A Brief History of Generative Models for Power Law and Lognormal Distributions

Michael Mitzenmacher*

Abstract

Recently, I became interested in a current debate over whether file size distributions are best modelled by a power law distribution or a a lognormal distribution. In trying to learn enough about these distributions to settle the question, I found a rich and long history, spanning many fields. Indeed, several recently proposed models from the computer science community have antecedents in work from decades ago. Here, I briefly survey some of this history, focusing on underlying generative models that lead to these distributions. One finding is that lognormal and power law distributions connect quite naturally, and hence it is not surprising that lognormal distributions have arisen as a possible alternative to power law distributions across many fields.

1 Introduction

Power law distributions (also often referred to as heavy-tail distributions, Pareto distributions, Zipfian distributions, etc.) are now pervasive in computer science; see, for example, [1, 8, 7, 9, 16, 19, 21, 22, 24, 25, 27, 28, 33, 34, 40, 41, 43, 45, 47, 50, 61, 69].[1]

This paper was specifically motivated by a recent paper by Downey [25] challenging the now conventional wisdom that file sizes are governed by a power law distribution. The argument was substantiated both by collected data and by the development of an underlying generative model which suggested that file sizes were better modeled by a lognormal distribution.[2] In my attempts to learn more about this question, I was drawn to the history the lognormal and power law distributions. As part of this process, I delved into past and present literature, and came across some interesting facts that appear not to be well known in the computer science community. This paper represents an attempt

*Supported in part by an Alfred P. Sloan Research Fellowship and NSF grant CCR-9983832. Harvard University, Division of Engineering and Applied Sciences, 33 Oxford St., Cambridge, MA 02138. michaelm@eecs.harvard.edu. A preliminary version of this work appeared as [64].

[1] I apologize for leaving out countless further examples.

[2] I elaborate on this specific model in another paper [63].

to disseminate what I have found, focusing specifically on the models of processes that generate these distributions.

Perhaps the most interesting discovery is that much of what we in the computer science community have begun to understand and utilize about power law and lognormal distributions has long been known in other fields, such as economics and biology. For example, models of a dynamically growing Web graph that result in a power law distribution for in- and out-degrees have become the focus of a great deal of recent study. In fact, as I describe below, extremely similar models date back to at least the 1950's, and arguably back to the 1920's.

A second discovery is the argument over whether a lognormal or power law distribution is a better fit for some empirically observed distribution has been repeated across many fields over many years. For example, the question of whether income distribution follows a lognormal or power law distribution also dates back to at least the 1950's. The issue arises for other financial models, as detailed in [59]. Similar issues continue to arise in biology [37], chemistry [67], ecology [4, 80], astronomy [82], and information theory [48, 70]. These cases serve as a reminder that the problems we face as computer scientists are not necessarily new, and we should look to other sciences both for tools and understanding.

A third discovery from examining previous work is that power law and lognormal distributions are intrinsically connected. Very similar basic generative models can lead to either power law or lognormal distributions, depending on seemingly trivial variations. There is therefore a reason why this argument as to whether power law or lognormal distributions are more accurate has arisen and repeated itself across a variety of fields.

The purpose of this paper is therefore to explain some of the basic generative models that lead to power law and lognormal distributions, and specifically to cover how small variations in the underlying model can change the result from one to the other. A second purpose is to provide along the way (incomplete) pointers to some of the recent and historically relevant scientific literature.

This survey is intended to be accessible to a general audience. That is, it is intended for computer science theorists, computer scientists who are not theorists, and hopefully also people outside of computer science. Therefore, while mathematical arguments and some probability will be used, the aim is for the mathematics to be intuitive, clean, and comprehensible rather than rigorous and technical. In some cases details may be suppressed for readability; interested readers are referred to the original papers.

2 The Distributions: Basic Definitions and Properties

We begin by reviewing basic facts about power law and lognormal distributions.

For our purposes, a non-negative random variable X is said to have a *power law* distribution if

$$\Pr[X \geq x] \sim cx^{-\alpha}$$

for constants $c > 0$ and $\alpha > 0$. Here $f(x) \sim g(x)$ represents that the limit of the ratios goes to 1 as x grows large. Roughly speaking, in a power law distribution asymptotically the tails fall according to the power α. Such a distribution leads to much heavier tails than other common models, such as exponential distributions. One specific commonly used power law distribution is the *Pareto distribution*, which satisfies

$$\Pr[X \geq x] = \left(\frac{x}{k}\right)^{-\alpha}$$

for some $\alpha > 0$ and $k > 0$. The Pareto distribution requires $X \geq k$. The density function for the Pareto distribution is $f(x) = \alpha k^{\alpha} x^{-\alpha-1}$. For a power law distribution, usually α falls in the range $0 < \alpha \leq 2$, in which case X has infinite variance. If $\alpha \leq 1$, then X also has infinite mean.

If X has a power law distribution, then in a log-log plot of $\Pr[X \geq x]$, also known as the *complementary cumulative distribution function*, asymptotically the behavior will be a straight line. This provides a simple empirical test for whether a random variable has a power law given an appropriate sample. For the specific case of a Pareto distribution, the behavior is exactly linear, as

$$\ln(\Pr[X \geq x]) = -\alpha(\ln x - \ln k).$$

Similarly, on a log-log plot the density function for the Pareto distribution is also a straight line:

$$\ln f(x) = (-\alpha - 1)\ln x + \alpha \ln k + \ln \alpha.$$

A random variable X has a *lognormal distribution* if the random variable $Y = \ln X$ has a normal (i.e., Gaussian) distribution. Recall that the normal distribution Y is given by the density function

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(y-\mu)^2/2\sigma^2}$$

where μ is the mean, σ is the standard deviation (σ^2 is the variance), and the range is $-\infty < y < \infty$. The density function for a lognormal distribution therefore satisfies

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \mathrm{e}^{-(\ln x-\mu)^2/2\sigma^2}.$$

Note that the change of variables introduces an additional $1/x$ term outside of the exponential term. The corresponding complementary cumulative distribution function for a lognormal distribution is given by

$$\Pr[X \geq x] = \int_{z=x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma z} \mathrm{e}^{-(\ln z-\mu)^2/2\sigma^2} dz.$$

We will say that X has parameters μ and σ^2 when the associated normal distribution Y has mean μ and variance σ^2, where the meaning is clear. The lognormal distribution is skewed, with mean $e^{\mu+\frac{1}{2}\sigma^2}$, median e^{μ}, and mode $e^{\mu-\sigma^2}$. A lognormal distribution has finite mean and variance, in contrast to the power law distribution under natural parameters.

Despite its finite moments, the lognormal distribution is extremely similar in shape to power law distributions, in the following sense: if X has a lognormal distribution, then in a log-log plot of the complementary cumulative distribution function or the density function, the behavior will be a straight line except for a large portion of the body of the distribution. Intuitively, for example, the complementary cumulative distribution function of a normal distribution appears close to linear. Indeed, if the variance of the corresponding normal distribution is large, the distribution may appear linear on a log-log plot for several orders of magnitude.

