

A Tale of the Tails: Power-laws in Internet Measurements

Aniket Mahanti¹, Niklas Carlsson², Anirban Mahanti³, Martin Arlitt^{4,5}, Carey Williamson⁵

¹University of Auckland, Auckland, New Zealand

²Linköping University, Linköping, Sweden

³NICTA, Eveleigh, New South Wales, Australia

⁴HP Labs, Palo Alto, California, U.S.A.

⁵University of Calgary, Calgary, Alberta, Canada

Abstract

Power-laws are ubiquitous in the Internet and its applications. This tutorial presents a review of power-laws with emphasis on observations from Internet measurements. First, we introduce power-laws and describe two commonly observed power-law distributions, namely the Pareto and Zipf distributions. Two frequently occurring terms associated with these distributions, specifically *heavy tails* and *long tails*, are also discussed. Second, the *preferential attachment* model, which is a widely used model for generating power-law graph structures, is reviewed. Subsequently, we present several examples of Internet workload properties that exhibit power-law behaviour. Finally, we explore several implications of power-laws in computer networks. Using examples from past and present, we review how researchers have studied and exploited power-law properties. We observe that despite the challenges posed, power-laws have been effectively leveraged by researchers to improve the design and performance of Internet-based systems.

I. INTRODUCTION

Power-laws are observed in many naturally occurring phenomena (e.g., earthquakes, precipitation, topography), as well as in many human-related behaviours (e.g., citations, urban population, wealth). Power-laws have been observed in many aspects of information systems, including software systems and computer networks. Early examples include memory referencing behaviour in virtual memory systems, database queries, and file usage patterns in file systems. More recently, several characteristics of the Internet and the Web have also been claimed to exhibit power-law characteristics such as the number of visitors to a Web site [1], the number of hyperlinks to a Web page [1], the sizes of Web objects [2], the number of links to routers on the Internet [3], and the number of friends of users on online social networks [4].

Power-law properties typically appear in high variance distributions wherein observations span many orders of magnitude, particularly if there is a pronounced skew of the distribution. Compared to exponential distribution which has been widely used in mathematically modeling telecommunication systems, power-law distributions decay more slowly. Presence of power-laws indicate that arbitrarily large values can occur with a non-negligible probability, and therefore, rather than ignoring these extreme values as “outliers” it is useful to study their statistical prevalence, if sufficiently many such samples are present in a large dataset.

The apparent abundance of power-law distributions in computing (and other) literature has drawn significant interest on understanding the origin and the implications of these power-law properties. For example, it has led to improved Web caching policies, better traffic routing and load balancing techniques, smarter search schemes, and sophisticated network topology generators. The ubiquity of power-laws has also evinced interest in developing models that generate power-law distribution, often with the goal of gaining insights on the processes behind the occurrence of power-laws. The presence of power-law and the accuracy of these models have been debated [5]. This debate has been fuelled by the discovery of measurement artifacts and the difficulty of deploying proper sampling techniques in large-scale systems. Due to the presence of many other highly skewed distributions, another active discussion topic is how to best identify the presence of power-laws from measurement data [2].

In this article, we review power-law relationships reported in the Internet measurement literature. We define power-law relationships in general, discuss approaches to identifying the presence of power-laws, and discuss two commonly used power-laws, namely, the Pareto and Zipf distributions. We provide examples of Internet

measurements that suggest power-law behaviour and discuss several examples from the literature highlighting how researchers have leveraged power-laws in an effective way to improve the design and performance of Internet-based systems and applications.

II. POWER-LAW RELATIONSHIPS

A power function is a scale-invariant function, $f(x)$, of the form $f(x) = \alpha x^{-\eta}$, where α and η are positive constants, and η is called the scaling exponent. Taking logarithms on both sides of the power function produces $\log(f(x)) = -\eta \log(x) + \log(\alpha)$. This expression exhibits a linear relationship with slope $-\eta$ and y-intercept $\log(\alpha)$. When plotted on a log-log scale, the function appears as a straight line. This observance is often considered as the distinctive feature of a power-law relationship.

In the computing literature a dataset is often said to follow a power-law if for large values the distribution follows the power function. More formally, a distribution is considered power-law if [2]:

$$f(x) \sim x^{-\eta} \quad (1)$$

where \sim is used to indicate asymptotic proportionality; i.e., $\frac{f(x)}{x^{-\eta}} \rightarrow c$, for some constant $c > 0$, when $x \rightarrow \infty$. In other words, the power-law distribution exhibits the power function for large values of x , typically referred to as the *tail of the distribution*.

A. Pareto Distribution

One commonly observed power-law distribution in Internet traffic measurements is the Pareto distribution. A random variable X is said to follow a Pareto distribution if the complementary cumulative distribution function (CCDF) indicating the probability of occurrence of an event being greater than x is inversely proportional to a power of x ; i.e., $P[X > x] \propto x^{-\kappa}$, where κ is called the shape parameter. A property of power functions is that the integral of a power function is also a power function. Due to this property, it is easy to show that the Pareto distribution (which itself has a power-law shape) and the power-law distribution is related by $\kappa = \eta - 1$.

Figure 1 illustrates the Pareto, exponential, and lognormal distributions. Figure 1(a) shows the probability distribution function. Figure 1(b) shows the CCDF on doubly logarithmic scales. We used the following shape parameters: 1.67 for Pareto, 0.96 for exponential, and 0.98 for lognormal. We note that the tail of the Pareto distribution gradually tapers off when compared to the exponential distribution. Note that in Figure 1(b), the lognormal distribution appears to exhibit linear relationship. In fact, there has been some debate of how to best determine whether a dataset follows lognormal, power-law, or other related distributions [1], [2]. In many cases, it is indeed difficult to ascertain whether or not a distribution is power-law, unless we observe a straight line across several orders of magnitude on a log-log scale. These debates have also resulted in development of more sophisticated methods for identification of power-laws [1].

