Double Pareto Lognormal Distributions in Complex Networks

Zheng Fang, Jie Wang, Benyuan Liu, Weibo Gong

Abstract This article elaborates the mathematical concept of double Pareto lognormal distribution and provides an overview of complex networks and natural phenomena that exhibit double Pareto lognormal distributions. These include the number of friends in social networks, the number of downloads on the Internet, Internet file sizes, stock market returns, wealth in human societies, human settlement sizes, oil field reserves, and areas burnt from forest wildfire.

1 Introduction

Power law distributions have been found in a good number of complex networks and natural phenomena of significant scientific interests. For example, population size distribution of cities, wealth distributions, intensities of earthquakes, and sizes of particles are all thought to follow the power law, and they cannot be correctly characterized by median or average values. For instance, the average population of

Z. Fang
Department of Computer Science, University of Massachusetts, Lowell, MA 01854. This author is supported in part by the NSF under grant CCF-0830314. e-mail: zfang@cs.uml.edu

J. Wang
Department of Computer Science, University of Massachusetts, Lowell, MA 01854. This author is supported in part by the NSF under grants CCF-0830314, CNS-0958477, and CNS-1018422. e-mail: wang@cs.uml.edu

B. Liu
Department of Computer Science, University of Massachusetts, Lowell, MA 01854. This author is supported in part by the NSF under grants CNS-0721626, CNS-0953620, and CNS-1018303. e-mail: bliu@cs.uml.edu

W. Gong
Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003. This author is supported in part by NSF under grant EFRI-0735974 and by Army Research Office under Contract W911NF-08-1-0233. e-mail: gong@ecs.umass.edu

all cities and towns is not a useful concept for most purposes because a significant fraction of the total population lives in big cities (e.g., New York, Los Angeles, and Chicago), which is substantially larger than those of many other cities by several orders of magnitude. Studies on complex networks such as the World Wide Web and online social networks also reveal that certain attributes of interests exhibit power law behaviors.

Pareto distribution, named after the Italian economist Vilfredo Pareto, is a commonly used canonical power law distribution. Pareto originally used this distribution to describe the allocation of wealth among individuals, for he observed that it depicts rather accurately the phenomenon that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. Pareto also used it to describe income distributions. This idea is sometimes referred to as the Pareto principle or the 80-20 rule, which says that 20% of the population controls 80% of the wealth [1]. Another popular power law distribution is the Zipfian distribution, also known as Zipf's law, which is widely cited in linguistic research of natural languages.

Pareto and Zipfian distributions only exhibit a single tail. It has come to people's attention in recent years that certain power-law phenomena, if observed more closely, actually exhibit two tails: a lower tail and an upper tail. While the two tails would share a similar shape with lognormal distributions, how to correctly characterize such phenomena remains a challenge, which has stimulated much debate among researchers. It is true that the lognormal model could better fit the bodies of empirical distributions, but it does not seem to fit well with the power law behavior in the tails.

The concept of double Pareto lognormal distributions has been shown useful in modeling distributions of various complex networks and natural phenomena that consist of a lognormal body and Pareto tails, including computer networks, social networks, economics, finance, geography, geology, and physical sciences [2, 3, 4]. In this article we will provide and elaborate mathematical background on the concept of double Pareto lognormal distribution, demonstrate how it would fit into empirical data, and present possible explanations of the causes for the power law behaviors.

The rest of this article is organized as follows. In Section 2 we will provide a brief overview of the classical power law and lognormal distributions. We will then introduce the double Pareto lognormal distribution and several models that generate it. In Section 3 we will present examples from diverse areas that fit the double Pareto (lognormal) distributions. These examples include wealth distributions in human societies, stock market returns, Internet file sizes, the number of friends distributions in social networks, downloads distributions on the Internet, human settlement sizes, oil field reserves, and areas burnt from forest wildfire. Section 4 concludes the article.

2 Generation Model of Double Pareto Distribution

We start this section by presenting the mathematical background for the power law distribution and lognormal distribution, and then derive the double Pareto lognormal distribution as an exponential mixture of lognormal distributions. We will also introduce several stochastic models of differential equations that converge to the double Pareto distribution under different sets of scenarios.

2.1 Power Law Distributions

Let X be a random variable. The complementary cumulative distribution function (ccdf) of X is defined by

$$\overline{F}_X(x) = Pr(X \geq x). \tag{1}$$

A non-negative random variable X is said to follow a power law distribution if its ccdf satisfies

$$\overline{F}_X(x) \sim x^{-\alpha} \tag{2}$$

for some constant $\alpha > 0$. Here $f(x) \sim g(x)$ denotes $\lim_{x\to\infty} f(x)/g(x) = c$ for some constant $c > 0$. The plot of the density of such random variable X has a long tail, which is referred to as *power tail* (a.k.a. *heavy tail*).

The Pareto distribution is one of the canonical power law distributions with the following ccdf on random variable X:

$$\overline{F}_X(x) = \begin{cases} (\frac{x}{x_m})^{-\alpha}, & x \geq x_m \\ 1, & x < x_m \end{cases} \tag{3}$$

for some $\alpha > 0$ and $x_m > 0$. Note that the Pareto distribution has the density function $f(x) = \alpha x_m^{\alpha} x^{-\alpha-1}$ when $x \geq x_m$. If α falls in the range of $(0,1]$, the function has an infinite mean. If α is in $(1,2]$, the function has a finite mean but infinite variance. Only when $\alpha > 2$ will the function have both finite mean and finite variance.

Another popular power law distribution is the Zipfian distribution (a.k.a. Zipf's law), which is widely used in linguistic studies of natural languages. Zipf's law states that frequency of occurrences of an event, as a function of the ranking of frequencies, is a power law function with the exponent α close to unity, where a higher frequency has smaller rank. For example, in the English language the word "the" has the highest frequency, and so its rank is equal to 1. The word "of" has the second highest frequency and so its rank is equal to 2. Zipf's law predicts that in a population of N elements, the frequency of elements that is ranked k is determined by the following formula:

$$f(k) = \frac{k^{-\alpha}}{\sum_{n=1}^{N} n^{-\alpha}}. \tag{4}$$

where α is the value of the exponent characterizing the distribution, typically close to 1.

It is customary to plot the power law distribution by taking logarithm on both sides, referred to as log-log plot, which produces a straight line for the asymptotic behavior of the ccdf. This is the basis for testing power-law behaviors. The same is true for the power law density function, which might be easier to work with mathematically under certain circumstances. For example, for the Pareto distribution, the logarithm of the density function is linear with the following form:

$$\ln f(x) = -(\alpha+1)\ln x + \alpha \ln x_m + \ln \alpha \tag{5}$$

Plotting ccdf in the logarithmic scale also emphasizes the tail region, providing a good visual when fitting empirical data into a power law model.

2.2 Lognormal Distributions

A positive random variable X is said to be lognormally distributed with parameters (μ, σ^2) if the random variable $Y = \ln X$ is normally distributed with mean μ and variance σ^2. The density function for a lognormal distribution is determined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{6}$$

The lognormal distribution, in contrast to the normal distribution, is skewed, with median e^{μ}, mean $e^{\mu+\frac{1}{2}\sigma^2}$, and mode $e^{\mu-\sigma^2}$. Although the lognormal distribution has finite moments and the Pareto distribution has infinite moments, their plot shapes are extremely similar in that a large portion of the body of the density function and the ccdf can appear linear. To be more specific, take logarithm on a lognormal distribution density function, we have

$$\ln f(x) = -\ln x - \ln \sqrt{2\pi}\sigma - \frac{(\ln x - \mu)^2}{2\sigma^2}. \tag{7}$$

When σ is sufficiently large, the value of $\ln f(x)$ is barely affected by the quadratic term for a large range of x values. Therefore, the logarithm of the density function will appear as a straight line for a large range of x values. The same holds true for the ccdf.

