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

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

Michael Mitzenmacher

TR-08-01

Computer Science Group
Harvard University
Cambridge, Massachusetts

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

Michael Mitzenmacher*
Harvard University
michaelm@eecs.harvard.edu

Abstract

Power law distributions are an increasingly common model for computer science applications; for example, they have been used to describe file size distributions and in- and out-degree distributions for the Web and Internet graphs. Recently, the similar lognormal distribution has also been suggested as an appropriate alternative model for file size distributions. In this paper, we briefly survey some of the history of these distributions, focusing on work in other fields. We find that several recently proposed models have antecedents in work from decades ago. We also find that lognormal and power law distributions connect quite naturally, and hence it is not surprising that lognormal distributions arise as a possible alternative to power law distributions.

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, e.g., [6, 8, 7, 13, 16, 18, 19, 21, 22, 24, 28, 29, 34, 35, 37, 38, 39, 50, 56].[1]

This paper was motivated by a recent paper by Downey [22] 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. I elaborate on this specific model in another paper [51]. Studying this work led me to learn more about 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 to disseminate what I have found.

*Supported in part by an Alfred P. Sloan Research Fellowship and NSF grant CCR-9983832.
[1] We apologize for leaving out countless further examples.

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, dynamic processes that generate the growth of the Web graph and result in power law distribution for in- and out-degrees have become the focus of a great deal of recent study. In fact, extremely similar models date back to at least the 1950's, and arguably back to the 1920's. Second, similar disagreements as to what type of distribution is a better fit for empirically determined distributions have been repeated across many fields over many years. The question of whether a lognormal or power law distribution best applies to income distribution, for example, also dates back to at least the 1950's. The issue arises for other financial models, as detailed in [48]. Similar issues continue to arise in biology [31], chemistry [54], ecology [3, 66], astronomy [67], and information theory [40, 57]. 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.

Another discovery from looking at previous work is that power law and lognormal distributions are intrinsically connected. Very similar and 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 arises and repeats itself across a variety of fields.

The purpose of this paper is to explain some of the basic 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 paper is intended to be accessible to a general computer science audience. 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. Also, it should be emphasized that this paper does not contain original work, but is a survey of the work of others.

2 Basic Definitions and Properties

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

If X has a power law distribution, then in a log-log plot of $\Pr[X \geq x]$, or 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. On a log-log plot the density function for the Pareto distribution also is 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 $-\infty < y < \infty$. Hence the density function for a lognormal distribution satisfies

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

and the 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 Y has mean μ and variance σ^2, where the meaning is clear. The lognormal distribution is skewed, with mean $\mathrm{e}^{\mu+\frac{1}{2}\sigma^2}$, median e^{μ}, and mode $\mathrm{e}^{\mu-\sigma^2}$. 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

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

If σ is sufficiently large, then the quadratic term above 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 A Model that Generates Power Law Distributions

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 graphs [6, 13, 16, 21, 34, 35, 37, 38]. Hence an important criterion for an appropriate random graph model is that it yields power law distributions on 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 [21, 35]. 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

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

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 $c_j \sim j^{-\frac{2-\alpha}{1-\alpha}}$, giving a power law. To see this, note that $c_j \sim j^{-\frac{2-\alpha}{1-\alpha}}$ implies

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

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 [68] 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 [61] 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.

It is worthwhile to point out that while these are the earliest references I have found to mathematical arguments explaining power law distributions, as one might expect from Simon's list of applications, power laws had been observed in a variety of fields for some time. The earliest apparent reference is to the work by Pareto [55] 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 [23], although generally the idea (and its elucidation) are attributed to Zipf [70, 71, 72]. 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 [5]. Lotka (circa 1926) found in examining the number of articles produced by chemists that the distribution followed a power law [42].

Mandelbrot had developed other arguments for deriving power law distributions based on information theoretic considerations somewhat earlier than Simon [44]. His argument is very similar in spirit to other recent optimization based arguments for heavy tailed distributions [14, 69]. We sketch Mandelbrot's framework. 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 get a power law for the p_j. Mandelbrot argues that a variation of this model matches empirical results for English quite well.

Indeed, Mandelbrot strongly argued against Simon's alternative assumptions and derivations. 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* [45, 62, 46, 63, 47, 64]. Economists, however, give the nod to Simon. Indeed, a recent popular economics text by Krugman [36] offers a derivation of the power law similar to that above.[2] A more academic treatment is given by Gabaix [25].

Finally, it is worth noting that before the Web graph became popular, the study of random trees had already led to power law distributions. 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 the 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 [43, 58, 65].

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. For instance, the current arguments based on martingales are much more rigorous than Simon's approach [12, 16, 38]. Indeed, the development of a 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 is extremely important for analysis [12, 65]. It has been shown that these models yield graphs with community substructures, a property not found in random graphs but amply found in the actual Web [34, 38]. The diameter of these random Web graphs have also been the subject of recent study [4, 11]. Still, it is interesting 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 A Model that Generates Lognormal Distributions

Lognormal distributions are generated by processes that follow what the economist Gibrat called the law of proportionate effect [26, 27]. 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, 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 them-

[2] 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 [9].

selves 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 [10]. In a simplified version of this setting [17, 30], 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 examples are given in [20]. More recently, Adamic and Huberman suggest that multiplicative processes may describe the growth of links on the Web as well as the growth of user traffic on Web sites [28, 29], and lognormal distributions have been suggested for file sizes [8, 7, 22].

The connection between multiplicative processes and the lognormal distribution can be traced back to Gibrat around 1930 [26, 27], although Kapteyn [32] described in other terms an equivalent process in 1903, and McAlister described the lognormal distribution around 1879 [49]. 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 which distribution gives the better fit was fully developed [1, 2]. 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 [52, 53], who offer a possible mathematical justification discussed below. Similar observations have been given for file sizes [8, 7].