B. Zipf Distribution

Another classical example of a power-law is the Zipf distribution, which was first used for modeling word frequencies in written texts, but has since been used to model the skewed popularities for library books, movies rentals, and Web objects. The Zipf distribution is a discrete distribution, defined in the rank-frequency domain by Zipf's law, which states that when items are ranked (\mathbb{R}) in descending order of their popularities, then the frequency (\mathbb{F}) of the item is inversely proportional to the rank of the item:

$$\mathbb{F} \propto \mathbb{R}^{-\theta} \quad (2)$$

The Zipf distribution exhibits a straight line with slope $-\theta$ on a log-log rank-frequency plot. The value of $\theta \approx 1$ for a pure Zipf distribution, but other values are possible while still exhibiting a straight-line behaviour. Degenerate forms of the Zipf distribution, in which the behaviour is piecewise linear, or only linear for a portion of the plot, are also often seen in Internet measurements. For example, the popularity of files in a peer-to-peer file sharing system have been found to be Zipf-like with deviation from the expected straight line for the most popular files arguably because of users' "fetch-at-most-once" approach to file sharing [6].

The Zipf distribution may be considered to be a discrete interpretation of the Pareto distribution. It can be represented by transforming the axes of the Pareto distribution. Thus, the Zipf distribution can be written as a Pareto distribution as follows: $\mathbb{R} \propto \mathbb{F}^{\frac{-1}{\theta}}$. To summarize, the Pareto, power-law, and Zipf parameters are related as:

$$\kappa = \eta - 1 = \frac{1}{\theta} \quad (3)$$

The Zipf distribution has a strong skew of references to a small but highly popular set of items. For example, it is not uncommon for a small subset of the items (e.g., 10-20%) to account for most of the referencing activity (e.g., 80-90%). The exact tradeoff depends on the shape parameter. In general, in many empirical studies similar skews have been observed as shown for example in Figure 2 for hosts making requests to a Web server. This phenomenon is often referred to as *Pareto's Law*, the *Pareto Principle*, or the 80/20 rule in the literature.

C. Heavy-tailed Distributions

Typical empirical distributions from Internet measurements can be divided into two parts: the *body* (small to medium-sized values that are responsible for much of the distribution) and the *tail* (large-sized values that are responsible for the rest). A probability distribution is said to have a heavy tail if the tail is not exponentially bounded. As illustrated in Figure 1(b), the Pareto distribution is a heavy-tailed probability distribution. A tail can be Pareto distributed (and heavy-tailed) even if the body of a distribution does not follow the power-law distribution. Such distributions are analyzed by looking at the tail of the distribution. Heavy tails are subject of interest because they highlight the presence of large-sized values. Change in occurrence of these (less frequent) large-sized values affects the distribution more than the change in the (abundant) small-sized values. This has an impact on modelling of the empirical distribution.

D. The Long Tail

The *long tail* is a manifestation of power-law relationships. Long tail exemplifies the statistical property that there are many more low frequency events compared to a Gaussian (Normal) distribution. For example, it has been argued that keywords used for searches often have a long tail. This means that if we order the keywords, from most popular to the least popular, based on how many times a keyword was used, we would find that there are only a few keywords that are often used and that there exists a very long list of infrequently used keywords. Since the list of keywords is very large, these infrequently occurring keywords could together account for a large fraction of keyword searches seen by a search engine. This term came into popular parlance from an article written by Chris Anderson in WIRED magazine (October, 2004) where he argued that online businesses such as Amazon, eBay, and Netflix have successfully leveraged the long tail. Anderson argued that these online businesses carry a wide variety of products, each of which may appeal to only a few customers. This is in contrast to standard retailers that mostly offer popular items because they are restricted by the size of their store inventory. Figure 3 shows an illustration of the long tail phenomena; the niche products sold would be in the green region or the long tail of the sales popularity distribution. As more sales are derived from the tail region, the body of the distribution becomes smaller. We note that the large number of items and their low individual popularity poses some technical challenges as discussed later in this article.

E. Keeping Track of the Tails

The observant reader may have noticed that the long and heavy tails associated with the power-law distribution, as described above, refer to two different ends of the distribution. This is often forgotten and simply discussed as *the tail*. While both the Pareto and Zipf distributions are power-law, we note that the way these distributions are plotted (and defined) often focuses on two different parts of the distributions. In particular, the tail of the Pareto probability distribution refers to the rare (but probable) occurrences of events with high values, whereas the tail of the Zipf distribution refers to the many occurrences of events with small values. This subtle but important difference is illustrated in Figure 4. Here, we note that the body of one distribution makes up for the tail of the other distribution, and vice versa. What further confuses this discussion is that a subclass of heavy-tailed probability distributions, defined in the probability domain, sometimes is referred to as long-tailed. This is in sharp contrast to the long tails referred to in the popular literature, which typically refers to the rank-frequency domain.

III. POWER-LAW EXAMPLES

We review some power-law examples from Internet measurements. One widely reported example of power-laws is Web objects access frequencies [7]. Another widely reported example is the Web-object size distribution, which has been shown to follow the Pareto distribution [2]. There has also been some debate as to whether Web-object sizes are power-law or follow other related distributions [2].