To see this, let us look the logarithm of the density function, which is easier to work with than the complementary cumulative distribution function (although the same idea holds). We have

$$\begin{aligned} \ln f(x) &= -\ln x - \ln \sqrt{2\pi}\sigma - \frac{(\ln x - \mu)^2}{2\sigma^2} && (1) \\ &= -\frac{(\ln x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right) \ln x - \ln \sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2}. && (2) \end{aligned}$$

If σ is sufficiently large, then the quadratic term of equation (2) will be small for a large range of x values, and hence the logarithm of the density function will appear almost linear for a large range of values.

Finally, recall that normal distributions have the property that the sum of two normal random variables Y_1 and Y_2 with μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively is a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. It follows that the product of lognormal distributions is again lognormal.

3 Power Laws via Preferential Attachment

We now move from mathematical definitions and properties to generative models. For the power law distribution, we begin by considering the World Wide Web. The World Wide Web can naturally be thought of as a graph, with pages corresponding to vertices and hyperlinks corresponding to directed edges. Empirical work has shown indegrees and outdegrees of vertices in this graph obey power law distributions. There has subsequently been a great deal of recent theoretical work on designing random graph models that yield Web-like graphs [7, 16, 19, 24, 40, 41, 43, 45]. An important criterion for an appropriate random graph model is that it yields power law distributions for the indegrees and outdegrees.

Most models are variations of the following theme. Let us start with a single page, with a link to itself. At each time step, a new page appears, with outdegree 1. With probability $\alpha < 1$, the link for the new page points to a page chosen uniformly at random. With probability $1-\alpha$, the new page points to page chosen proportionally to the indegree of the page. This model exemplifies what is often called *preferential attachment*; new objects tend to attach to popular objects. In the case of the Web graph, new links tend to go to pages that already have links.

A simple if slightly non-rigorous argument for the above model goes as follows [7, 24, 41, 45]. Let $X_j(t)$ (or just X_j where the meaning is clear) be the number of pages with indegree j when there are t pages in the system. Then for $j \geq 1$ the probability that X_j increases is just

$$\alpha X_{j-1}/t + (1-\alpha)(j-1)X_{j-1}/t;$$

the first term is the probability a new link is chosen at random and chooses a page with indegree $j-1$, and the second term is the probability that a new link is chosen proportionally to the indegrees and chooses a page with indegree $j-1$. Similarly, the probability that X_j decreases is

$$\alpha X_j/t + (1-\alpha)jX_j/t.$$

Hence, for $j \geq 1$, the growth of X_j is roughly given by

$$\frac{dX_j}{dt} = \frac{\alpha(X_{j-1} - X_j) + (1-\alpha)((j-1)X_{j-1} - jX_j)}{t}.$$

Some mathematical purists may object to utilizing a continuous differential equation to describe what is clearly a discrete process. This intuitively appealing approach can be justified more formally using martingales [45] and in particular the theoretical frameworks of Kurtz and Wormald [24, 46, 83].

The case of X_0 must be treated specially, since each new page introduces a vertex of indegree 0.

$$\frac{dX_0}{dt} = 1 - \frac{\alpha X_0}{t}.$$

Suppose in the steady state limit that $X_j(t) = c_j \cdot t$; that is, pages of indegree j constitute a fraction c_j of the total pages. Then we can successively solve for the c_j. For example,

$$\frac{dX_0}{dt} = c_0 = 1 - \frac{\alpha X_0}{t} = 1 - \alpha c_0,$$

from which we find $c_0 = \frac{1}{1+\alpha}$. More generally, we find using the equation for dX_j/dt that for $j \geq 1$,

$$c_j(1 + \alpha + j(1-\alpha)) = c_{j-1}(\alpha + (j-1)(1-\alpha)).$$

This recurrence can be used to determine the c_j exactly. Focusing on the asymptotics, we find that for large j

$$\frac{c_j}{c_{j-1}} = 1 - \frac{2-\alpha}{1+\alpha+j(1-\alpha)} \sim 1 - \left(\frac{2-\alpha}{1-\alpha}\right)\left(\frac{1}{j}\right).$$

Asymptotically, for the above to hold we have $c_j \sim cj^{-\frac{2-\alpha}{1-\alpha}}$ for some constant c, giving a power law. To see this, note that $c_j \sim cj^{-\frac{2-\alpha}{1-\alpha}}$ implies

$$\frac{c_j}{c_{j-1}} \sim \left(\frac{j-1}{j}\right)^{\frac{2-\alpha}{1-\alpha}} \sim 1 - \left(\frac{2-\alpha}{1-\alpha}\right)\left(\frac{1}{j}\right).$$

Strictly speaking, to show it is a power law, we should consider $c_k^* = \sum_{j \geq k} c_j$, since we desire the behavior of the tail of the distribution. However, we have

$$c_k^* \sim \sum_{j \geq k} cj^{-\frac{2-\alpha}{1-\alpha}} \sim \int_{j=k}^{\infty} cj^{-\frac{2-\alpha}{1-\alpha}} dj \sim c' k^{-\frac{1}{1-\alpha}}$$

for some constant c'. More generally, if the fraction of items with weight j falls roughly proportionally to $j^{-\alpha}$, the fraction of items with weight greater than or equal to j falls roughly proportionally $j^{1-\alpha}$, a fact we make use of throughout.

Although the above argument was described in terms of degree on the Web graph, this type of argument is clearly very general and applies to any sort of preferential attachment. In fact the first similar argument dates back to at least 1925. It was introduced by Yule [84] to explain the distribution of species among genera of plants, which had been shown empirically by Willis to satisfy a power law distribution. While the mathematical treatment from 1925 is different than modern versions, the outline of the general argument is remarkably similar. Mutations cause new species to develop within genera, and more rarely mutations lead to entirely new genera. Mutations within a genus are more likely to occur in a genus with more species, leading to the preferential attachment.

A clearer and more general development of how preferential attachment leads to a power law was given by Simon [75] in 1955. Again, although Simon was not interested in developing a model for the Web, he lists five applications of this type of model in his introduction: distributions of word frequencies in documents, distributions of numbers of papers published by scientists, distribution of cities by population, distribution of incomes, and distribution of species among genera. Simon was aware of Yule's previous work, and suggests his work is a generalization. Simon's argument, except for notation and the scaling of variables, is painfully similar to the outline above.

