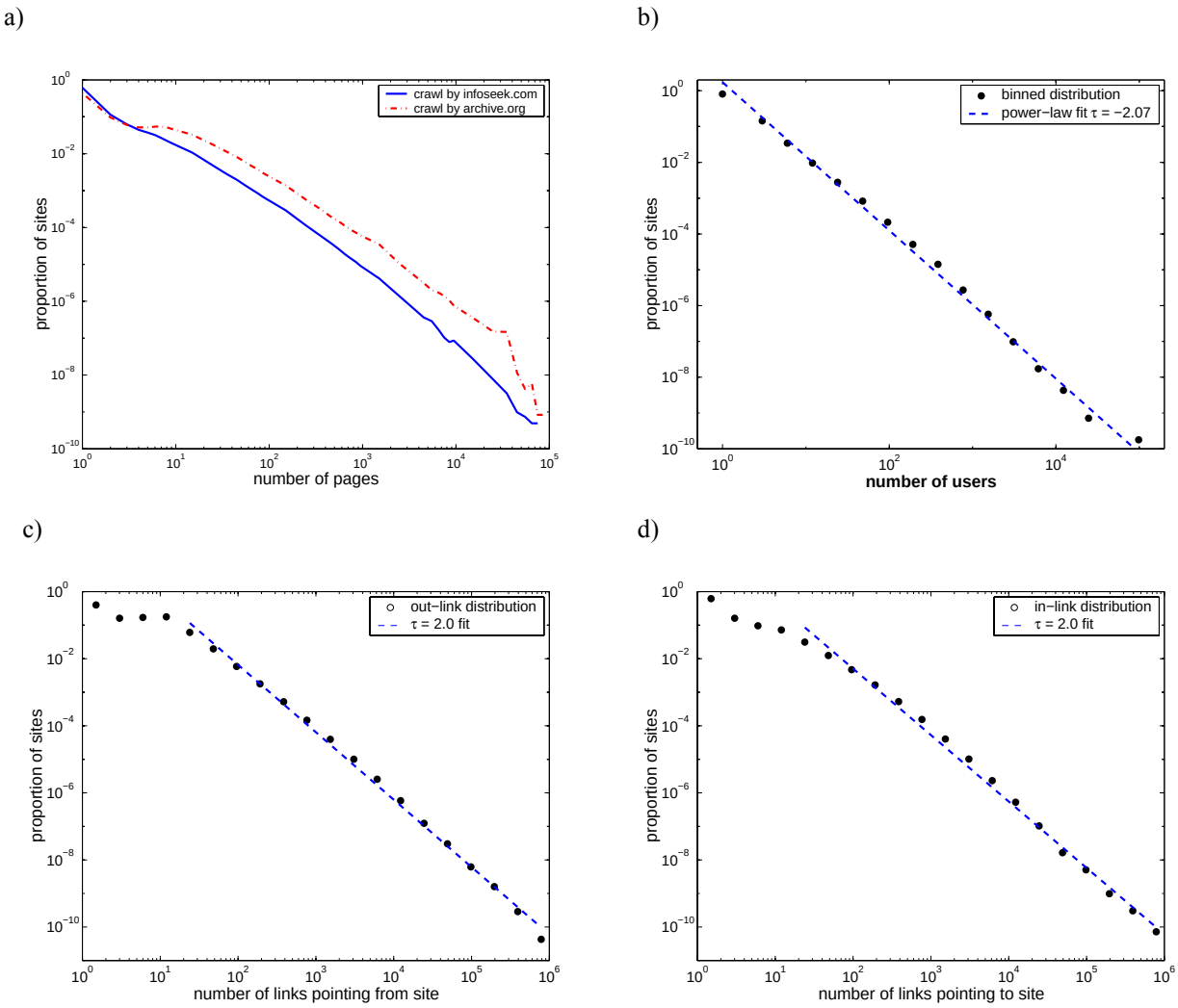

Figure 1. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links, measured in 1997.

Although the distributions plotted above are given in terms of the probability density function (PDF), they can also be easily recast in terms of Zipf's ranked distribution. In fact, any purely power-law probability density function will yield a Zipf ranked distribution as follows:

Let the PDF be $p(x) = Cx^{-\tau}$. Then the probability that a website is of size y or lager $P(x > y) = \sum_{y}^{\infty} Cx^{-\tau} \approx Ay^{-\tau+1}$, C and A are constants. If there are N websites total, the expected number of sites greater than N is given by $r = NAy^{\tau-1}$. Solving for y, we find that the size of the r^{th} ranked variable is proportional to $r^{-\frac{1}{\tau-1}} = r^{-\alpha}$, α being the Zipf rank exponent. While the PDF emphasizes the count of small elements, the ranked distribution emphasizes the size of the largest ones.