More recently, studies of YouTube video popularity (e.g., [8]) have found that video popularity both at an edge network and as observed by the YouTube servers appears to follow Zipf-like distributions. There are, however, noticeable differences at the tail of the distribution. Power-law characteristics have also been identified for other user-generated video sharing services, with the number of short-term video views exhibiting power-law characteristics, while long-term video views being better modelled using a power-law distribution with an exponential cutoff.

Other works have analyzed Internet TV workloads and found that the popularity of TV channels can be approximated using a Zipf-like distribution [9]. This observation highlights user proclivity towards watching the same channel. The channel popularity distribution followed the Pareto principle, with the top 10% of channels accounted for nearly 80% of viewers. While the audience demographics changed, the Pareto principle was found to be consistently true through different times of the day.

Various Internet-related power-law structures have also been identified. One such example is the number of (online) friends per user in online social networks, which has been shown to follow power-law distributions [4]. Another interesting observation for these networks is that there typically is a highly connected (core) group of users. This significantly reduces the number of friend-of-friends needed to connect to arbitrary people. Similar observations have been made for real-world networks (e.g., the Web and offline social networks). For these networks, the highly connected core allows for efficient dissemination of information and data.

IV. A GENERATIVE MODEL FOR POWER-LAWS

Despite being widely observed, the origin of power-laws is an open problem and an active discussion topic. Generative models have been developed in order to understand the underlying processes that cause the observed power-laws to occur. The *preferential attachment* (or “rich-get-richer”) model [10] is one particularly popular generative model although other models have also been developed [2]. The preferential attachment model has received much attention in the context of graph structures in which the vertices have a power-law degree distribution.

As an example, consider Web pages and the hyperlinks among them. The Web pages may be viewed as vertices of a graph and the hyperlinks as directed edges between the vertices. A simple preferential attachment model is as follows. Suppose that we begin with a single page with a hyperlink to itself. At each time step, a new page is created, and this page is assumed to create a new hyperlink. The new link is formed to one of the existing pages, chosen uniformly at random with probability $p < 1$ and chosen proportionally to the number of incoming links with probability $1 - p$. Iterating this process for many vertex additions generates a power-law graph.

Figure 5 shows a comparative illustration a power-law and a random graph, each consisting of 150 vertices. The power-law graph was created using the preferential attachment model, where each vertex creates one outbound edge at a time. The random network is based on the Erdos-Renyi model where the probability that any two vertices are connected have an equal probability p . The figure illustrates how the high-degree vertices in the power-law graph can be critical for good connectivity and may make the network sensitive to attacks (e.g., targeted node elimination). Similar to the power-law structures discussed here, and as observed in various social and physical networks, preferential attachment and rich-get-richer behaviour have been considered as a potential explanation for content popularity and Web server workloads, among many other things.

V. SOME IMPLICATIONS OF POWER-LAWS

A. Web Caching

Effective Web caching relies on the presence of a non-uniform popularity distribution of Web objects and their sizes. Web accesses have been shown to follow Zipf’s law [7]. This property has proved important in the design of Web cache architectures, since it allows designers to calculate approximate cache sizes to achieve desired hit rates. The appropriate cache size along with the appropriate replacement policy could achieve high cache hit rates.

Zipf’s law can be useful for predicting the probability of access of an object. Researchers have found that deploying caching hierarchies may be undesirable as they suffer from diminishing returns on hit rates. This is

because the objects most worth caching are cached multiple times at levels closest to the users. Furthermore, deeper caching hierarchies may increase document access latencies. Content delivery networks (CDNs) can take advantage of more proactive delivery schemes.

B. Search Schemes

Power-law node connectivity distribution has helped improve search in the Web. As described in the previous section, the Web can be considered a directed graph of Web pages and hyperlinks. Measurements of the structure of the Web graph have identified the presence of key nodes. On any given topic, some Web pages may have a high out-degree and others may have high in-degree [11]. Web pages that work as information aggregators typically have high out-degree and are referred to as hubs, whereas Web pages with high in-degree are typically referred to as authorities. The idea of hubs and authorities was used for ranking Web pages in one of the early search mechanisms [11]. Similar approaches have also been exploited for searching in online social networks and peer-to-peer (P2P) systems.

C. The Long Tail and Business Practices

Online businesses have taken advantage of the mildly popular items to increase sale volumes. Both Amazon and Netflix have developed smarter recommendation schemes that expose users to items of personal interest based on purchasing history of the user and other users with similar interest. This allows them to potentially recommend niche content to its customer. This is in contrast to search engines that use popularity measures to rank Web pages, and users formulate their decisions based on the top-ranked pages.

D. The Long Tail and System Design

The long tails observed in power-law can impact system efficiency. For example, a new object storage system has been designed to optimize Facebook's Photos application with the aim of serving the long tail (requests for less popular photos) [12]. These optimizations are important, as requests from the long tail accounted for a significant amount of their traffic and since the low individual request rates and high miss rates caused most of these requests being served from the main photo storage server, rather than by Facebook's CDN.

The long tail also poses challenges in the context of peer-to-peer file sharing systems such as BitTorrent. In particular, the mildly popular files may not have sufficient popularity to have an active torrent. One approach to improve the resulting file-availability problem is to group multiple files into bundles such that the bundles become sustainable torrents [13]. There exist both static and dynamic bundling approaches, including adaptive bundling policies that take the current file popularities into consideration.