As one might expect from Simon's list of applications, power laws had been observed in a variety of fields for some time; Simon was attempting to give a mathematical argument explaining these observations. The earliest apparent reference is to the work by Pareto

[68] in 1897, who introduced the Pareto distribution to describe income distribution. The first known attribution of the power law distribution of word frequencies appears to be due to Estoup in 1916 [26], although generally the idea (and its elucidation) are attributed to Zipf [86, 87, 88]. Similarly, Zipf is often credited with noting that city sizes appear to match a power law, although this idea can be traced back further to 1913 and Auerbach [6]. Lotka (circa 1926) found in examining the number of articles produced by chemists that the distribution followed a power law [52]; indeed, power laws of various forms appear in many places in informetrics [15].

Although we now associate the argument above with the Web graph, even before the Web graph became popular, more formal developments of the argument above had been developed as part of the study of random trees. Specifically, consider the following recursive tree structure. Begin with a root node. At each step, a new node is added; its parent is chosen from the current vertices with probability proportional to one plus the number of children of the node. This is just another example of preferential attachment; indeed, it is essentially equivalent to the simple Web graph model described above with the probability α of choosing a random node equal to $1/2$. That the degree distribution of such graphs obey a power law (in expectation) was proven in 1993 in works by Mahmoud, Smythe, and Szymański [54]. See also the related [53, 81, 71, 79].

Of course, in recognizing the relationship between the recent work on Web graph models and this previous work, it would be remiss to not point out that modern developments have led to many new insights. Perhaps most important is the development of a connection between Simon's model, which appears amenable only to limiting analysis based on differential equations, and purely combinatorial models based on random graphs [14, 54, 79]. Such a connection is important for further rigorous analysis of these structures. Also, current versions of Simon's arguments based on martingales provide a much more rigorous foundation [14, 19, 45, 53]. More recent work has focused on greater understanding of the structure of graphs that arise from these kinds of preferential attachment model. It has been shown that in the Web graph model described above where new pages copy existing links, the graphs have community substructures [45], a property not found in random graphs but amply found in the actual Web [32, 44]. The diameter of these random Web graphs have also been the subject of recent study [5, 13]. Still, it is important to note how much was already known about the power law phenomenon in various fields well before the modern effort to understand power laws on the Web, and how much computer scientists had to reinvent.

4 Power Laws via Optimization

Mandelbrot had developed other arguments for deriving power law distributions based on information theoretic considerations somewhat earlier than Simon [55]. His argument is very similar in spirit to other recent optimization based arguments for heavy tailed distributions [17, 27, 85].

We sketch Mandelbrot's framework, which demonstrates a power law in the rank-frequency distribution of words. That is, the frequency p_j of the jth most-used word, given as a fraction of the time that word appears, follows a power law in j, so $p_j \sim cj^{-\alpha}$. This is a slightly different flavor than the type power law than we considered previously; Simon's model considers the fraction of words that appear j times. But of course the two are related. We clarify this by following an argument of Bookstein [15].

Suppose we have a text where the number of words q_k that appear k times is given by $q_k = Qk^{-\alpha}$ for $\alpha > 1$. Further suppose for convenience we have one most frequent word that appears k_m times, so that we may write $q_k = (k/k_m)^{-\alpha}$. The number of words that appear k or more times is then approximately

$$\int_k^{k_m} \left(\frac{x}{k_m}\right)^{-\alpha} dx,$$

and hence the rank j of a word that appears k times is approximately

$$j = \frac{j_m}{\alpha - 1}\left[\left(\frac{j_m}{k}\right)^{\alpha-1} - 1\right].$$

Now solving for k in terms of j, we find that the jth most-used word appears approximately

$$k = j_m\left[\frac{(\alpha-1)j}{j_m} + 1\right]^{-1/(\alpha-1)}$$

times, yielding a power law for the frequency p_j as a function of j.

We now begin Mandelbrot's argument. Consider some language consisting of n words. The cost of using the jth word of the language in a transmission is C_j. For example, if we think of English text, the cost of a word might be thought of as the number of letters plus the additional cost of a space. Hence a natural cost function has $C_j \sim \log_d j$ for some alphabet size d. Suppose that we wish to design the language to optimize the average amount of information per unit transmission cost. Here, we take the average amount of information to be the entropy. We think of each word in our transmission as being selected randomly, and the probability that a word in the transmission is the jth word of the language is p_j. Then the average information per word is the entropy $H = -\sum_{j=1}^{n} p_j \log_2 p_j$, and the average cost per word is $C = \sum_{j=1}^{n} p_j C_j$. The question is how would the p_j be chosen to minimize $A = C/H$. Taking derivatives, we find

$$\frac{dA}{dp_j} = \frac{C_j H + C\log_2(\mathrm{e}p_j)}{H^2}.$$

Hence all the derivatives are 0 (and A is in fact minimized) when $p_j = 2^{-HC_j/C}/\mathrm{e}$. Using $C_j \sim \log_d j$, we obtain a power law for the p_j.[3] Mandelbrot argues that a variation of

[3] The eagle-eyed reader might note that technically the result above does not quite match a power law as we have defined it; just because $C_j \sim \log_d j$ does not strictly give us $p_j \sim j^{-\alpha}$. In this case this is a minor point; really C_j is within an additive constant of $\log_d j$, and we therefore find that p_j is within a constant multipilcative factor of a power law. We ignore this distinction henceforth.

this model matches empirical results for English quite well.

Carlson and Doyle suggest a similar framework for analyzing file sizes and forest files [17]. Fabrikant, Koutsoupias, and Papadimitriou introduce combinatorial models for the Internet graph (which should not be confused with the Web graph; the Internet graph consists of the servers and links between them as opposed to Web pages) and file sizes based on local optimization that also yield power laws [27].

As an aside, I found when reviewing the literature that Mandelbrot strongly argued against Simon's alternative assumptions and derivations based on preferential attachment when his article came out. This led to what is in retrospect an amusing but apparently at the time quite heated exchange between Simon and Mandelbrot in the journal *Information and Control* [56, 76, 57, 77, 58, 78].[4]

It is worth noting that economists appear to have given the nod to Simon and the preferential attachment model. Indeed, a recent popular economics text by Krugman [42] offers a derivation of the power law similar to Simon's argument.[5] A more formal treatment is given by Gabaix [29].