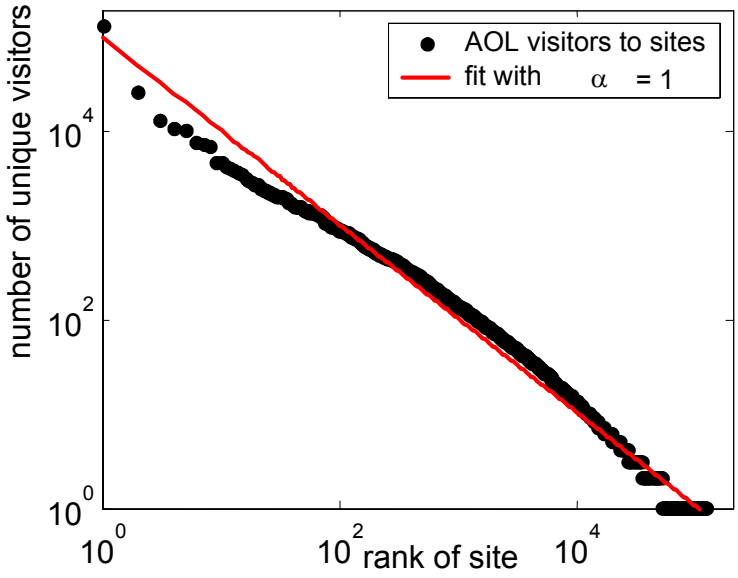

Figure 2. Sites ranked by the number of unique AOL visitors they received Dec. 1, 1997. AOL (America Online) is the largest Internet service provider in the United States. The fit is a Zipf distribution $n_r \sim r^{-1}$

Figure 2, for example, shows the ranked distribution of visitors to web sites corresponding to the PDF plotted in Figure 1b. The distribution shows mild concavity and a ranked exponent of 1: Zipf's law. As Table 1 shows, a small number of sites such as Yahoo are extremely popular and capture a disproportionate amount of traffic.

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

%sites	%user volume
0.1	32.36
1	55.63
5	74.81
10	82.26
50	94.92

A Growth Model

The pervasiveness of Zipf distributions on the Internet can be explained by an intuitive growth model (Huberman 1999) that incorporates three simple assumptions. Let us formulate the argu-ment in terms of the number of web pages hosted on a website. Similar arguments can be applied just as easily to the number of visitors or links. The first assumption is that of proportional growth or preferential attachment, i.e. the number of pages added to or removed from the site is proportional to the number of pages already present. For example, a site with a million pages might have a whole team of webmasters or generate its content automatically. It could easily gain or shed a several thousand pages on any given day. On the other hand, it would be surprising, but not impossible, for a site with only a handful of pages to suddenly add a thousand more.

This multiplicative stochastic growth process yields a lognormal distribution in the number of pages at a site after a fixed period of time. However, the World Wide Web is anything but fixed. Its first decade was a period of rapid growth, with sites appearing at an exponential rate. It so happens that when one computes an exponentially weighted mixture of lognormals one obtains a power-law distribution exactly!

While the exponential growth in the number of websites and their stochastic addition of pages alone can produce power law distributions, a key ingredient is still missing. For in spite of the random nature of the growth, if one were taking a mixture of lognormals depending only on a time variable, one would expect that the sites established early would have grown to greater sizes than recently founded ones. However, studies have found only weak correlation between the size of a site and its age (equivalently some very popular sites were founded more recently, while sites present at the very start of the Internet boom did not necessarily acquire a wide audience). The missing assumption is that sites can grow at different rates, depending on the type of content and interest that they generate. Incorporating variability in growth rates again yields power law distributions with varying exponents. The greater the difference in growth rates among sites, the lower the exponent τ, which means that the inequality in site sizes increases. In summary, a very simple assumption of stochastic multiplicative growth, combined with the fact that sites appear at different times and/or grow at different rates, leads to an explanation for the scale free behavior so prevalent on the Web (Huberman 2001).

Caching

Computer scientists have gone beyond observations and explanations of Zipf's law to apply it to the design of content delivery on the Internet. A problem Internet service providers (ISP's) face is devising ways to support rapidly growing web traffic while maintaining quality of service in the form of fast response time for file requests. In order to quickly satisfy users' request for web content, ISP's utilize caching, whereby frequently used files are copied and stored "near" to users on the network. It is important to note, however, that the effectiveness of caching relies heavily on the existence of Zipf's law.

Let's say that there is a web server in the United States serving a page that is extremely popular in a town in Europe. In the absence of caching, every time someone in that town requests the page, their request travels across the Atlantic, reaches the US server, which in turn sends the page back across the Atlantic to the requester in Europe.