While the normal distribution can be thought of as an additive accumulation of a number of independent random variables, a variable could be thought of as lognormal if it is the multiplicative product of a number of independent, positive random variables, for

$$\ln Y = \ln X_1 + \ln X_2 + \cdots + \ln X_n = \ln(X_1 X_2 \cdots X_n) \tag{8}$$

For example, a long-term discount factor in a financial market can be derived from the product of short-term discount factors. In wireless communications, for another example, the attenuation caused by shadowing or slow fading from random objects is often assumed to be lognormally distributed.

This phenomenon may be better illustrated using the following example. Suppose we start with an initial state X_0. At each step, the state may change with a certain percentage (e.g., changes of price, size, and volume) denoted by an independent random variable C_i. Thus, each state X_t can be described as

$$X_t = X_{t-1}C_t = X_0 \prod_{i=1}^{t} C_i. \tag{9}$$

Taking logarithm on both sides we have

$$\ln X_t = \ln(X_{t-1}C_t) = \ln X_0 + \sum_{i=1}^{t} \ln C_i. \tag{10}$$

Since each C_i is independent from each other, applying the Central Limit Theorem to the summation term yields a convergence to a normal distribution for sufficiently large t. Hence, random variable X_t is well approximated by a lognormal distribution. Lognormal distributions are natural distributions for modeling growth of population, growth of wealth, growth of organisms, and any process where the underlying change rate is a random factor over a time step independent of the current size.

2.3 Double Pareto Distribution, A Mixture of Lognormals

A double Pareto distribution can be generated by mixing a number of lognormal distributions. This can be done based on a model defined in Section 2.2 that yields a lognormal distribution.

Suppose in $X_t = X_{t-1}C_t$ we have $X_0 = c$ for a constant $c > 0$, and C_t is a lognormally distributed random variable with parameters (μ, σ^2). We may view the subscript of X as a moment in time, then at time $t = T$, the random variable X_T is also lognormally distributed with parameters $(\mu T, \sigma^2 T)$. We may also view the time t itself as a random variable and let the process run for some random time k and obtain a random variable that comes from a mixture of lognormal distributions with parameters $(\mu k, \sigma^2 k)$. Specifically, if the process stops at a constant rate λ, the time variable is exponentially distributed with density

$$f_t(k) = \lambda e^{-\lambda k}, \quad k \geq 0.$$

Reed et al [5] show that this mixture of an exponentially distributed number of lognormal distributions exhibits power law behavior for both the upper tail and the lower tail regions, and name it double Pareto distribution. The resulting density function for this mixture is

$$\begin{aligned} f(x) &= \int_{k=0}^{\infty} \lambda e^{-\lambda k} \frac{1}{\sqrt{2\pi k}\sigma x} e^{-\frac{(\ln x - k\mu)^2}{2k\sigma^2}} dk \\ &= \frac{2\lambda e^{\mu \ln x/\sigma^2}}{\sqrt{2\pi} x\sigma} \int_{u=0}^{\infty} e^{-\left(\lambda + \frac{\mu^2}{2\sigma^2}\right)u^2} e^{-\frac{\ln^2 x}{2\sigma^2 u^2}} du \\ &= \frac{\lambda}{\sigma\sqrt{(\mu/\sigma)^2 + 2\lambda}} \begin{cases} x^{-1+\mu/\sigma^2+\sqrt{(\mu/\sigma)^2+2\lambda}/\sigma}, & 0 < x \leq 1 \\ x^{-1+\mu/\sigma^2+\sqrt{(\mu/\sigma)^2+2\lambda}/\sigma}, & x \geq 1. \end{cases} \end{aligned}$$

Reed [5] suggests the following simpler form of double Pareto distribution density function:

$$f(x) = \frac{\alpha\beta}{\alpha+\beta} \begin{cases} x^{\beta-1}, & 0 < x \leq 1 \\ x^{-\alpha-1}, & x \geq 1 \end{cases} \tag{11}$$

where α and $-\beta$ ($\alpha > 0, \beta > 0$) are the two roots of the following quadratic equation

$$\frac{\sigma^2}{2} z^2 + \left(\mu - \frac{\sigma^2}{2}\right) z - \lambda = 0.$$

Instead of using exponential distribution, Mitzenmacher [22] shows that using a geometric distribution to randomly sample lognormal distributions can lead to a distribution that is extremely similar to a double Pareto distribution. This approach assumes that the random process stops with a probability p. That is, the process stops at time k with probability $p(1-p)^{k-1}$. Using this discrete geometric mixture we can obtain the following distribution density:

$$f(x) = \sum_{k=1}^{\infty} p(1-p)^{k-1} \left(\frac{1}{\sqrt{2\pi k} x\sigma} e^{-\frac{(\ln x - k\mu)^2}{2k\sigma^2}} \right). \tag{12}$$

This summation can be nicely approximated using the following integral when the absolute value of $\ln x$ is sufficiently large.

$$f(x) \approx \int_{k=1}^{\infty} \frac{p}{\sqrt{2\pi k} x\sigma(1-p)} e^{k\ln(1-p) - \frac{(\ln x - k\mu)^2}{2k\sigma^2}} dk. \tag{13}$$

The geometric approach produces essentially the same power-tail behavior as the exponential mixture does. Technically, the geometric mixing of lognormal distributions only yields an approximation to the double Pareto distribution according to Reed's definition. However, the tails of the cumulative distribution function (cdf) and the ccdf of the geometric mixture are each bounded by two power law distributions that differ only by a constant factor [22]. Thus, the geometric mixture produces a valid and practical approximation to the double Pareto distribution.

To the extent of double Pareto distribution, Reed [5] also suggests a more generalized form called double Pareto lognormal distribution, by removing the constraint that the initial state X_0 must be a constant. Assume that the initial state X_0 also follows a lognormal distribution with parameters (ν, τ^2). Mixing with the exponential-time distribution we can show that the distribution of random variable $Y = \ln X$ can be represented as the sum of two independent random variables, where one variable is normally distributed and the other is double exponentially distributed. It follows that the density function $f_X(x)$ can be obtained from the density function $f_Y(y)$ using $f_X(x) = f_Y(\ln x)/x$, which in turn can be found by convolving a normal density and a double exponential density. The final density function of X is

$$\begin{aligned} f_X(x) &= \frac{\alpha\beta}{\alpha+\beta}(A+B) \\ A &= x^{-\alpha-1} e^{\alpha\nu+\alpha^2\tau^2/2} \Phi\left(\frac{\ln x - \nu - \alpha\tau^2}{\tau}\right) \\ B &= x^{\beta-1} e^{-\beta\nu+\beta^2\tau^2/2} \Phi^c\left(\frac{\ln x - \nu + \beta\tau^2}{\tau}\right) \end{aligned} \tag{14}$$

where Φ is the cdf of the standard normal distribution and Φ^c the compliment of Φ, that is, Φ^c is the ccdf of the standard normal distribution. Figure 1 shows the density of the double Pareto lognormal distributions with $\beta > 1$ and $\beta < 1$ (Note that changing the value of α does not alter the general shape of the distribution).