5 Multiplicative processes

Lognormal distributions are generated by processes that follow what the economist Gibrat called the law of proportionate effect [30, 31]. We here use the term *multiplicative process* to describe the underlying model. In biology, such processes are used to described the growth of an organism. Suppose we start with an organism of size X_0. At each step j,

[4] At the risk of offending the original authors, a few excerpts from the exchange are worth citing to demonstrate the disagreement. The abstract of Mandelbrot's initial note begins, "This note is a discussion of H. A. Simon's model (1955) concerning the class of frequency distributions generally associated with the name of G. K. Zipf. The main purpose is to show that Simon's model is analytically circular in the case of the linguistic laws of Estouf-Zipf and Willis-Yule." [56] The abstract of Simon's rebuttal begins, "This note takes issue with a recent criticism by Dr. B. Mandelbrot of a certain stochastic model to explain word-frequency data. Dr. Mandelbrot's principal empirical and mathematical objections to the model are shown to be unfounded." [76] Mandelbrot begins his "Final Note" in response to Simon's rebuttal as follows: "In a "Note" published in this Journal in 1959 (Mandelbrot, 1959), we had shown the impossibility of ever explaining the Pareto-Yule-Zipf class of skew distribution functions by using the model due to H. A. Simon (1955). That model was most ingenious and tempting but it turned out to be totally inadequate to derive the desired results." [57] Simon's further rebuttal contains the sentence, "Thus we have come to the end of the list of Dr. Mandelbrot's objections to my approximation without finding a single one that is valid." [77] In the final volley of the series of exchanges (Mandelbrot felt it necesary to add a "Post Scriptum" after his "Final Note") the abstracts are short and to the point. Mandelbrot says, "My criticism has not changed since I first had the privilege of commenting upon a draft of Simon (1955)." [58] Simon's final word is: "Dr. Mandelbrot has proposed a new set of objections to my 1955 models of the Yule distribution. Like his earlier objections, these are invalid." [78]

[5] As an interesting example of the breadth of the scope of power-law behavior, one review of Krugman's book, written by an urban geographer, accuses the author of excessive hubris for not noting the significant contributions made by urban geographers with regard to Simon's model [11].

the organism may grow or shrink, according to a random variable F_j, so that

$$X_j = F_j X_{j-1}.$$

The idea is that the random growth of an organism is expressed as a percentage of its current weight, and is independent of its current actual size. If the $F_k, 1 \leq k \leq j$, are all governed by independent lognormal distributions, then so is each F_j, inductively, since the product of lognormal distributions is again lognormal.

More generally, lognormal distributions may be obtained even if the F_j are not themselves lognormal. Specifically, consider

$$\ln X_j = \ln X_0 + \sum_{k=1}^{j} \ln F_k.$$

Assuming the random variables $\ln F_k$ satisfy appropriate conditions, the Central Limit Theorem says that $\sum_{k=1}^{j} \ln F_k$ converges to a normal distribution, and hence for sufficiently large j, X_j is well approximated by a lognormal distribution. In particular, if the $\ln F_k$ are independent and identically distributed variables with finite mean and variance, then asymptotically X_j will approach a lognormal distribution.

Multiplicative processes are used in biology and ecology to describe the growth of organisms or the population of a species. In economics, perhaps the most well-known use of the lognormal distribution derives from the Black-Scholes option pricing model [12], which is a specific application of Ito's lemma (see, e.g., [35, 36]). In a simplified version of this setting [20, 35], the price of a security moves in discrete time steps, and the price X_j changes according to $X_j = F_j X_{j-1}$, where F_j is lognormally distributed. Using this model, Black and Scholes demonstrate how to use options to guarantee a risk-free return equivalent to the prevailing interest rate in a perfect market. Other applications in for example geology and atmospheric chemistry are given in [23]. More recently, as described below, Adamic and Huberman suggest that multiplicative processes may describe the growth of sites on the Web as well as the growth of user traffic on Web sites [33, 34]. Lognormal distributions have also been suggested for file sizes [8, 9, 25].

The connection between multiplicative processes and the lognormal distribution can be traced back to Gibrat around 1930 [30, 31], although Kapteyn described in other terms an equivalent process in 1903 [38], and McAlister described the lognormal distribution around 1879 [60]. Aitchison and Brown suggest that the lognormal distribution may be a better fit for income distribution than a power law distribution, representing perhaps the first time the question of whether a power law distribution or a lognormal distribution gives a better fit was fully developed [2, 3]. It is interesting that when examining income distribution data, Aitchison and Brown observe that for lower incomes a lognormal distribution appears a better fit, while for higher incomes a power law distribution appears better; this is echoed in later work by Montroll and Schlesinger [65, 66], who offer a possible mathematical justification discussed below. Similar observations have been given for file sizes [8, 9].

5.1 Multiplicative Models and Power Law Distributions

Although the multiplicative model is used to generate lognormal or approximately distributions, only a small change from the lognormal generative process yields a generative process with a power law distribution. To provide a concrete example, we consider the interesting history of work on income distributions.

Recall that Pareto introduced the Pareto distribution in order to explain income distribution at the tail end of the nineteenth century. Champernowne [18], in a work slightly predating Simon (and acknowledged by Simon, who suggested his work generalized and extended Champernowne), offered an explanation for this behavior. Suppose that we break income into discrete ranges in the following manner. We assume there is some minimum income m. For the first range, we take incomes between m and γm, for some $\gamma > 1$; for the second range, we take incomes between γm and $\gamma^2 m$. We therefore say that a person is in class j for $j \geq 1$ if their income is between $m\gamma^{j-1}$ and $m\gamma^j$. Champernowne assumes that over each time step, the probability of an individual moving from class i to class j, which we denote by p_{ij}, depends only on the value of $j - i$. He then considers the equilibrium distribution of people among classes. Under this assumption, Pareto distributions can be obtained.

Let us examine a specific case, where $\gamma = 2$, $p_{ij} = 2/3$ if $j = i - 1$, and $p_{ij} = 1/3$ if $j = i + 1$. Of course the case $i = 1$ is a special case; in this case $p_{11} = 2/3$. In this example, outside of class 1, the expected change in income over any step is 0. It is also easy to check that in this case the equilibrium probability of being in class k is just $1/2^k$, and hence the probability of being in class greater than or equal to k is $1/2^{k-1}$. Hence the probability that a person's income X is larger than $2^{k-1}m$ in equilibrium is given by

$$\mathrm{P}r[X \geq 2^{k-1}m] = 1/2^{k-1},$$

or

$$\mathrm{P}r[X \geq x] = m/x$$

for $x = 2^{k-1}m$. This is a power law distribution.

Note, however, the specific model above looks remarkably like a multiplicative model. Moving from one class to another can be thought of as either doubling or halving your income over one time step. That is, if X_t is your income after t time steps, then

$$X_t = F_t X_{t-1},$$

where F_t is 1/2 with probability 2/3 and 2 with probability 1/3. Again, $E[X_t] = E[X_{t-1}]$. Our previous discussion therefore suggests that X_t should converge to a lognormal distribution for large t.

What is the difference between the Champernowne model and the multiplicative model? In the multiplicative model, income can become arbitrarily close to zero through successive decreases; in the Champernowne model, there is a minimum income corresponding to the lowest class below which one cannot fall. This small change allows one model to produce a power law distribution while the other produces a lognormal. As long as there is a bounded minimum that acts as a lower reflective barrier to the multiplicative model, it will yield a power law instead of a lognormal distribution. The theory of this phenomenon is more fully developed in [29, 39].