To avoid sending unnecessary cross-Atlantic requests, the Internet service provider serving the European town can place a proxy server near the town. The proxy server's role is to accept requests from the users and forward them on their behalf. Now, when the first user requests the document, the request goes to the proxy. If the proxy cache does not contain the document, it makes a request to the US server, which replies to the proxy. The proxy then sends the file to the requesting user, and stores the file locally in a cache. When additional users send their requests for the file to the proxy, the proxy can serve them the file directly from its cache, without having to contact the webserver in the US. Of course, files that are updated frequently, such as the front page of a news site, have an expiration time after which the file is considered 'stale'. The cache uses the expiration time to determine when to request a new version from the origin server.

Caching has two advantages. First, since the requests are served immediately from the cache, the response time can be significantly faster than contacting the origin server. Second, caching conserves bandwidth by avoiding redundant transfers along remote internet links. The benefits of caching are confirmed by its wide use by ISPs. They benefit because they are able

to reduce the amount of inter-ISP traffic that they have to pay for. Caching by proxies benefits not only the ISPs and the users, but also the websites holding the original content. Their content reaches the users more quickly and they avoid being overloaded themselves by two many direct requests.

However, since any cache has a finite size, it is impossible for the cache to store all of the files users are requesting. Here Zipf's law comes into play. Several studies (Cunha 1995, Breslau 1999) have found the popularity of files requested follows a Zipf distribution. Hence, the cache need only store the most frequently requested files in order to satisfy a large fraction of users requests.

Networks

The Internet is comprised of networks on many levels, and some of the most exciting consequences of Zipf's law have been discovered in this area. The World Wide Web is a network of interconnected webpages and the Internet backbone is a physical network used to transmit data, including web pages, in the form of packets, from one location to another. Measurements on both the World Wide Web (Adamic 1999, Jeong 1999) and the Internet backbone (Faloutsos 1999, Albert 2000) have shown that they differ significantly from the classic Erdös-Rényi model of random graphs (Erdös 1960). While the traditional Erdös-Rényi model has a Poisson node degree distribution, with most nodes having a characteristic number of links, these networks approximately follow a Zipf or scale-free degree distribution $p(k) \sim k^{-\tau}$, where k is the node degree, and τ is the scale-free exponent. To account for these observations, new random graph growth models have been developed that rely on the above mentioned idea of preferential attachment (Albert 2002).

The scale free degree distribution of the Internet backbone, shown in Figure 3, implies that some nodes in the network maintain a large number of connections (proportional to the total size of the network), while for the most part nodes have just one or two connections. This is a two edged sword when it comes to resilience of the network. It means that if a node fails at random, it is most likely one with very few connections, and its failure won't affect the performance of the network overall. However, if one were to specifically target just a few of the high degree nodes, the network could be adversely affected. Because many routes pass through the high degree nodes, their removal would require rerouting through longer and less optimal paths. Once a sufficient number of high degree nodes are removed, the network itself can become fragmented, without a way to communicate from one location to another.

On a different level, one of the recent developments in the use of the Internet has been the emergence of peer-to-peer (P2P) networks. These networks are used by millions of users daily to exchange a variety of files directly with one another. Examples of P2P networks include Napster, Gnutella, and Kazaa. Although Napster was immensely popular, it was forced to shut down by the recording industry over concerns that users were trading copyrighted music files. Part of the reason Napster could so easily be shut down is that it operated with a central server. The users would report which files they were sharing to the central server, and when they looked for additional files, they would query the central server to locate other users who had those files.

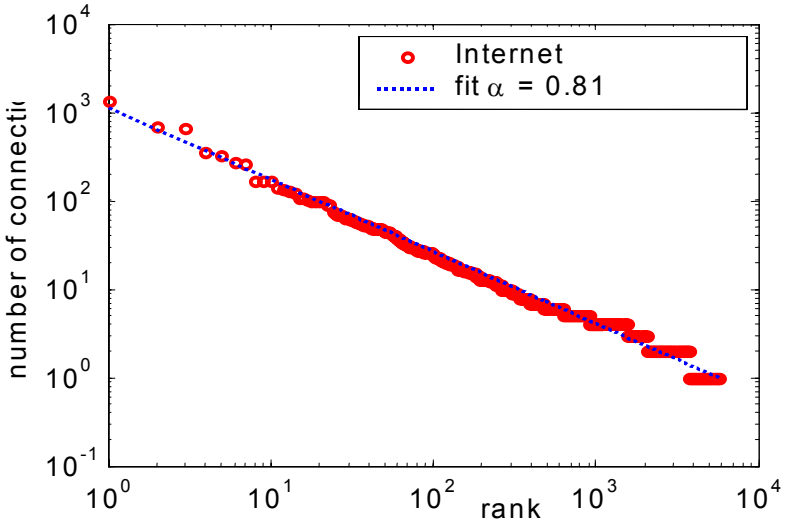

Figure 3. The connectivity of the internet backbone at the autonomous system (AS level). Each AS is itself a network corresponding to a single ISP, business entity or educational institution.