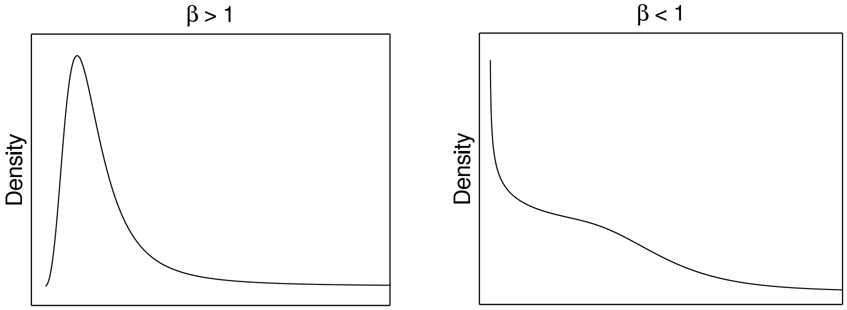

Fig. 1 The double Pareto lognormal distributions have completely different shapes with $\beta > 1$ and $\beta < 1$.

The double Pareto lognormal distribution provides a reasonable model to describe a random process that allows random variables to start from different initial values, as long as they obey the same lognormal distribution. Releasing the initial value constraint would make the model more useful in empirical studies. For example, in surveying personal wealth accumulations, it would be more reasonable to assume that different people begin their asset accumulation with different starting salaries. Thus, the double Pareto lognormal distribution might allow closer matches with empirical data distributions.

The double Pareto distribution exhibits power law behavior at both the upper and lower tails. That is, both the ccdf and cdf each has a linear tail on a log-log plot. This is an important characteristic, which is often used to test if a distribution has a double Pareto distribution. That is, check whether the ccdf and the cdf of a distribution each have a linear tail on a log-log plot.

The double Pareto distribution falls nicely between the lognormal distribution and the Pareto distribution. The Pareto distribution and the double Pareto distribution both are power law distributions, but they have the following distinction: The log-log plot of the density function of the Pareto distribution is a single straight line, and the log-log plot of the density function of the double Pareto distibution 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. Hence, an appropriate double Pareto distribution can closely match the body of a lognormal distribution and the tail of the Pareto distribution. Figure 2 shows the ccdfs of lognormal distribution, Pareto, and double Pareto distributions in the log-log plot. We can see that the double Pareto distribution matches well with the lognormal distribution in the body, and matches well with the Pareto distribution in the tail.

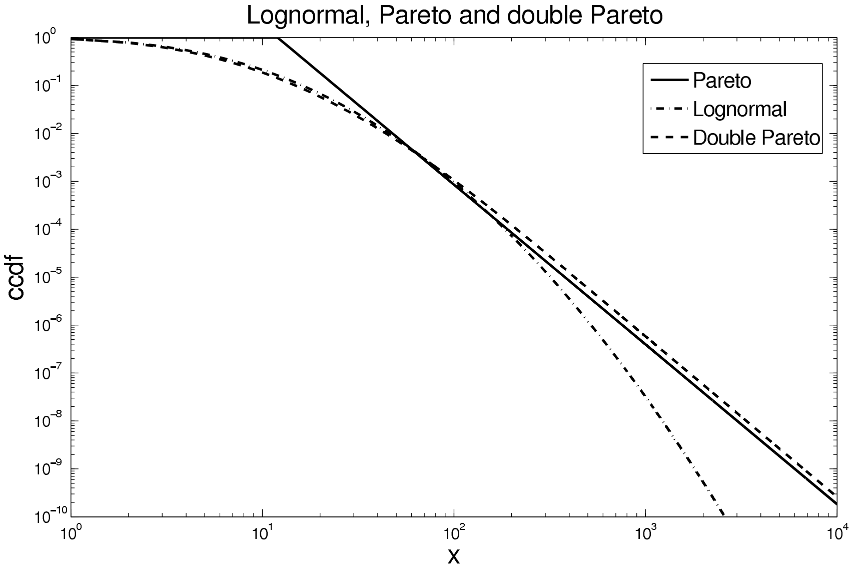

Fig. 2 Tail comparisons of lognormal, Pareto, and double Pareto.

2.4 Power Law through Stochastic Differential Equations

The lognormal distribution plays an essential role in generating a double Pareto (lognormal) distribution. However, how a lognormal distribution is generated is an interesting topic in its own right. More specifically, one would like to know what kind of processes that model natural phenomena will result in a lognormal distribution.

A stochastic process, a.k.a. a random process, describes the probability distribution of possible realities of how the process might evolve over time [6]. A stochastic differential equation (SDE) is a differential equation in which one or more of the terms is a stochastic process, thus resulting in a solution that is itself a stochastic process. SDEs are used to model diverse phenomena such as personal income figures, human settlement sizes, fluctuating stock prices, or physical systems subject to thermal fluctuations. Typically, SDEs incorporate white noise that can be thought of as the derivative of the Brownian motion (or the Wiener process). However, it should be mentioned that other types of random fluctuations are also possible, including jump processes and Poisson counters.

The first SDE we would like to discuss is a Geometric Brownian Motion (GBM) process. Reed [2, 5, 3, 10] uses this model to explain the double Pareto lognormal distribution. A random variable X is said to follow GBM if its behavior over time is governed by the following differential equation

$$dX = (\mu dt + \sigma dB)X, \tag{15}$$

where dB is the increment of a standard Brownian motion (a.k.a. the white noise). For a GBM the proportional increment of X in time dt comprises a systematic component μdt, which is a steady contribution to X, and a random component σdB, which is fluctuated over time. Thus the GBM can be seen to be a stochastic version of simple exponential growth.

Assuming a constant initial state X_0. Applying Itō's lemma [8] on this SDE produces the following equation at time $t = T$:

$$\ln X_T = \ln X_0 + \left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma B_T. \tag{16}$$

Since B_T is normally distributed with parameters $(0, T)$, it is evident that the random variable $\ln X_T$ is also normally distributed with mean

$$\ln X_0 + \left(\mu - \frac{1}{2}\sigma^2\right)T$$

and variance $\sigma^2 T$, which means that the random variable X_T is lognormally distributed with the same set of parameters. Since in a random process the stopping time (or a starting time) T for each individual instance may be different, it is a random variable. If it is exponentially distributed, then it can be shown that, as discussed in Section 2.3, the mixture of states follows a double Pareto distribution. If we further assume that the initial state X_0 follows a lognormal distribution, the distribution of X becomes a double Pareto lognormal distribution.

In addition to the GBM model, Jiang et al. [9] propose several other generalized forms of first order SDEs that also exhibit power law behaviors. In particular, they consider a different scenario involving the steady state density associated with an

SDE. The SDE describes a scenario where the quantity of interest decays to zero exponentially, that is,

$$dX = -\alpha X,$$

but is incremented by a fixed amount of σ at a random moment of time. Under the assumption that the random variable of time is exponentially distributed, they show that for a range of parameter values the steady state distribution of X exhibits a lower-tail power law, and through a simple transformation $Y = X^{-1}$ the distribution can be converted into an upper-tail power law. In their model, the random fluctuation component is implemented by a Poisson counter.

Poisson counter was first studied by Brockett with the following SDE [7],

$$dX = -\alpha X dt + \sigma dN, \tag{17}$$

where $\alpha > 0, \sigma > 0$, and N is a Poisson process of intensity λ.

Jiang et al. [9] showed that, through the transformation that converts the lower-tail power law into an upper-tail power law, the resulting distribution converts to a Pareto distribution as $t \to \infty$.