6 Monkeys Typing Randomly

We return now to Mandelbrot's optimization argument for the power law behavior of word frequency in written language. A potentially serious objection to Mandelbrot's argument was developed by the psychologist Miller [62], who demonstrated that the power law behavior of word frequency arises even without an underlying optimization problem. This result, explained below, should perhaps serve as warning: just because one finds a compelling mechanism to explain a power law does not mean that there are not other, perhaps simpler explanations.

Miller suggests the following experiment. A monkey types randomly on a keyboard with n characters and a space bar. A space is hit with probability q; all other characters are hit with equal probability $(1-q)/n$. A space is used to separate words. We consider the frequency distribution of words.

It is clear that as the monkey types each word with c (non-space) characters occurs with probability

$$q_c = \left(\frac{1-q}{n}\right)^c q,$$

and there are n^c words of length c. (We allow the empty word of length 0 for convenience.) The words of longer length are less likely and hence occur lower in the rank order of word frequency. In particular, the word with frequency ranks $1 + (n^j - 1)/(n-1)$ to $(n^{j+1} - 1)/(n-1)$ have j letters. Hence, the word with frequency rank $r_j = n^j$ occurs with probability

$$q_j = q\left(\frac{1-q}{n}\right)^{\log_n r_j} = q\,(r_j)^{\log_n(1-q)-1},$$

and the power law behavior is apparent. Hence the power law associated with word frequency requires neither preferential attachment nor optimization; monkeys typing randomly would produce it.

Bell, Cleary, and Witten observe empirically that when the probabilities of each letter are not equal, a smoother match to the power law develops [10]. I am currently unaware of a proof similar to the one above demonstrating that power law behavior occurs when the probabilities for each of the letters are arbitrary. Indeed, to confuse the issue, one paper on the subject claims that if the letter frequencies are not equal, a lognormal distribution occurs [70] (see also [51], where this claim is repeated). It is worth examining this argument more carefully, since it demonstrates the confusion that can arise in trying to distinguish models that generate power law and lognormal distributions.

Perline notes that in the experiment with monkeys typing randomly, if we consider words only of some fixed length m, for m sufficiently large their frequency-rank distribution will approximate a lognormal distribution, following the paradigm of multiplicative processes. To see this, let the probabilities for our n characters be $p_1, p_2, \ldots, p_n$. Consider the generation a random m-letter word. Let X_i take on the value p_j if the ith letter is j. Then $Y_m = X_1 X_2 \ldots X_m$ is a random variable whose value corresponds to the probability that a word chosen uniformly at random from all m-letter words appears as the monkeys type. We have that $\ln Y_m = \sum_{k=1}^{m} \ln X_i$; since the X_i are independent and identically distributed, $\log Y_m$ converges to a normal distribution by the Central Limit Theorem, and hence Y_m converges to a lognormal distribution. Notice that this holds true even if all letter frequencies are equal, although in this case the resulting distribution is trivial.

Perline then argues that if we consider all words of length up to m, we still obtain asymptotic convergence to a lognormal distribution. This follows from a generalization of the Central Limit Theorem due to Anscombe. Intuitively, this is because most words have length close to m, so the words with small length are just noise in the distribution. This result does require that the probability some two letters have different probabilities of being hit.

From this, it might be tempting to conclude that the distribution if the word length is unrestricted is also lognormal when letters do not all have the same probabilities. However, this does not follow. The problem is that for each value of m we obtain a slightly different lognormal distribution. Hence it is not necessarily true that in the limit as m increases we are getting closer and closer to some final lognormal distributions. Rather, we have a sequence of lognormal distributions that is converging to some distribution. To justify that the result need not be lognormal, I present an amusing example of my own devising.

Consider an alphabet with two letters: “a” occurs with probability q, “b” occurs with probability q^2, and a space occurs with probability $1-q-q^2$. The value q must be chosen so that $1 - q - q^2 > 0$. In this case, every valid word the monkey can type occurs with probability $q^j(1 - q - q^2)$ for some integer j. Let us say a word has pseudo-rank j if it occurs with probability $q^j(1 - q - q^2)$. There is 1 word with pseudo-rank 0 (the empty word), 1 with pseudo-rank 1 (“a”), 2 with pseudo-rank 2 (“aa” and “b”), and so on. A simple induction yields that the number of words with pseudo-rank k is in fact the $(k + 1)$st Fibonacci number F_{k+1} (where here we start with $F_0 = 0$ and $F_1 = 1$). This follows obviously from the fact that to obtain the words with pseudo-rank k we append an “a” to a word with pseudo-rank $k - 1$, or a “b” to a word with pseudo-rank $k - 2$.

Recall that $F_k \approx \phi^k/\sqrt{5}$ for large k, where $\phi = (1 + \sqrt{5})/2$. Also $\sum_{i=1}^{k} F_k = F_{k+2} - 1$. Now the argument is entirely similar to the case where all items have the same probability. If we ask for the frequency of the $r_j = F_j$th most frequent item, it has pseudo-rank $j - 2$, and hence its frequency is

$$q^{j-2}(1 - q - q^2) \approx q^{\log_\phi \sqrt{5} r_j - 2}(1 - q - q^2) = r_j^{\log_\phi q} q^{\log_\phi \sqrt{5} - 2}(1 - q - q^2),$$

and again we have power law behavior.

There is nothing special about having two characters for this example; one could easily expand it to include more complex generalized Fibonacci sequences. A suitable generalization is in fact appears feasible for any probabilities $p_1, p_2, \ldots, p_n$ associated with the n characters, although a formal proof is beyond the scope of this survey.[6] Roughly, let p_1 be the largest of the p_i, and let $p_j = p_1^{\gamma_j}$ for $j \geq 1$. Then the number of words with frequency greater than or equal to p_1^k grows approximately proportionally to $(1/c)^k$, where c is the unique real root between 0 and 1 of $\sum_{j=1}^{n} x^{\gamma_j} = 1$. This is all we need for the monkeys to produce a power law distribution, following the arguments above.

7 Double Pareto Distributions

Interestingly, there is another variation on the multiplicative generative model also yields power law behavior. Recall that in the multiplicative model, if we begin with value X_0 and every step yields an independent and identically distributed multiplier from a lognormal distribution F, then any resulting distribution X_t after t steps is lognormal. Suppose, however, that instead of examining X_t for a specific value of t, we examine the random variable X_T where T itself is a random variable. As an example, when considering income distribution, in seeing the data we may not know how long each person has lived. If different age groups are intermixed, the number of multiplicative steps each person may be thought to have undergone may be thought of as a random variable.