E. Measurement Issues

Large-scale graphs such as the Internet topology or online social networks present many measurement challenges. Some of these challenges have helped fuel the debate about the authenticity of power-law nature of Internet graphs. For example, the scale of these networks often limits the fraction of the network that can be captured. It has been suggested that the bias in the partial crawling of online social graphs, which results in only a subset of the graph being captured, can underestimate the power-law scaling exponent [4]. To alleviate this problem, recently a multi-dimensional random walk algorithm [14] was proposed, which captured dynamic real world networks better and reduced scaling exponent estimation errors.

Other researchers have suggested that incomplete measurement data (using traceroute, for example) result in missing large number of Internet topology connections causing it to exhibit power-law behaviour. The debate regarding whether the power-law degree distribution is an integral property of the Internet topology or an artifact of biased sampling has attracted significant attention [3], [5], [15]. Such deliberations point towards the challenges in measurement and accentuates the need for appropriate sampling techniques.

The hazards of improper sampling have also been investigated and discussed in the contexts of access patterns and other workloads which have been argued to possess power-law characteristics. Ideally, content popularity measurements should be based on probability sampling methods with known biases, such that the underlying distribution can be recreated based on the samples. Unfortunately, probability sampling is often difficult to apply

to large-scale dynamic systems. The impact of sampling methods is embodied by the differences in the content popularity distribution observed when applying different definitions of popularity, sampling methods, or the length of the measurement interval.

VI. CONCLUSIONS

Power-laws are apparent in several aspects of Internet measurements. Power-laws pose some challenges, however, they have been efficaciously leveraged by researchers in design and optimization of Internet-based systems. We described a simple generative power-law model, although several other models exist in the literature that can lead to power-law behaviour. We touched upon the ongoing debate regarding the authenticity of power-law nature of some Internet attributes. Nevertheless, these deliberations point to the significance of power-laws in computer networks, which cannot be ignored.

REFERENCES

- [1] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.
- [2] M. Mitzenmacher, “A Brief History of Generative Models for Power Law and Lognormal Distributions,” *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, 2003.
- [3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On Power-law Relationships of the Internet Topology,” in *Proc. ACM SIGCOMM Conference*, Cambridge, U.S.A., 1999.
- [4] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and Analysis of Online Social Networks,” in *Proc. ACM SIGCOMM International Measurement Conference (IMC)*, San Diego, U.S.A., 2007.
- [5] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, “The Origin of Power Laws in Internet Topologies Revisited,” in *Proc. IEEE INFOCOM Conference*, New York, U.S.A., 2002.
- [6] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, “Measurement, Modeling, and Analysis of a Peer-to-Peer File-sharing Workload,” in *Proc. ACM Symposium on Operating Systems Principles (SOSP)*, Bolton Landing, U.S.A., 2003.
- [7] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, “Web Caching and Zipf-like Distributions: Evidence and Implications,” in *Proc. IEEE INFOCOM Conference*, New York, U.S.A., 1999.
- [8] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, “Youtube Traffic Characterization: A View from the Edge,” in *Proc. ACM SIGCOMM International Measurement Conference (IMC)*, San Diego, U.S.A., 2007.
- [9] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatrian, “Watching Television Over an IP Network,” in *Proc. ACM SIGCOMM International Measurement Conference (IMC)*, Vouliagmeni, Greece, 2008.
- [10] A.-L. Barabasi and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [11] J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” in *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Philadelphia, U.S.A., 1998.
- [12] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel, “Finding a Needle in Haystack: Facebook’s Photo Storage,” in *Proc. USENIX Symposium on Operating System Design and Implementation (OSDI)*, Vancouver, Canada, 2010.
- [13] D. S. Menasche, A. A. Rocha, B. Li, D. Towsley, and A. Venkataramani, “Content Availability and Bundling in Swarming Systems,” in *Proc. ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, Rome, Italy, 2009.
- [14] B. Ribeiro and D. Towsley, “Estimating and Sampling Graphs with Multidimensional Random Walks,” in *Proc. ACM SIGCOMM International Measurement Conference (IMC)*, Melbourne, Australia, 2010.
- [15] W. Willinger, D. Alderson, and J. Doyle, “Mathematics and the Internet: A Source of Enormous Confusion and Great Potential,” *Notices of the AMS*, vol. 56, no. 5, pp. 586–599, 2009.

PLACE
PHOTO
HERE

Aniket Mahanti is a Lecturer in the Department of Computer Science at the University of Auckland, New Zealand. He holds a B.Sc. (Honours) in Computer Science from the University of New Brunswick, Canada, and an M.Sc. and Ph.D. in Computer Science from the University of Calgary, Canada. His research interests include Internet traffic measurement and performance evaluation.

PLACE
PHOTO
HERE

Niklas Carlsson is an Assistant Professor at Linköping University, Sweden. He received his M.Sc. degree in Engineering Physics from Umeå University, Sweden, and his Ph.D. in Computer Science from the University of Saskatchewan, Canada. He has previously worked as a Postdoctoral Fellow at the University of Saskatchewan, Canada, and as a Research Associate at the University of Calgary, Canada. His research interests are in the areas of design, modeling, characterization, and performance evaluation of distributed systems and networks.

PLACE
PHOTO
HERE

Anirban Mahanti is a Senior Researcher at NICTA, Australia. He received the B.E. degree in Computer Science and Engineering from the Birla Institute of Technology (at Mesra), India, and the M.Sc. and Ph.D. degrees in Computer Science from the University of Saskatchewan, Canada. His research interests are in the areas of network measurements, network protocols, performance evaluation, and distributed systems.