It is also possible to let Brownian motion and Poisson counter co-exist in the SDE. Adding a Brownian motion component, the SDE becomes

$$dX_t = (\mu dt + \sigma dB_t)X_t + (x_0 - X_{t-})dN_t, \tag{18}$$

which is a geometric Brownian motion with Poisson jumps that always reset the motion to a fixed state x_0. This SDE is similar to the one analyzed by Reed [2, 5], and the result is also similar, for Reed showed that as $t \to \infty$ the steady-state density distribution of X_t converts to a double Pareto distribution.

Jiang et al. [9] also studied the power-law behavior near a critical point and derived an SDE comprises of two independent bi-directional Poisson counters to demonstrate this behavior. They showed that a discontinuity at this critical point occurs in a surprising way, which might be of interest in statistical physics.

Early studies have all indicated that random multiplication with exponentially distributed stopping time will lead to power law behaviors. This result may serve as a guidelines in explaining real-world phenomena that obey the power law.

3 Power-Law Behaviors in Complex Networks and Natural Phenomena

The most effective way to evaluate the accuracy a power-law model is to use it to explain the cause of certain power-law phenomenon and validate the prediction using empirical data. The goodness of fit is a typical measure of a theoretical model. A good model should make this an intuitive process. In this section, we will present the possible explanations to various real-world complex networks and natural phenomena that exhibit power-law behaviors, and demonstrate the goodness-of-fit figures with empirical data.

3.1 Income

It is well known that the distribution of a population's income and wealth follows the power law. The original observation was made in 1907 by Pareto. He noticed that 80% of the wealth in Italy was owned by 20% of the population [1]. He then surveyed a number of various types of countries and found to his surprise that a similar income distribution applied. For over a century most of the studies have been focusing on the distribution itself as well as the impacts imposed by Pareto's principle. Not much has been done to explain the underlying reasons that cause this phenomenon. Reed's GBM model for the double Pareto lognormal distribution based was the first-known model that provides an intuitive explanation for the income distribution.

The distribution of incomes over a population is the same as the probability distribution of the income of an individual randomly selected from that population. Thus a stochastic model for the generation of the income of such an individual can be used to explain the observed distribution of incomes in a population or in random samples. For an individual, the more income he or she currently receives, the more income he or she will accumulate in the next time interval based on an expected rate of increase (e.g., interests of deposits in saving accounts and annual raises of salaries). Similarly, finance uncertainties (e.g., good or bad investment, market depression, marriage or divorce) will directly affect his or her income as well. This argument suggests that the income behavior over time could be modeled as a GBM model, which was discussed in Section 2.4 and is presented here for convenience:

$$dX = (\mu dt + \sigma dB)X,$$

where μ represents the instantaneous rate of increase on a riskless asset, σ the volatility of the income, and dB the infinitesimal change in a Brownian motion over the next instant of time (a.k.a. white noise). Assume that individual's income follows GBM process X with initial state X_0 being a constant, then for a randomly selected individual from the group of people with the same working time T, his or her income is lognormally distributed.

If an individual is randomly selected from the entire workforce, the time T that he or she has been in the workforce will be a random variable. To find the distribution of time T, we consider a simple case when the workforce or population is growing at a constant rate ν. We assume that all individuals will eventually merge into the workforce at certain time. A simple analysis gives us the following equalities:

$$F_T(t) = 1 - Pr(T \geq t)$$

$$\begin{aligned} &= 1 - \frac{N_{T-t}}{(1+\nu)^t N_{T-t}} \\ &= 1 - (1+\nu)^{-t}, \end{aligned}$$

$$\begin{aligned} f_T(t) &= F_T'(t) \\ &= \ln(1+\nu)(1+\nu)^{-t} \\ &= \lambda e^{-\lambda t}. \end{aligned}$$

In this case the time T has an exponential distribution with a probability density function $f_T(t) = \lambda e^{-\lambda t}$. Therefore, the income distribution from the entire workforce will be an exponential mixture of lognormal distributions, leading to a double Pareto distribution. Now let us consider the initial state X_0. It would be more realistic to assume that individuals' initial incomes will also vary and evolve over time, which can be described by another GBM. In this case, the income distribution is changed to a double Pareto lognormal distribution [5].

Figure 3 demonstrates the double Pareto lognormal distribution fitted to empirical income data (originally fitted by Reed [10]): the United States household income of 1997 and the Canadian personal earnings of 1996, respectively. The data fits in the theoretical curves quite well, not only in the upper tail region, but also in the lower tail region.

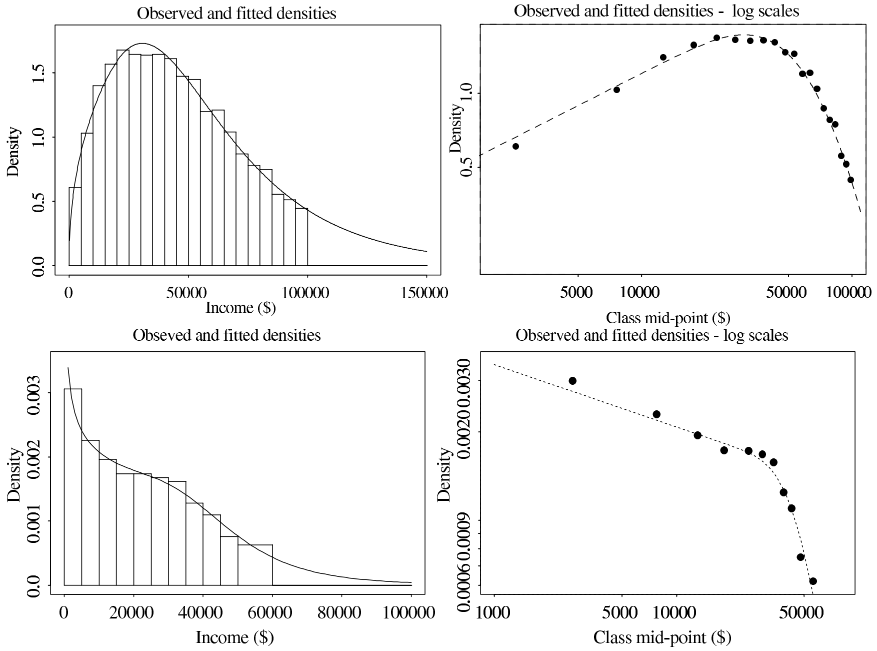

Fig. 3 The double Pareto lognormal distribution fitted to the US household income (1997) data and the Canadian personal earnings (1996) data [10]

3.2 Stock Market Returns

Stock market returns is another example that fits the double Pareto distribution and can potentially be explained using a GBM model. During the last several decades researchers have investigated the statistical distribution of returns and have concluded that returns are “fat tailed”, which is a key power law characteristic. Several studies have been devoted studying to the stochastic behaviors of the stock price changes [12, 13, 14, 15, 16]. Bachelier in his 1900 dissertation [11] suggested that stock prices in successive periods follow a random process that is best described by a Wiener process (i.e., a Brownian motion).

This suggests that we could treat the return rate of each transaction as an independent random variable. Thus, at certain time the total return could be modeled as a sequence of multiplication of such independent random variables with initial values, which is analogous to a long-term price discount, except that it should not always be a discount. The stochastic multiplicative process yields a lognormally distributed variable representing the stock returns at a fixed time. If we consider a random killing process with a constant killing rate, by killing here it means to cash out the stocks and settle, and the returns is observed at the killing time, then the time span a process is kept alive is exponentially distributed. Thus, if we observe the population of stock returns, the distribution will be the mixture of lognormal distributions with exponentially distributed time span, which is a double Pareto distribution. The initial wealth can be viewed as any wealth naturally accumulated at certain time, which is lognormally distributed as the result of multiplicative process. Hence, the distribution of stock returns can be extended to the double Pareto lognormal distribution.