This effect was noticed as early as 1982 by Montroll and Schlesinger [65, 66]. They show that a mixture of lognormal distributions based on a geometric distribution would have essentially a lognormal body but a power law distribution in the tail. Huberman and Adamic suggest a pleasantly simple variation of the above result; in the case where the time T is an exponential random variable, and we may think of the number of multiplicative steps as being continuous, the resulting distribution of X_T has a power law distribution [33, 34]. Huberman and Adamic go on to suggest that this result can explain the power law distribution observed for the number of pages per site. As the Web is growing exponentially, the age of a site can roughly be thought of as distributed like an exponential random variable. If the growth of the number of pages on a Web site follows a multiplicative process, the above result suggests a power law distribution.

In more recent independent work, Reed provides the correct full distribution for the above model, which yields what he calls a double Pareto distribution [72]. Specifically, the resulting distribution has one Pareto tail distribution for small values (below some point) and another Pareto tail distribution for large values (above the same point).[7]

[6] I am currently constructing a formal treatment of this argument, which appears to require some non-trivial analytic number theory. This work will hopefully appear in the near future.

[7] For completeness we note that Huberman and Adamic concentrate only on the tail of the density function, and correctly determine the power law behavior. However, they miss the two-sided nature of the distribution. Reed gives the complete correct form, as we do below.

For example, consider for simplicity the case where if we stop a process at time t the result is a lognormal random variable with mean 0 and variance t. Then if we stop the process at an exponentially distributed time with mean $1/\lambda$, the density function of the result is

$$f(x) = \int_{t=0}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{1}{\sqrt{2\pi t} x} \mathrm{e}^{-(\ln x)^2/2t} dt.$$

Using the substitution $t = u^2$ gives

$$f(x) = \frac{2\lambda}{\sqrt{2\pi} x} \int_{u=0}^{\infty} \mathrm{e}^{-\lambda u^2 - (\ln x)^2/2u^2} du.$$

An integral table gives us the identity

$$\int_{z=0}^{\infty} \mathrm{e}^{-az^2 - b/z^2} = \frac{1}{2}\sqrt{\frac{\pi}{a}} \mathrm{e}^{-2\sqrt{ab}},$$

which allows us to solve for the resulting form. Note that in the exponent $\sqrt{2ab}$ of the identity we have $b = (\ln x)^2/2$. Because of this, there are two different behaviors, depending on whether $x \geq 1$ or $x \leq 1$. For $x \geq 1$, $f(x) = \left(\sqrt{\lambda/2}\right) x^{-1-\sqrt{2\lambda}}$, so the result is a power law distribution. For $x \leq 1$, $f(x) = \left(\sqrt{\lambda/2}\right) x^{-1+\sqrt{2\lambda}}$.

The double Pareto distribution falls nicely between the lognormal distribution and the Pareto distribution. Like the Pareto distribution, it is a power law distribution. But while the log-log plot of the density of the Pareto distribution is a single straight line, for the double Pareto distribution the log-log plot of the density consists of two straight line segments that meet at a transition point. This is similar to the lognormal distribution, which has a transition point around its median e^{μ} due to the quadratic term, as shown in equation (1). Hence an appropriate double Pareto distribution can closely match the body of a lognormal distribution and the tail of a Pareto distribution. For example, Figure 1 shows the complementary cumulative distribution function for a lognormal and a double Pareto distribution. (These graphs have only been minimally tuned to give a reasonable match.) The plots match quite well with a standard scale for probabilities, as shown on the left. On the log-log scale, however, one can see the difference in the tail behavior. The double Pareto distribution follows a power law; the lognormal distribution has a clear curvature.

Reed also suggests a generalization of the above called a double Pareto-lognormal distribution with similar properties [73]. The double Pareto-lognormal distribution has more parameters, but might allow closer matches with empirical distributions.

It seems reasonable that in many processes the time an object has lived should be considered a random variable as well, and hence this model may prove more accurate for many situations. For example, that the double Pareto tail phenomenon could explain why income distributions and file size distributions appear better modeled by a distribution with a lognormal body and a Pareto tail [2, 8, 9, 65, 66]. Reed presents empirical evidence

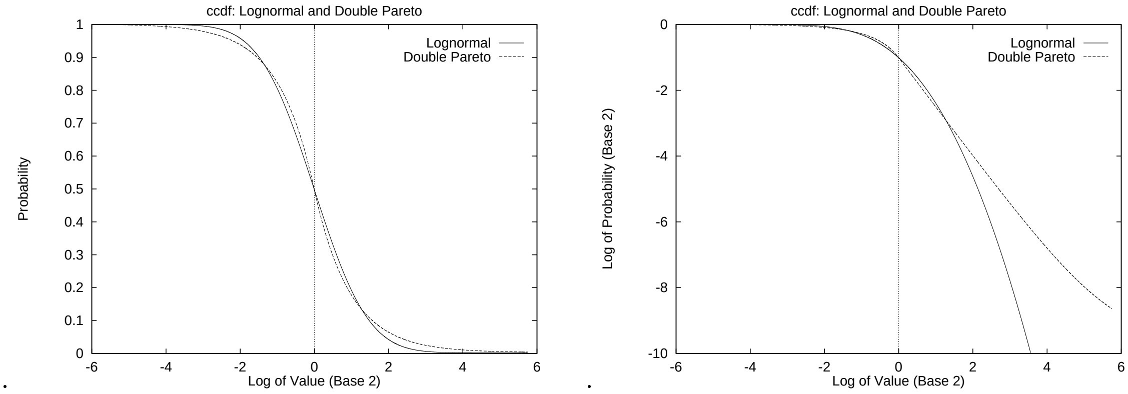

Figure 1: Shapes of lognormal and double Pareto distributions.

in favor of using the double Pareto and double-Pareto lognormal distributions for incomes and other applications [72, 73].

To give an idea of why it might be natural for the time parameter to be (roughly) exponentially distributed, I briefly describe a model that I introduced in [63]. This model combines ideas from the theory of recursive trees, preferential attachment, and the double Pareto framework. Consider a graph process that works as follows: at each step, with probability γ, a new node is introduced that becomes the root of a new tree. Each new node has an associated size chosen independently and uniformly at random from a distribution $\mathcal{D}_1$. With probability $1 - \gamma$, an existing node is chosen uniformly at random, and it generates a child. The size of a child is equal to the size of its parent, multiplied by some multiplicative factor chosen by a distribution $\mathcal{D}_2$. It is easy to show that the distribution of the depths of the nodes generated in this manner converges to a geometric distribution. Along each branch of the tree, the size of the nodes follows a multiplicative process. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are lognormal distributions, then the size of a randomly chosen node is a geometric mixture of lognormally distributed random variables, which closely matches the exponential mixture required for a double Pareto distribution. In fact, the tail behaviors are the same. I use this model to explain file size distributions in [63]; [74] analyzes other similar models.

This line of thought also ties back into the discussion of monkeys typing randomly. In the case of unrestricted word lengths and unequal letter probabilities, the word length is geometrically distributed, and the probability of a word of any (large) fixed length is approximately lognormal, with the appropriate mean and variance being proportional to the length of the word. Hence the underlying distribution of word lengths is a geometric mixture of approximately lognormal random variables as in the framework above, and hence the resulting power law is unsurprising.