PLACE
PHOTO
HERE

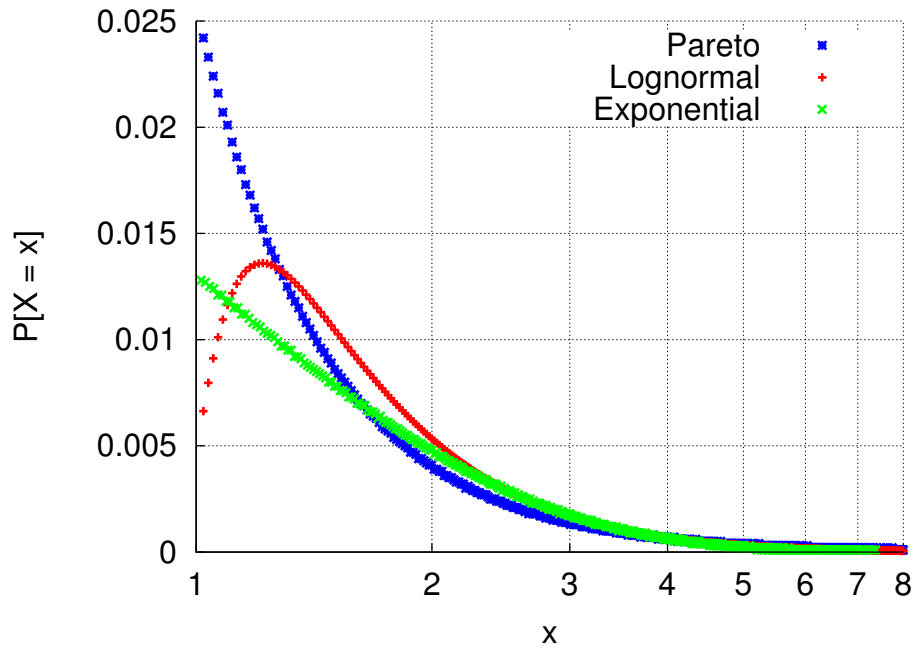
Martin Arlitt is a Senior Research Scientist at Hewlett-Packard Laboratories (HP Labs) in Palo Alto, California, where he has been working since 1997. His general research interests are workload characterization and performance evaluation of distributed computer systems. He is the creator of the ACM SIGMETRICS GreenMetrics workshop, and co-chair of the IEEE Computer Society's Special Technical Community on Sustainable Computing. Since 2001, Martin has been living in Calgary, Alberta, where he is also a Senior Research Associate in the Department of Computer Science at the University of Calgary, Canada.

PLACE
PHOTO
HERE

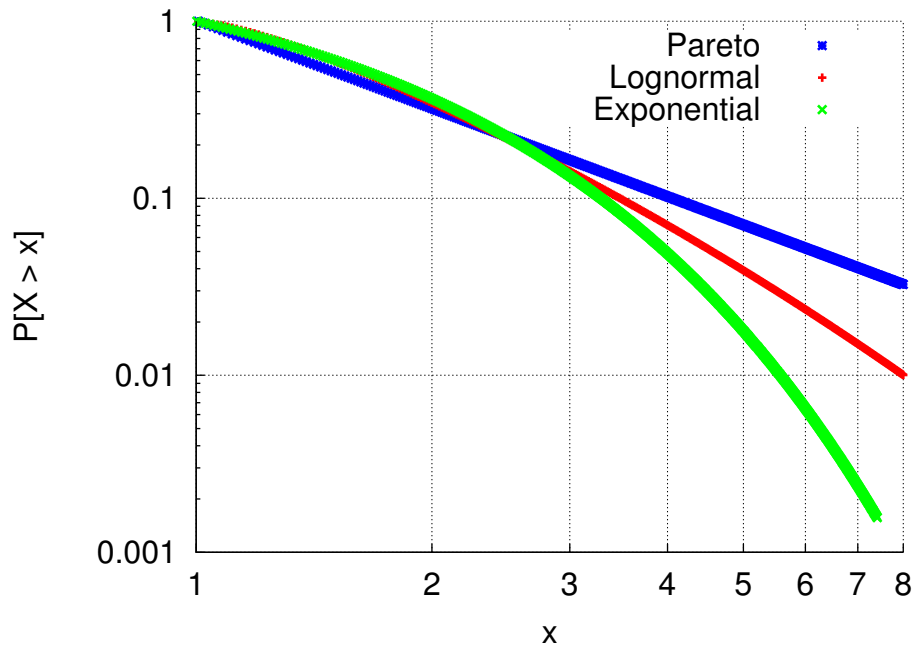
Carey Williamson is a Professor in the Department of Computer Science at the University of Calgary, Canada. He holds a B.Sc. (Honours) in Computer Science from the University of Saskatchewan, Canada, and a Ph.D. in Computer Science from Stanford University, USA. His research interests include Internet protocols, wireless networks, network traffic measurement, network simulation, and Web performance.