5 Power Law versus Lognormal Distributions

Although the generative models of the power law and lognormal distributions given above appear different, they are actually very closely connected. Only small changes from the lognormal generative process modifies it to a heavy-tailed 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 [15], 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

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

or

$$\Pr[X \geq x] = m/x$$

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 [25, 33].

Interestingly, another seemingly minor variation on the multiplicative generative model also yields power law behavior, although this derivation is significantly more recent. 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 [52, 53]. 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 [28, 29].

In more recent independent work, Reed provides the correct full distribution for the above model, which yields what he calls a double Pareto distribution [59]. 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).[3]

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

[3]For completeness we note that Huberman and Adamic give an incorrect form of the density function; they miss the two-sided nature of the distribution. Reed gives the correct form, as we do below.

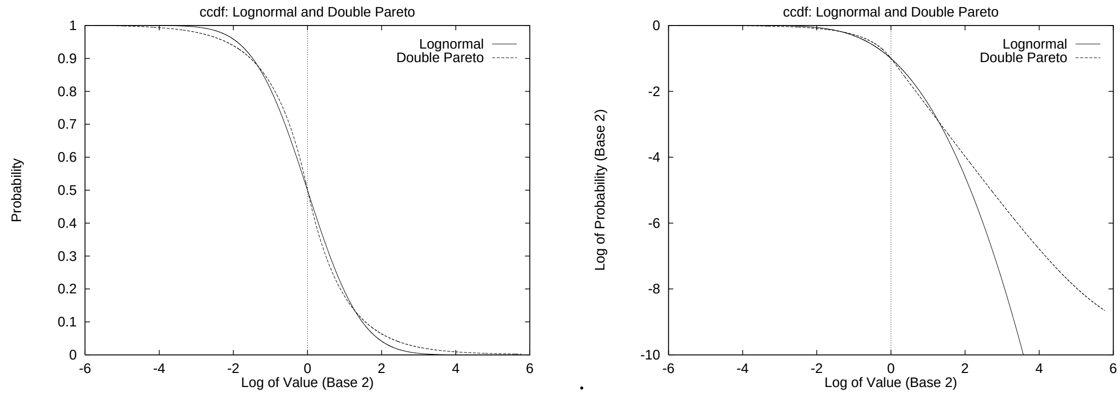

Figure 1: Shapes of lognormal and double Pareto distributions.

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, but on the log-log scale one can see the difference in the tail behavior.

Reed also suggests a generalization of the above called a double Pareto-lognormal distribution with similar properties [60]. 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 [1, 8, 7, 52, 53]. Reed presents empirical evidence for the double Pareto and double-Pareto lognormal distributions for incomes and other applications [59, 60].

More generally, the above result shows that natural mixtures of lognormal distributions may lead to power law distributions. Finding other interesting similar cases is an open problem.

6 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 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 [59, 60], 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. Determining guidelines for cases where a power law distribution cannot be suitably approximated by a lognormal for simulation or other practical purposes would be useful.

7 Acknowledgments

The author would like to thank Alan Frieze, Allen Downey, Mark Crovella, and John Byers for suggestions on how to improve this work. For further reading, the author strongly recommends the article by Xavier Gabaix [25], which provides 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 [48]. Wentian Li has a Web page devoted to Zipf's law which is an excellent reference [41]. For lognormal distributions, useful sources include the text by Aitchison and Brown [2] or the modern compendium edited by Crow and Shimizu [20].

References

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

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

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

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

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

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

[7] 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.

[8] P. Barford and M. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *Proceedings of ACM SIGMETRICS*, pp 151-160, 1998.

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

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

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

[12] 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.

[13] 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.

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

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

[16] C. Cooper and A. Frieze. On a general model of Web graphs. Submitted, 2001.

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

[18] 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.

[19] 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, pp. 3-26, Chapman and Hall, 1998.

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

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

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

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

[24] 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.

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

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

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

[28] 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*, 399, p. 130, 1999.

[29] B. A. Huberman and L. A. Adamic. The Nature of Markets in the World Wide Web. *Quarterly Journal of Economic Commerce*, vol 1., pp. 5-12, 2000.

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

[31] 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.

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

[33] H. Kesten. Random Difference Equations and Renewal Theory for Products of Random Matrices. *Acta Mathematica*, CXXXI:207-248, 1973.

[34] 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.

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

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

[37] 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.

[38] 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*, pp. 57-65, 2000.

[39] 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* pp. 1-15, 1994.

[40] W. Li. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory*, 38(6):1842-1845, 1992.

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

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

[43] H. Mahmoud, R. Smythe, and J. Szymański. On the structure of plane-oriented recursive trees and their branches. *Random Structures and Algorithms*, 3:255-266, 1992.

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

[45] 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.

[46] 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.

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

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

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

[50] 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.

[51] M. Mitzenmacher. Dynamic Models for File Size Distributions and Double Pareto Distributions. Preprint, available at http://www.eecs.harvard.edu/~michaelm/NEWWORK/papers/.

[52] 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.

[53] 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.

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

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

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

[57] 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.

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

[59] W. J. Reed The Pareto law of incomes - an explanation and an extension. Submitted to Journal of Business and Economic Statistics. 2000. Available at http://www.math.uvic.ca/faculty/reed/index.html.

[60] 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.

[61] H. A. Simon On a Class of Skew Distribution Functions. *Biometrika*, 42(3/4):425-440, 1955.

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

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

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

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

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

[67] 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.

[68] 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).

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

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

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

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