8 Conclusions

Power law distributions and lognormal distributions are quite natural models and can be generated from simple and intuitive generative processes. Because of this, they have appeared in many areas of science. This example should remind us of the importance of seeking out and recognizing work in other disciplines, even if it lies outside our normal purview. Since computer scientists invented search engines, we really have little excuse. On a personal note, I was astounded at how the Web and search engines have transformed the possibilities for mining previous research; many of the decades-old articles (including the 1925 article by Yule!) cited here are in fact available on the Web.

It is not clear that the above discussion settles one way or another whether lognormal or power law distributions are better models for things like file size distributions. Given the close relationship between the two models, it is not clear that a definitive answer is possible; it may be that in seemingly similar situations slightly different assumptions prevail. The fact that power law distributions arise for multiplicative models once the observation time is random or a lower boundary is put into effect, however, may suggest that power laws are more robust models. Indeed, following the work of Reed [72, 73], we recommend the double Pareto distribution and its variants as worthy of further consideration in the future.

From a more pragmatic point of view, it might be reasonable to use whichever distribution makes it easier to obtain results. This runs the risk of being inaccurate; perhaps in some cases the fact that power law distributions can have infinite mean and variance are salient features, and therefore substituting a lognormal distribution loses this important characteristic. Also, if one is attempting to predict future behavior based on current data, misrepresenting the tail of the distribution could have severe consequences. For example, large files above a certain size might be rare currently, and hence both lognormal and power law distibutions based on current data might capture these rare events adequately. As computer systems with more memory proliferate, and even larger files become more frequent, the prediction from two models may vary more substantially. The recent work [51] argues that for at least some network applications the difference in tails is not important. We believe that formalizing this idea is an important open question. Specifically, it would be useful to know in a more formal sense in what situations the small differences between power laws and lognormal distributions manifest themselves in vastly different qualitative behavior, and in what cases a power law distribution can be suitably approximated by a lognormal distributions.

9 Acknowledgments

The author would like to thank John Byers, Mark Crovella, Allen Downey, Alan Frieze, Jon Kleinberg, and Eli Upfal for suggestions on how to improve this work. For further reading, the author strongly recommends the article by Xavier Gabaix [29], which pro-

vides both underlying mathematics and an economic perspective and history. Similarly, Mandelbrot provides both history about and his own perspective on lognormal and power law distributions in a recent book [59]. Wentian Li has a Web page devoted to Zipf's law which is an excellent reference [49]. For lognormal distributions, useful sources include the text by Aitchison and Brown [3] or the modern compendium edited by Crow and Shimizu [23].

References

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pages 171-180, 2000.

[2] J. Aitchison and J. A. C. Brown. On criteria for descriptions of income distribution. *Metroeconomica*, 6:88-98. 1954.

[3] J. Aitchison and J. A. C. Brown. **The Lognormal Distribution**. Cambridge University Press, 1957.

[4] A. P. Allen, B. Li, and E. L. Charnov. Population fluctuations, power laws and mixtures of lognormal distributions. *Ecology Letters*, 4:1-3, 2001.

[5] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World Wide Web. *Nature*, 401:130-131, 1999.

[6] F. Auerbach. Das Gesetz der Bevolkerungskonzentration. *Petermanns Geographische Mitteilungen*, LIX (1913), 73-76.

[7] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, vol. 272, pages 173-189, 1999.

[8] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: characteristics and caching implications. *World Wide Web*, 2:15-28, 1999.

[9] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *Proceedings of ACM SIGMETRICS*, pages 151-160, 1998.

[10] T. C. Bell, J. G. Cleary, and I. H. Witten. **Text Compression**. Prentice-Hall, Englewood Cliffs, New Jersey, 1990.

[11] B. Berry. Déjà vu, Mr. Krugman. *Urban Geography*, vol 20, 1, pages 1-2, 1999.

[12] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economics*, 81:637-654, 1973.

[13] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. To appear.

[14] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random process. *Random Structures and Algorithms*, vol 18(3): 279-290, 2001.

[15] A. Bookstein. Informetric Distributions, Part I: Unified Overview. *Journal of the American Society for Information Science*, 41(5):368-375, 1990.

[16] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web: experiments and models. In *Proc. of the 9th World Wide Web Conference*, 2000.

[17] J. M. Carlson and J. Doyle. Highly optimized tolerance: a mechanism for power laws in designed systems. *Physics Review E*, 60(2):1412-1427, 1999.

[18] D. Champernowne. A model of income distribution. *Economic Journal*, 63:318-351, 1953.

[19] C. Cooper and A. Frieze. On a general model of undirected Web graphs. In *Proceedings of the 9th Annual European Symposium on Algorithms*, pages 500-511, 2001.

[20] J. Cox, S. Ross, and M. Rubinstien. Option pricing: a simplified approach. *Journal of Financial Economics*, 7:229-265, 1979.

[21] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835-846, 1997.

[22] M. Crovella, M. S. Taqqu, and A. Bestavros. Heavy-tailed probability distributions in the world wide web. In *A Practical Guide to Heavy Tails*, editors R. J. Adler, R. E. Feldman, M. S. Taqqu. Chapter 1, pages 3-26, Chapman and Hall, 1998.

[23] E. L. Crow and K. Shimizu (editors). **Lognormal Distributions: Theory and Applications**. Markel Dekker, Inc., New York, 1988.

[24] E. Drinea, M. Enachescu, and M. Mitzenmacher. Variations on random graph models of the Web. Harvard Computer Science Technical Report TR-06-01.

[25] A. B. Downey. The structural causes of file size distributions. To appear in *MASCOTS 2001.* Available at http://rocky.wellesley.edu/downey/filesize/

[26] J. B. Estoup. **Gammes Stenographiques**. Institut Stenographique de France, Paris, 1916.

[27] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically Optimized Tradeoffs: A new paradigm for power laws in the Internet. In *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, 2002.

[28] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM 1999 Conference*, pages 251-261, 1999.

[29] X. Gabaix. Zipf's law for cities: an explanation. *Quarterly Journal of Economics*, 114:739-767. 1999.

[30] R. Gibrat. Une loi des réparations économiques: l'effet proportionnel. *Bull. Statist. gén Fr.*. 19:469, 1930.

[31] R. Gibrat. **Les inegalites economiques**. Libraire du Recueil Sirey, Paris France, 1931.

[32] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pp. 225-234, 1998.

[33] B. A. Huberman and L. A. Adamic. Evolutionary dynamics of the World Wide Web. Technical Report, Xerox Palo Alto Research Center, 1999. Appears as a brief communication in *Nature*, 401, p. 131, 1999.