LIST OF FIGURES

1	Comparison of the Pareto versus lognormal and exponential distributions. In this example, the Pareto, exponential, and lognormal distributions have shape parameters of 1.67, 0.96, and 0.98, respectively. Note the scale of the axes in the plot; Figure 1(a) uses log-scale on the x axis, while Figure 1(b) uses log-scales on both the x and y axes.	9
2	The Pareto principle: The illustration shows the distribution of requests to the WWW 2007 conference Web server over a one-year period. We observe that the top 10% of the hosts account for 80% of the total Web requests. It exemplifies the Pareto principle, where most of the Web requests are made by few hosts.	10
3	The long tail of item sales: This illustration is based on the proposition by Chris Anderson that sales of niche items (area in green) collectively can earn more revenue than the sale of the few popular items (area in red). As more sales are derived from the tail section, the body of the distribution becomes smaller. This illustration has been adapted from the public domain picture by Hay Kranen available at Wikipedia.org	11
4	Tracking the tails: The illustration shows the distribution of visitors arriving at YouTube from referring Web sites. Figure 4(a) shows that the number of visitors to YouTube is directly proportional to the ordered rank of the referring Web site. Figure 4 (b) shows that there are a few sites that provide a bulk of the referrals to YouTube. The data was collected from Compete.com , which omitted values for referrals that resulted in less than 3,000 visitors arriving at YouTube.	12
5	Comparison of power-law and random graphs: Each graph consists of 150 vertices. A vertex is represented by a red dot and the edge is shown using a solid grey line. The graphs were simulated using the <code>NetworkX</code> package in <code>Python</code> , and visualized using <code>Graphviz</code>	13



(a) Probability Distribution Function



(b) Complementary Cumulative Distribution Function

Fig. 1. Comparison of the Pareto versus lognormal and exponential distributions. In this example, the Pareto, exponential, and lognormal distributions have shape parameters of 1.67, 0.96, and 0.98, respectively. Note the scale of the axes in the plot; Figure 1(a) uses log-scale on the x axis, while Figure 1(b) uses log-scales on both the x and y axes.

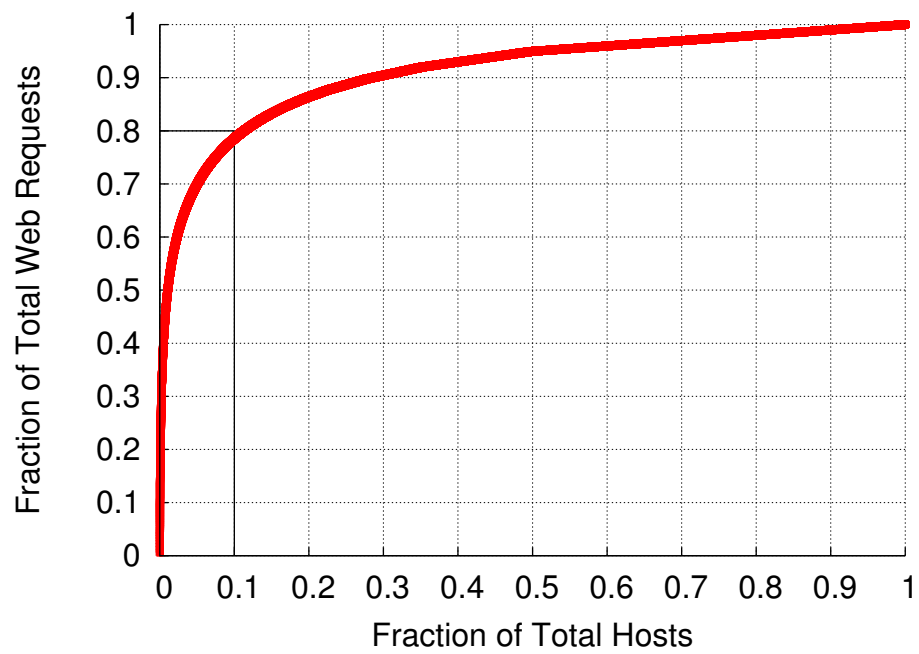


Fig. 2. The Pareto principle: The illustration shows the distribution of requests to the WWW 2007 conference Web server over a one-year period. We observe that the top 10% of the hosts account for 80% of the total Web requests. It exemplifies the Pareto principle, where most of the Web requests are made by few hosts.