Figure 4 shows a good fit to the returns of IBM’s ordinary stocks from Jan 1, 1999 to Sept 18, 2003, originally plotted by Reed [5]. The figure at the left-hand side shows a density histogram and the fitted double Pareto lognormal density; the figure at the right-hand side shows the fitted density in logarithmic scale.

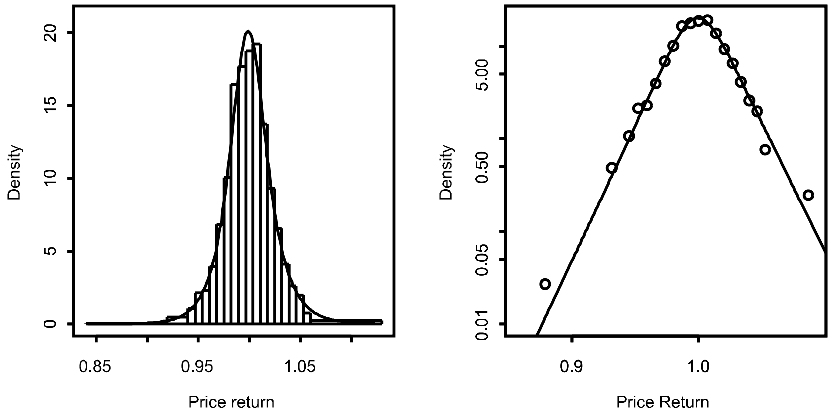

Fig. 4 The double Pareto lognormal distribution fitted to stock market returns [5]

3.3 Internet File Sizes

It has been observed that the file size distribution over the internet seem to follow a double Pareto distribution [17, 18, 19]. Figure 5 shows a file size statistics from a Web server at the University of Calgary, from traces collected by Arlitt and Williamson [20]. There have been attempts to explain the file size distribution. Among them, Downey's multiplicative file size model [21] and Mitzenmacher's recursive forest file model [22] have received much attention. The latter can be viewed as an improvement over Downey's model by introducing a dynamic insertion and deleting process.

Downey's model [21] is based on the following observation: Users tend to create new files from old files by copying, editing, or filtering in some way, and the size of the new file will differ from that of the old file by a multiplicative factor f from a given distribution D. That is, the system begins with a single root file and repeatedly performs the following actions: randomly choose a file and create a new file from it with size equal to fs, where s is the size of the chosen file.

The assumption behind this model is that creating a new file from a template file through copying, editing, or filtering will cause the size of the new file to differ from that of the old file by a factor from a given distribution D. For any file in this system, the history of the creation can always be traced back to the original root file. Thus, the size of the file can be viewed as the result of a random multiplicative process. Downey therefore suggest that the entire file size distribution is lognormal. Downey's model, however, does not address the situations of insertion and deletion, and the result of empirical fitting on these two operations is not satisfiable.

To overcome this problem, Mitzenmacher [22] introduces a recursive forest file model by modifying Downey's model to include dynamic insertion and deleting process. In this model, the system begins with a collection of one or more files, whose sizes are drawn from a distribution D_1. New files are generated repeatedly as follows.

1. With probability γ, add a new file with size chosen from a given distribution D_1.

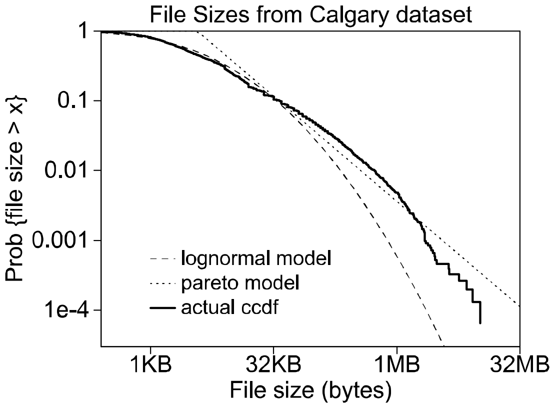

Fig. 5 ccdf of file sizes statistics from University of Calgary Web server. Compare with a lognormal distribution and a Pareto distribution [18]. The empirical data has the lognormal body and Pareto tail, which indicates a double Pareto distribution.

2. With probability η, select a file uniformly at random, delete this file.
3. With probability $1-\gamma-\eta$, select a file S with size s uniformly at random, choose a multiplicative factor f from a given distribution D_2, and create a new file S' with size fs.

Mitzenmacher reasons [22] that the file size density in this model converges to a double Pareto distribution if D_2 is a lognormal distribution D_2, and a double Pareto lognormal distribution if D_1 is also a lognormal distribution. This model explains why file size distribution may appear to have a lognormal body and a Pareto tail, making it more appealing compared to other attempted explanations. It also provides a flexible framework that may be extend to handle additional operations in the file system.

3.4 Friends in MySpace

The observation that complex networks often exhibit power-law behaviors has attracted much attention in recent years [18, 19, 23, 24]. Since Huberman and Adamic [25] suggested in 1999 that the exponential growth of the World Wide Web network could explain its power-law degree distribution, many studies have attempted to migrate this idea to explain other complex networks [26, 27, 28, 29]. The rise of online social networks has generated overwhelmingly huge amount of data, making it possible to carry out quantitative analysis on human social behaviors in a large scale.

Ribeiro et al. [30] recently investigated MySpace and showed that the distribution of the number of friends follows a double Pareto like distribution, which is shown in Figure 6.

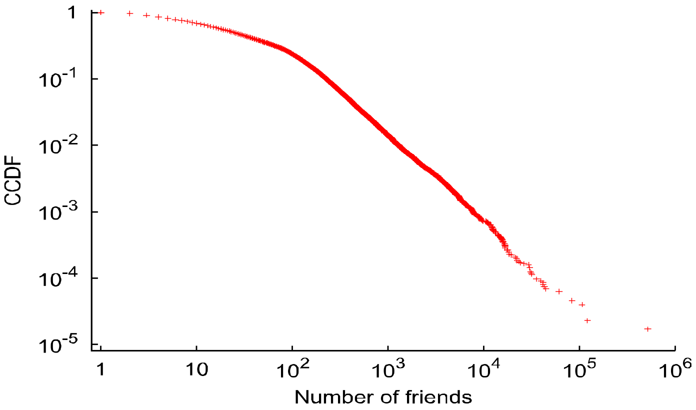

Fig. 6 Empirical ccdf of the number of friends in MySpace, plotted in the log-log scale [30]

Making friends in a virtual world is much easier, for the click of “add as friend” button really does not need any social skill. Thus, the meaning of friends in an online social network may be weak, which may simply mean “somehow related”. However, this observation still suggests a reasonable assumption that for a user with a large number of friends, most of the friends are added through passive referrals, assuming that all requests of making friends are automatically approved. In addition, if we assume that every user has a different referral probability to different people and the referral is always accepted, the growth of the number of friends can be modeled roughly as a multiplicative process $X_t = P_t X_{t-1}$, where each P_t can be computed as the average referral probability of the current group of friends. According to the analysis presented in Section 2.2, for a fixed period of time the number of friends follows a lognormally distribution.

Ribeiro et al. [30] also observe that the time span of MySpace accounts is distributed exponentially. This scenario might be explained by assuming a fixed increase rate of MySpace accounts. Combining a exponentially distributed time span with the lognormally distributed number of friends at fixed time, the overall friend distribution for all MySpace accounts converges to a double Pareto distribution.

3.5 Downloads from SourceForge

SourceForge is a Web-based source code repository that provides a centralized location for software developers to control and manage open source software development. As of February 2009, the SourceForge repository hosts more than 230,000 projects and has more than 2 million registered users.