[34] B. A. Huberman and L. A. Adamic. The nature of markets in the World Wide Web. *Quarterly Journal of Economic Commerce*, vol 1., pages 5-12, 2000.

[35] J. C. Hull. **Introduction to futures and options markets (third edition)**. Prentice-Hall, Inc., New Jersey, 1997.

[36] K. Itô. **Stochastic Differential Equations**. Memoirs of the American Mathematical Society, 4, 1951.

[37] R. Jain and S. Ramakumar. Stochastic dynamics modeling of the protein sequence length distribution in genomes: implications for microbial evolution. *Physica A*, 273:476-485, 1999.

[38] J. C. Kapteyn. **Skew Frequency Curves in Biology and Statistics**. Astronomical Laboratory, Noordhoff, Groningen, 1903.

[39] H. Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, CXXXI:207-248, 1973.

[40] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: measurements, models, and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, 1999.

[41] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63, 066123-1 – 066123014, 2001.

[42] P. Krugman. **The Self-Organizing Economy**. Blackwell, Cambridge MA, 1996.

[43] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large scale knowledge bases from the Web. In *Proceedings of the 25th VLDB Conference*, 1999.

[44] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th International World Wide Web Conference*, pp. 403-415, 1999.

[45] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57-65, 2000.

[46] T. G. Kurtz, **Approximation of Population Processes**, SIAM, 1981.

[47] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking* pages 1-15, 1994.

[48] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842-1845, 1992.

[49] W. Li. References on Zipf's Law. At http://linkage.rockerfeller.edu/wli/zipf/

[50] M. Mihail and C. H. Papadimitriou. On the eigenvalue power law. In *Proceedings of RANDOM 2002*, pages 254-262, 2002.

[51] W. Gong, Y.Liu, V. Misra, and D. Towsley. On the tails of Web filesize distributions. In *Proceedings of the Thirty-Ninth Annual Allerton Conference on Communication, Control, and Computing*, pages 192-201, 2001.

[52] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16:317-323, 1926.

[53] J. Lu and Q. Feng. Strong consistency of the number of the vertices of given degrees in nonuniform random recursive trees. *Yokohama Math. J.*, 45:61-69, 1998.

[54] H. Mahmound, R. Smythe, and J. Szymański. On the structure of plane-oriented recursive trees and their branches. *Random Structures and Algorithms*, 3:255-266, 1993.

[55] B. Mandelbrot. An informational theory of the statistical structure of languages. In **Communication Theory**, edited by W. Jackson, Betterworth, pages 486-502, 1953.

[56] B. Mandelbrot. A note on a class of skew distribution function: analysis and critique of a paper by H.A. Simon. *Information and Control*, 2:90-99, 1959.

[57] B. Mandelbrot. Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon. *Information and Control*, 4:198-216, 1961.

[58] B. Mandelbrot. Post scriptum to "final note". *Information and Control*, 4:300-304, 1961.

[59] B. Mandelbrot. **Fractals and Scaling in Finance**. Springer-Verlag, New York, 1997.

[60] D. McAlister. The law of the geometric mean. *Proceedings of the Royal Society*, 29:367, 1879.

[61] A. Medina, I. Matta, and J. Byers. On the origin of power laws in Internet topologies. *Computer Communication Review*, 30(2), pages 18–28, 2000.

[62] G. A Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70:311-314, 1957.

[63] M. Mitzenmacher. Dynamic models for file size distributions and double Pareto distributions. Preprint, available at http://www.eecs.harvard.edu/∼michaelm/NEWWORK/papers/.

[64] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. In *Proceedings of the Thirty-Ninth Annual Allerton Conference on Communication, Control, and Computing*, pages 182-191, 2001.

[65] E. W. Montroll and M. F. Shlesinger. On $1/f$ noise and other distributions with long tails. *Proceedings of the National Academy of Sciences, USA*, 79:3380-3383, 1982.

[66] E. W. Montroll and M. F. Shlesinger. Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails. *Journal of Statistical Physics*, 32:209-230, 1983.

[67] T. Nakajima and A. Higurashi. A use of two-channel radiances for an aeresol characterization from space. *Geophysical Research Letters*, 25:3815-3818, 1998.

[68] V. Pareto. **Cours d'Economie Politique**. Droz, Geneva Switzerland, 1896.

[69] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, vol. 3, pages 226-244, 1995.

[70] R. Perline. Zipf's law, the central limit theorem, and the random division of the unit interval. *Physical Review E*, 54(1):220-223, 1996.

[71] B. Pittel. Note on the height of recursive trees and m-ary search trees. *Random Structures and Algorithms*, 5:337-347, 1994.

[72] W. J. Reed. The Pareto law of incomes - an explanation and an extension. Submitted. Available at http://www.math.uvic.ca/faculty/reed/index.html.

[73] W. J. Reed. The double Pareto-lognormal distribution - A new parametric model for size distribution. 2001. Available at http://www.math.uvic.ca/faculty/reed/index.html.

[74] W. J. Reed and B. D. Hughes. From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature. 2002. Available at http://www.math.uvic.ca/faculty/reed/index.html.

[75] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425-440, 1955.

[76] H. A. Simon. Some further notes on a class of skew distribution functions. *Information and Control*, 3:80-88, 1960.

[77] H. A. Simon. Reply to “final note”. *Information and Control*, 4:217-223, 1961.

[78] H. A. Simon. Reply to Dr. Mandelbrot’s post scriptum, *Information and Control*, 4:305-308, 1961.

[79] R. Smythe and H. Mahmound. A survey of recursive trees. *Theoretical Probability and Mathematical Statistics*, 51:1-27, 1995.

[80] R. V. Sole, D. Alonso, and A. McKane. Connectivity and Scaling in S-species model ecosystems. *Physica A*, 286:337-344, 2000.

[81] J. Szymański. On a nonuniform random recursive tree. *Annals of Discrete Mathematics*, 333:297-306, 1987.

[82] M. S. Wheatland and P. A. Sturrock. Avalanche models of solar flares and the distribution of active regions. *The Astrophysical Journal*, 471:1044-1048, 1996.

[83] N. C. Wormald, “Differential Equations for Random Processes and Random Graphs”, *Annals of Appl. Prob.*, Vol 5, 1995, pp. 1217–1235.

[84] G. Yule. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London (Series B), 213:21-87 (1925).

[85] X. Zhu, J. Yu, and J. Doyle. Heavy tails, generalized coding, and optimal web layout. In *Proceedings of IEEE INFOCOM*, 2001.

[86] G. Zipf. **Selective Studies and the Principle of Relative Frequency in Language**. Harvard University Press, Cambridge, MA, 1932.

[87] G. Zipf. **The psycho-biology of language: an introduction to dynamic philology**. Houghton Mifflin Company, Boston, MA, 1935.

[88] G. Zipf. **Human Behavior and the Principle of Least Effort**. Addison-Wesley, Cambridge, MA, 1949.