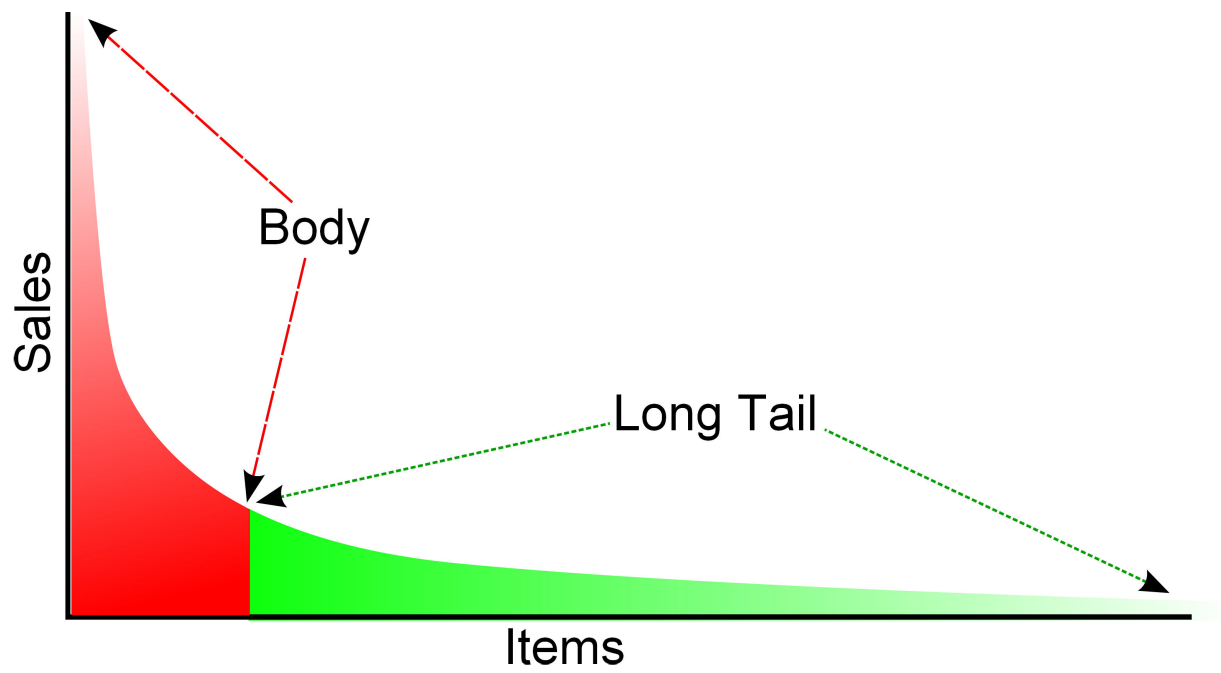
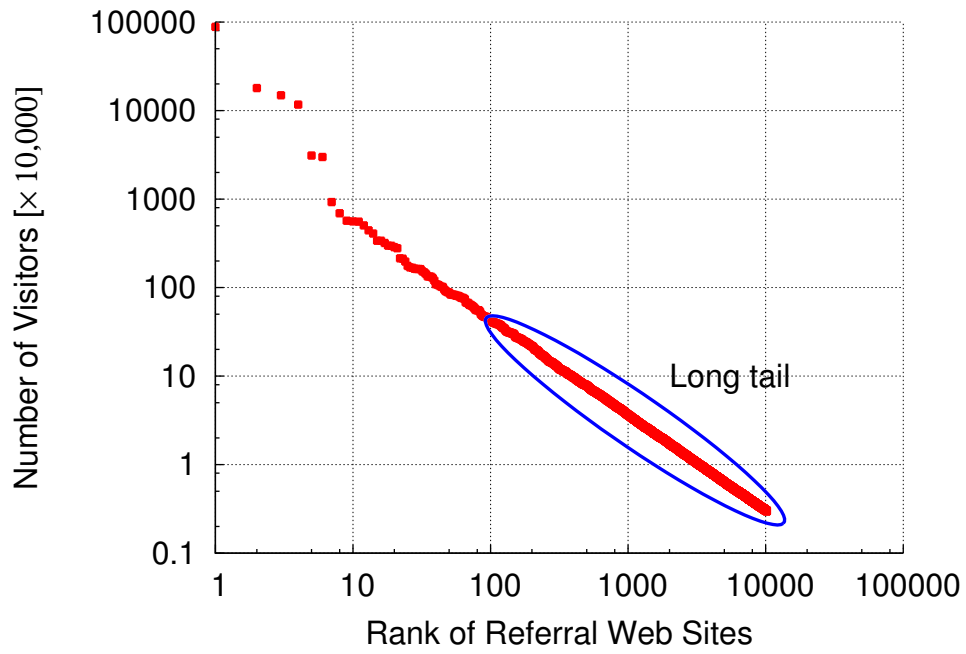
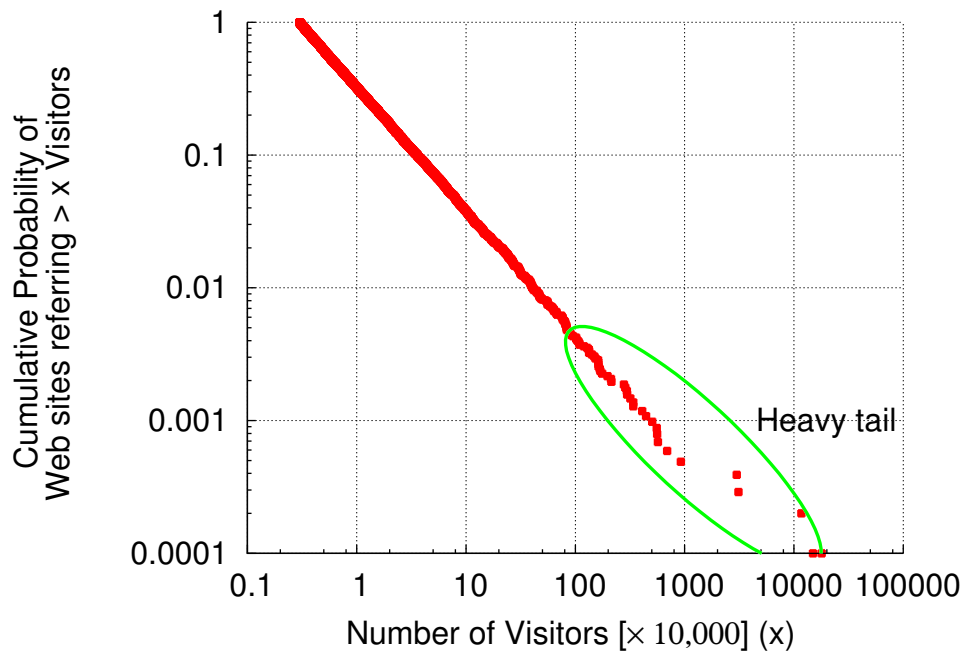


Fig. 3. The long tail of item sales: This illustration is based on the proposition by Chris Anderson that sales of niche items (area in green) collectively can earn more revenue than the sale of the few popular items (area in red). As more sales are derived from the tail section, the body of the distribution becomes smaller. This illustration has been adapted from the public domain picture by Hay Kranen available at Wikipedia.org.