Hunt and Johnson [31] use SourceForce as a downloading platform and study the statistics of downloads. The data on the number of downloads was collected from all projects listed on the most active project list on October 22, 2001 for 30 days, which was partially shown in Figure 7. The distribution obtained is heavily skewed which exhibits a significant sign of power tail. Taking a closer look at the few download region, it is evident that the download distribution also exhibits a lower tail, which implies that the double Pareto model could be able to provide an explanation to both of these tails.

It is reasonable to assume that during these 30 days every download for a single project comes from different users for the following reasons: People would not download the same file over and over again unless the download is not successful or the file has being updated, and the 30 days time span is considered to be a relatively short cycle that a file only has a small chance of being altered. Popular projects would attract more user downloads and people tend to broadcast satisfying user experience. Thus, the popular downloads are more likely to be introduced to new users and become even more popular, which would in turn increase the download counts. This process is similar to the growth of the number of MySpace friends. The same model can be applied to conclude that the number of downloads of projects by the same age group would exhibit a lognormal distribution.

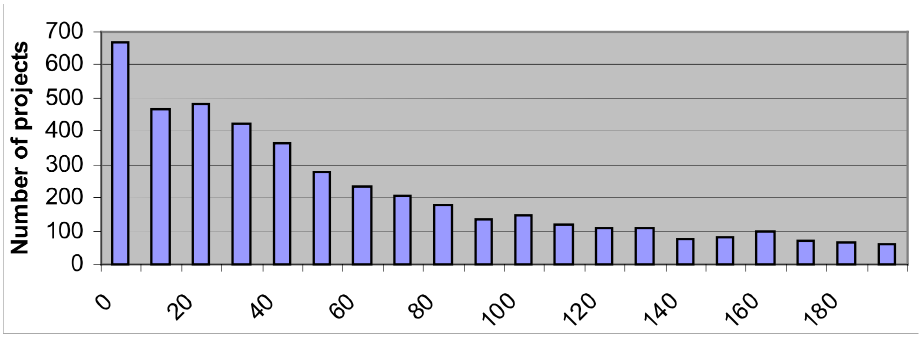

Fig. 7 The distribution of downloads in 30 days for active projects [31]

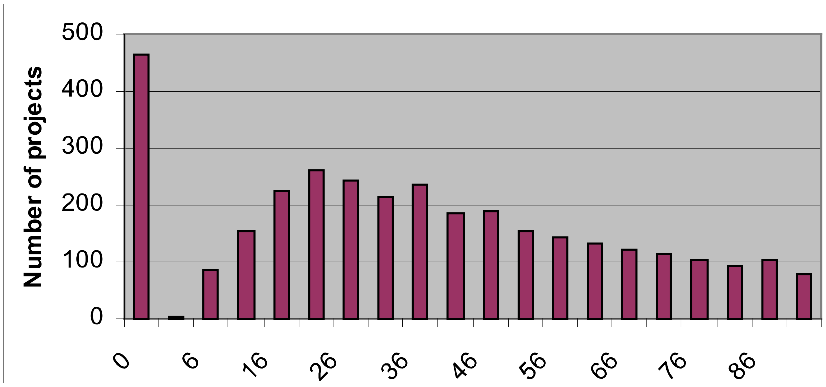

Fig. 8 A closer look at the distribution of downloads [31]

No public data is available concerning the distribution of ages for all the projects. If we assume that the total number of projects increase approximately at a fix rate, then the age of users would follow a exponential distribution. Therefore, the download distribution over all active projects is a double Pareto distribution, which exhibit both the lower tail and the upper tail.

3.6 Sizes of Human Settlements

Auerbach (1913) was the first to discover that the distribution of city sizes can be well approximated by a power-law distribution. That is, if we rank the cities based on population, then the size of the largest city is twice as that of the second largest city, three times as that of the third largest city, and so on. This distribution was believed to follow Zipf's law, a.k.a. the rank-size distribution. A number of studies have since contributed evidence and provided support to this statement until recently [34, 35, 37, 38]. In a wide spread article [32, 33] Eeckhout points out that the old evidence is problematic since the early studies only dealt with truncated samples and only focused on large cities, and the Zipf's law does not hold when taking all settlements of a country into consideration. The Zipf's law is still valid for the tail behavior since the tail mostly consists of large cities. When adding more cities of smaller sizes to the distribution, it gradually shows a lognormal body. Eeckhout [32, 33] then model the growth of settlements using the pure form of Gibrat's law (a.k.a Gibrat's rule of proportionate growth) and generate lognormal size distribution. Gibrat's law states that the size and growth rate are independent. However, the lognormal distribution does not fit well with the power-law tail.

The double Pareto lognormal seems more appropriate since it comprises a lognormal body and power law tails. Reed [36] suggests a GBM model, similar to the one that models personal incomes, for obtaining the settlement size distribution. Individual human settlements grow in many different ways. At the macro level a GBM process can be used to model the size growth by assuming a steady systematic growing rate and a random component. The steady growing rate reflects the average growth rate over all settlements and times, and the random component reflects the variability of the growth rate. The time when a city is founded varies from settlement to settlement. If we assume in the time interval $(t, t+dt)$ any existing settlement can form a new satellite settlement with probability λdt, the creation of settlements is a Yule process [39], which was first proposed as a model for the creation of new biological species. Under Yule process, the expected number of settlements is $e^{\lambda t}$ after t time since the first settlement. That is, the number of settlements is growing at rate λ. Therefore, the existing time for all settlements is exponentially distributed. It is straightforward to conclude that under GBM and Yule processes, the overall settlements size distribution will is a double Pareto distribution. If we further assume a lognormal initial settlement size, the result will converge to the double Pareto lognormal distribution.

Figure 9 shows empirical and fitted double Pareto lognormal distribution originally plotted by Reed [36] on the West Virginia data and California data obtained in 1998. The left-hand panel demonstrates the empirical density histogram in logarithmic size scale and fitted theoretical curve. The right-hand panel shows the fitted double Pareto lognormal distribution density in the log-log scale. It is evident that the double Pareto lognormal distribution provides a nice fit to the data in each region.

3.7 Volumes of Oil Field Reserves

The distribution of oil field sizes (i.e., the volumes of oil field reserves) has been the subject of much study for decades. Allais [40] was the first to propose to use a lognormal distribution for mineral resources and Kaufman [41] used this distribution for a population of oil or gas field in a petroleum basin. After that, the Pareto distribution has also been commonly used [42, 43, 44, 45] since the petroleum exploration practice has indicated that the probability of discovering large oil pools is low while