Having learned from Napster's troubles, current peer-to-peer networks tend to be decentralized. That is, nodes connect directly to one another rather than to a central server. The distribution in the number of computers a computer has connections to is a Zipf distribution (recently it has shifted into a two-sided Zipf distribution, with a shallower exponent for the high degree nodes and a steeper exponent for the low degree ones) (Ripeanu 2002). The presence of Zipf's law has implications for the search strategies used in P2P networks. Currently, most P2P networks use a broadcast method of locating files. Because there is no central server that queries can be sent to, each node broadcasts the query to all of its neighbors who in turn broadcast to all of their neighbors, out to some fixed distance from the originating node. As one can well imagine, the network can become quite congested with broadcasted queries. Recent research has shown, however, that routing queries to the high degree nodes may provide a degree of congestion relief, while maintaining a short response time (Adamic 2001). Again, knowledge of Zipf's law in the connectivity distribution has offered a solution to an Internet communication problem.

Finally, it has been shown that scale-free networks are more susceptible to viruses than networks with a more even degree distribution. Namely, a virus spreading in a random network needs to surpass a threshold of infectiousness in order not to die out. However, if the network has a Zipf degree distribution, the virus can persist in the network indefinitely, no matter what level of its infectiousness (Pastor-Satarros 2002).

Both email (Ebel 2002) and instant messaging networks (Smith 2002) have been shown to be scale free. Some individuals have a large number of email contacts but most individuals would keep only a few addresses in their contact lists. This wide variance in the connectivity of electronic communication reflects the different degrees of communicativeness in people and their different roles at work and in society overall. Over the past few years, email viruses have plagued the Internet, no doubt facilitated by hubs, or individuals with large contact lists. An email virus can be passed on as an attachment in email messages. Once the attachment is opened, the virus can activate and cause the email program to send numerous infected emails to email addresses from the person's contact list. The "I love you" email virus alone infected over 500,000 individual systems in May of 2000[2]. Sometimes the sheer quantity of viral

[2]Source: CERT® Advisory, http://www.cert.org/advisories/CA-2000-04.html

email can affect the Internet's performance. But just as hubs (individuals or computers with many contacts) can facilitate the spreading of a virus, they can also aid in preventing their spread. Carefully im-munizing the hubs could stop the virus in its tracks.

Conclusions

On the Internet, Zipf's law appears to be the rule rather than the exception. It is present at the level of routers transmitting data from one geographic location to another and in the content of the World Wide Web. It is also present at the social and economic level, in how individuals select the websites they visit and form peer-to-peer communities. The ubiquitous nature of Zipf's law in cyberspace has lead to a deeper understanding of Internet phenomena, and has consequently influenced the way in which it has evolved.

Acknowledgements
We would like to thank T.J. Giuli, Eytan Adar and Rajan Lukose for their comments and suggestions.

References

Adamic, L.A. (1999). The Small World Web, Proceedings of ECDL'99. *Lecture Notes in Computer Science 1696, 443-452*. Berlin: Springer.

Adamic, L.A., Lukose, R.M., Puniyani, A.R., and Huberman, B.A. (2001). Search in Power-Law Networks. *Physical Review E 64: 046135*.

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Attack and error tolerance of complex networks. *Nature 406, 378*.

Albert R. and Barabasi A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics 74, 47-94*.

Breslau L. et al. (1999). Web Caching and Zipf-like Distributions: Evidence and Implications. *Proceedings of INFOCOM '99, 126-134*.

Cunha, C.R., Bestavros A., and Crovella, M.E. (1995). Characteristics of WWW Client-based Traces". *Technical Report TR-95-010*. Boston University Computer Science Department.

Ebel, H, Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *cond-mat/0201476*.

Erdös, P. and Rényi, A. (1960). On the Evolution of Random Graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17-61*.

Faloutsos M., Faloutsos P. and Faloutsos C. (1999). On Power-Law Relationships of the Internet Topology. *Proceedings of ACM SIGCOMM '99, 251-262*.

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

Huberman, B. and Adamic, L. (1999). Growth Dynamics of the World Wide Web. *Nature 401, 131*.

Huberman, B. A. (2001). *The Laws of the Web.* The MIT Press.

Pareto, V. (1896). *Cours d'Economie Politique*. Genève: Droz.

Pastor-Satarros, R. and Vespignani, A. (2001). Epidemic spreading in Scale Free Networks. *Physical Review Letters 86, 3200*.

Ripeanu M., Foster, I. and Iamnitchi A. (2002). Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal special issue on peer-to-peer networking, vol. 6(1), 50-57.*

Smith, R.D. (2002). Instant Messaging as a Scale-Free Network. *cond-mat/0206378*.

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA.: Harvard University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.