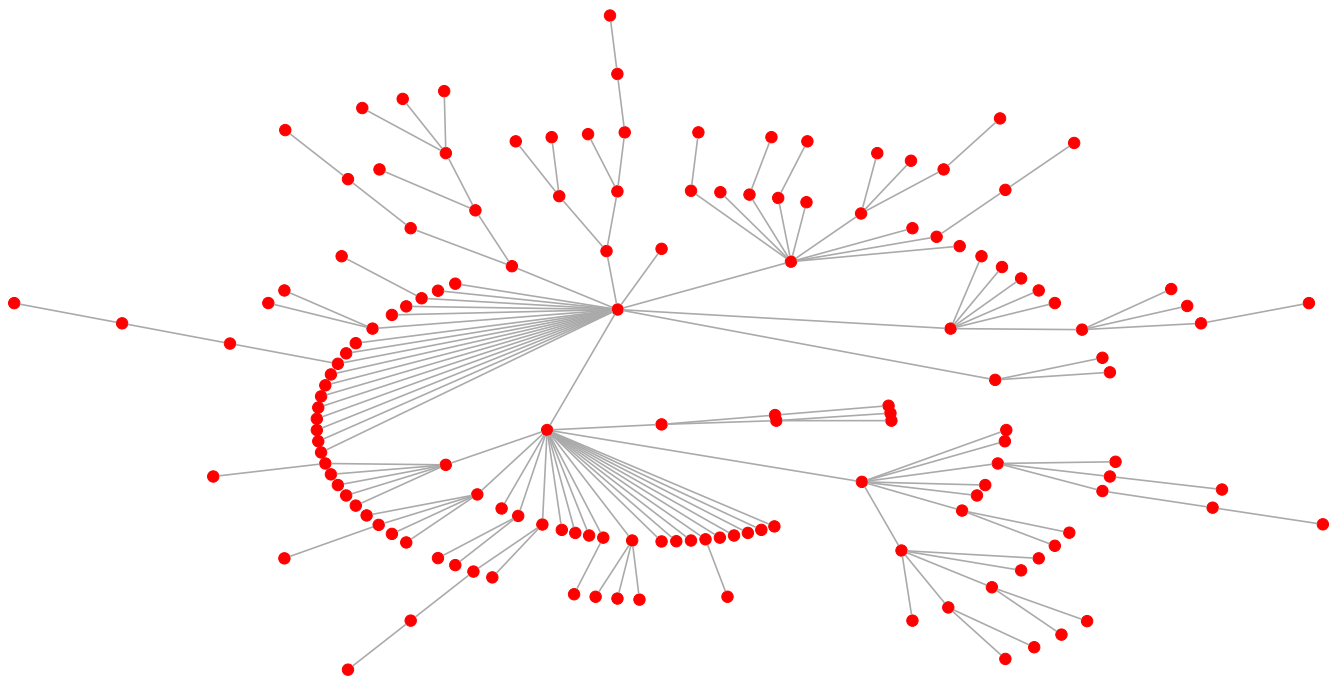


(a) Rank Size Distribution

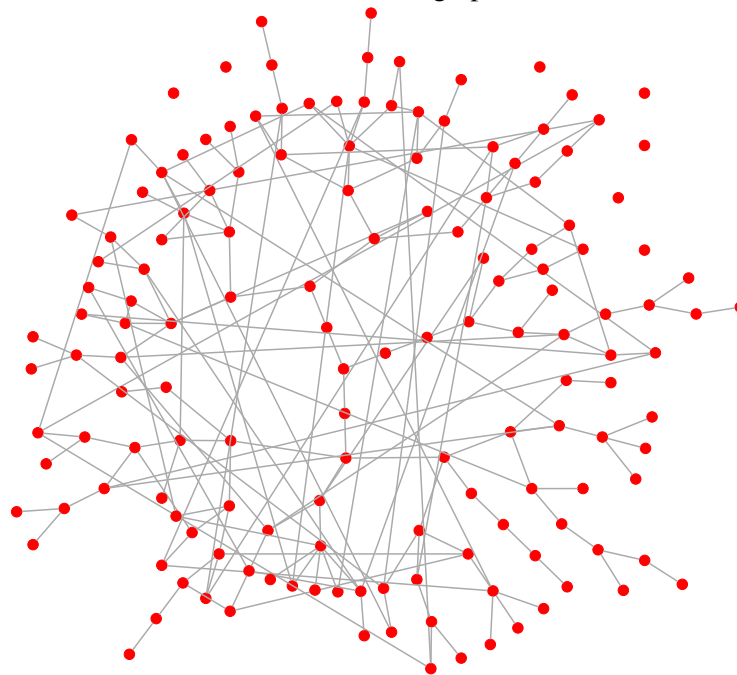


(b) Complementary Cumulative Distribution Function

Fig. 4. Tracking the tails: The illustration shows the distribution of visitors arriving at YouTube from referring Web sites. Figure 4(a) shows that the number of visitors to YouTube is directly proportional to the ordered rank of the referring Web site. Figure 4 (b) shows that there are a few sites that provide a bulk of the referrals to YouTube. The data was collected from `Compete.com`, which omitted values for referrals that resulted in less than 3,000 visitors arriving at YouTube.



(a) Power-law graph



(b) Random graph

Fig. 5. Comparison of power-law and random graphs: Each graph consists of 150 vertices. A vertex is represented by a red dot and the edge is shown using a solid grey line. The graphs were simulated using the `NetworkX` package in `Python`, and visualized using `Graphviz`.