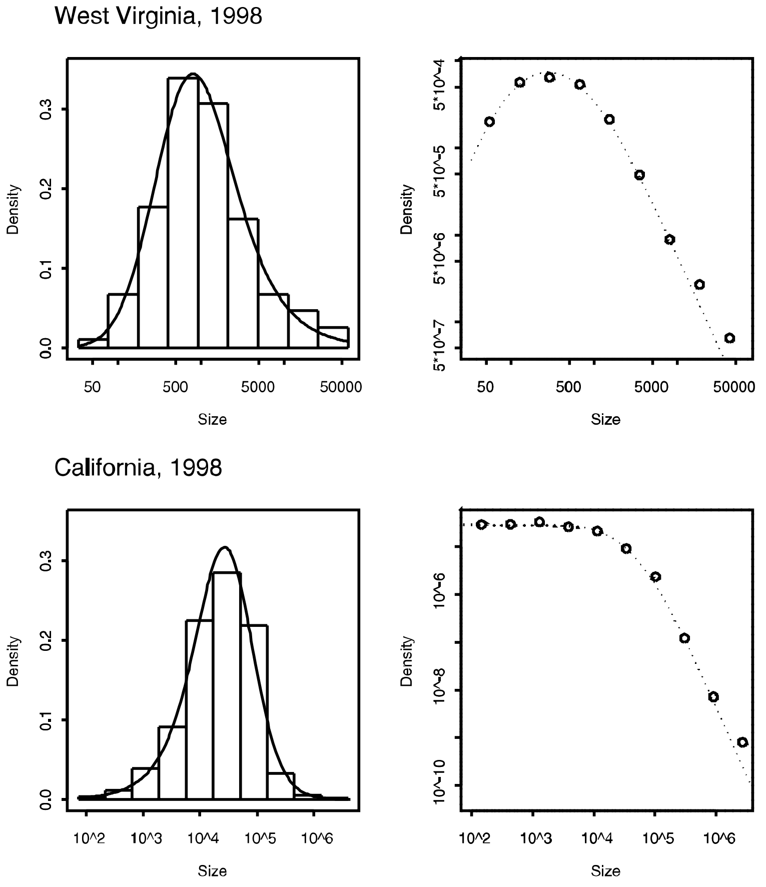

Fig. 9 The double Pareto lognormal distribution fitted to empirical city size data [36]

the probability of discovering medium and small-sized pools is high and the Pareto distribution is consistent with this type of structure. However, distinguishing lognormal and Pareto features that are both shown in the oil field size distribution may be difficult, which suggests that the double Pareto lognormal distribution might be a better choice.

An oil field can be thought of as a percolation cluster. The percolation theory provides a useful model of connectivity and dynamics in complex geometries(see [46] for a comprehensive introduction). The typical problem of percolation is to consider a lattice of $n \times n$ sites, each of which is either occupied or unoccupied with a certain probability. Clusters are formed when neighboring sites are occupied. The objective is to understand the relationship among groups of clusters. If we think of an oil field as a percolation cluster, the growth of oil field sizes can then be considered as a stochastic process of merging adjacent regions. Thus, the growth of an oil field size can be assumed to be proportional to its current size, for larger oil fields would have more possible regions to merge, which implies that the size distribution follows a lognormal distribution for a fixed percolating time.

The initial percolation cluster was formed by burying a huge amount of organically rich materials (e.g., plants and animal bodies), which could be caused by geologic hazards such as earthquake, landslides, and mudflows. Such extreme event occurs with small probability and percolation clusters are formed randomly with a small constant rate. Thus, the total percolating time span for a cluster is distributed exponentially. If we also assume a lognormal distribution on the initial burying amount, the overall size distribution of oil field would be a double Pareto lognormal distribution.

Figure 10 shows empirical and fitted double Pareto lognormal distribution by Reed [5] for the volume of 634 oil fields in the West Siberian basin. The left-hand figure demonstrates the empirical density histogram in logarithmic volume scale and fitted theoretical curve. The right-hand figure shows the fitted double Pareto lognormal distribution density in the log-log scale.

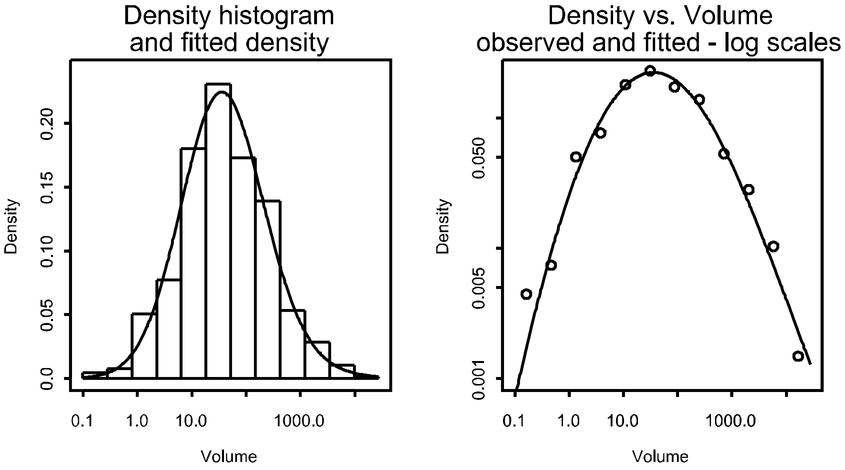

Fig. 10 The double Pareto lognormal distribution fitted to sizes of oil fields [5]

3.8 Areas Burned from Forest Wildfire

A number of researches have examined the distribution of burned areas from forest fires [47, 49, 50, 51, 52, 53] and claimed that it exhibits power-law behaviors, as shown in Figure 11. Studying the fire size distribution would be useful to help construct a wildfire spreading and distinguishing model. Finding a good model that could precisely describe the process of a forest fire is still an active and open topic

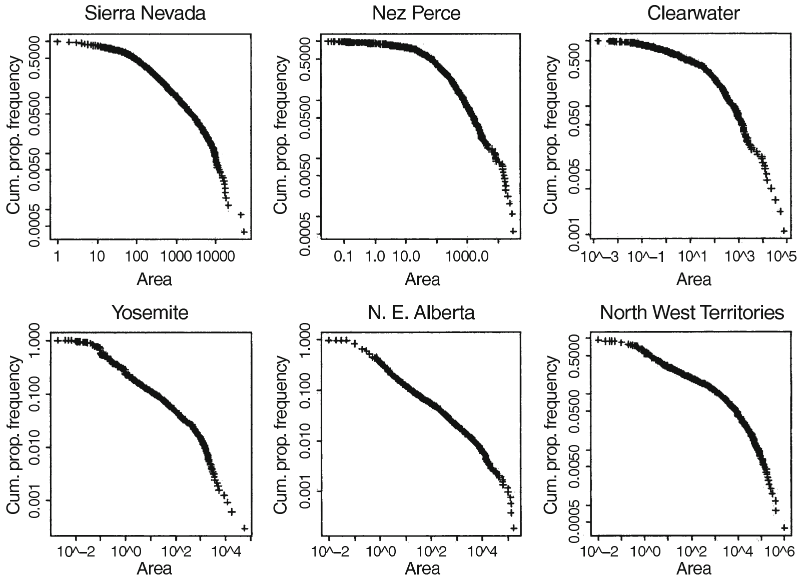

Fig. 11 A log-log plot of cdf for the size distribution of areas burned for the six dataset [47], measured in hectares

in ecological science. We only provide a sketch model to demonstrate a forest fire process and omit the details.

The start of a forest fire could just be a random lit up. The growth of a fire is a very complicated affair, depending on water, topology, temperature, humidity, plant types (i.e. fuel types), and many other factors. Percolation theory models forest fire spreads as a multiplicative stochastic process, which is similar to many biological entities that grow and die in a monotonic and stochastic fashion. That is, the size of burning area grows proportional to its current burning size as time spans. Therefore, the sizes of burnt area follows approximately a lognormal distribution for a fixed burning duration.

The causes of the forest fire extinguishment varies under different situations. For a small fire it could extinguish because of lack of fuel in surrounding areas (burned up or changes of materials). Another possible cause of extinguishment could come from intervention of human. However, people involved in fire fighting commonly believe that suppression cannot put out very large fires. The effect of suppression is more likely to reduce the spread (e.g., by putting separation lines around the back and sides of fires) rather than to actually extinguish the fires [48]. The only effective cause of extinguishment is precipitation. A few days of heavy rainfall could put off any size of forest fire. A weaker precipitation might only slow down the spread, which could be viewed as similar to that of human intervention. If we assume an equal chance that fires can be put off by natural causes, the time a forest fire lasts would follow an exponential distribution. Mixing the lognormal distribution of burned area size and the exponential time span will lead to a double Pareto distribution.

4 Conclusion and Future Directions

In this article we presented an overview of recent significant research results in the studies of power law that occur in complex networks and natural phenomena, explored a two-tailed power-law model called the double Pareto (lognormal) model and presented a number of real-world examples that can be explained using this model. The diversity of these examples shows the robustness of the double Pareto (lognormal) model. We expect that this model can be further applied to other areas that have yet to explore, including the knowledge networks. A number of complex networks have been discovered to obey the power law, and so we would like to know whether the linkage of human knowledge also exhibits the power law. If we could manage to divide the knowledge networks into subject domains, would the different bodies of domain knowledge share the structural similarity? If so, can we even control the knowledge accumulation process? These questions seem interesting, for they may help discover new knowledge. The world is expecting continuous excitement from new findings on power law research.

References

1. Pareto, V., Page, A.N.: Translation of Manuale di economia politica (”Manual of political economy”), A.M. Kelley, ISBN 9780678008812 (1971)
2. Reed, W.: The Pareto, Zipf and other power laws. Economics Letters, Vol. 74, No. 1, 15-19 (2001)
3. Reed, W., Hughes, B.D.: From gene families and genera to incomes and internet file sizes: Why power laws are so common nature. Physical Review E, Vol. 66, No. 6 (2002)
4. Mitzenmacher, M.: A history of and new directions for power law research. Invite talk at University at Buffalo (2008)
5. Reed, W., Jorgensen, M.: The double Pareto-lognormal distribution - A new parametric model for size distributions. Commun. in Statistics - Theory and Methods, Vol. 33, No. 8 (2004)
6. Ross, S.M.: Stochastic Processes, ISBN 9780471120629 (1983)
7. Brockett, R.W.: Talk at Kyoto University, Kyoto, Japan (2007)
8. Itō, K.: On stochastic differential equations. Memoirs, American Mathematical Society 4, 151 (1951)
9. Jiang, B., Brockett, R., Gong, W., Towsley D.: Stochastic differential equations for power law behaviors. Applied Probability Trust (2010)
10. Reed, W.: A parametric model for income and other size distributions and some extensions. International Journal of Statistics, Vol. LXIV, No. 1, 93-106 (2006)
11. Bachelier, L.: Théorie de la spculation, Annales Scientifiques de lÉcole Normale Suprieure, Vol. 3, No.17, 2186 (1900)
12. Gu, G., Chen, W., Zhou, W.: Empirical distribution of Chinese stock returns at different microscopic timescales. Physica A, 387, 495-502 (2008)
13. Onour, I.A.: Extreme risk and fat-tails distribution model: Empirical analysis. Journal of Money, Investment and Banking, Issue 13 (2010)

14. Klass, O.S., Biham, O., Levy, M., Malcai, O., Solomon, S.: The Forbes 400 and the Pareto wealth distribution. Economics Letters, Vol. 90, 290-295 (2006)
15. Jondeau, E., Rocklinger, M.: The tail behavior of stock returns: Emerging versus mature markets. Les Cahiers de Recherche 668 (1999)
16. Levy, M.: Market efficiency, the Pareto wealth distribution, and the Levy distribution of stock returns. Economy as an Evolving Complex System III, Oxford University Press (2006)
17. Park, K., Kim, G., Crovella, M.E.: On the relationship between file sizes, transport protocols, and self-similar network traffic. In proceedings of the 4th International Converence on Network Protocols, 171-180 (1996)
18. Downey, A.B.: Evidence for long-tailed distributions in the internet. In Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement (IMW '01), 229-241 (2001)
19. Downey, A.B.: Lognormal and Pareto distributions in the Internet. Comput. Commun. 28, 7, 790-801 (2005)
20. Arlitt, M.F., Williamson, C.L.: Web server workload characterization: the search for invariants. In Proceedings of the ACM SIGMETRICS'96, 126-137 (1996)
21. Downey, A.B.: The structural causes of file size distributions. In Proceedings of the 9th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 361-370 (2001)
22. Mitzenmacher, M.: Dynamic models for file sizes and double pareto distributions. Internet Mathematics, Vol. 1, No. 3, 305-333 (2004)
23. Crovella, M.E., Taqqu, M.S., Bestavros, A.: Heavy-tailed probability distribution in the World Wide Web. In A Practical Guide to Heavy Tails, Chapmen & Hall, 3-26 (1998)
24. Gong, W., Liu, Y., Misra, V., Towsley, D.: Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. Comput. Netw. 48, 3, 377-399 (2005)
25. Huberman, B., Adamic, L.: Growth dynamics of the World Wide Web. Nature, page 130-130 (1999)
26. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network Motifs: Simple building blocks of complex networks. Science, Vol. 298, No. 5594, 824-827 (2002)
27. Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, Vol. 311, No. 3-4, 590-614 (2002)
28. Ebel, H., Mielsch, L., Bornholdt, S.: Scale-free topology of e-mail networks. Phys. Rev. E Vol. 66, No. 3 (2002)
29. Dorogovtsev, S.N., Mendes, J.F.F.: Language as an evolving word web. Proc. R. Soc. Lond. B 22, Vol. 268, No. 1485, 2603-2606 (2001)
30. Ribeiro, B., Gauvin, W., Liu, B., Towsley, D.:On MySpace Account Spans and Double Pareto-Like Distribution of Friends. Second International Workshop on Network Science for Communication Networks (NetSciCom), pp1-6 (2010)
31. Hunt, F., Johnson, P.: On the Pareto distribution of SourceForge projects. In Proceedings of Open Source Software Development Workshop, 122-129 (2002)
32. Eeckhout, J.: Gibrat's law for (all) cities. The American Economic Review, Vol. 94, No. 5, 1429-1451 (2004)
33. Eeckhout, J.: Gibrat's law for (all) cities: reply. The American Economic Review, Vol. 99, No. 4, 1676-1683 (2009)
34. Nishiyama, Y., Osada, S., Morimune, K.: Estimation and testing for rank size rule regression under pareto distribution. In Proceedings of the International Environmental Modelling and Software Society (2004)
35. Chen, Y., Zhou, Y.: Multi-fractal measures of city-size distributions based on the three-parameter Zipf model. Chaos, Solitons & Fractals, Vol. 22, No. 4, 793-805 (2004)
36. Reed, W.: On the rank-size distribution for human settlements. Journal of Regional Science Vol. 42, No. 1, 117 (2002)
37. Giesen, K., Zimmermann, A., Suedekum, J.: The size distribution across all cities double Pareto lognormal strikes. Journal of Urban Economics, Vol. 68, 129-137 (2010)

38. Jiang, B., Jia, T.: Zipf's law for all the natural cities in the United States: A geospatial perspective. Preprint, http://arxiv.org/abs/1006.0814
39. Yule, G.U.: 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, Ser. B 213: 2187 (1925)
40. Allais, M.: Methods of appraising economic prospects of mining exploration over large territories. Management Science, Vol.3, 284-357 (1957)
41. Kaufman, G.M.: Statistical decision and related techniques in oil and gas exploration. Prentice Hall, Englewood Cliffs (1963)
42. Houghton, J.C.: Use of the truncated shifted Pareto distribution in assessing size distribution of oil and gas fields. Mathematical Geology, Vol. 20, No. 8 (1988)
43. Greenman, J.V., Fryer, M.J.: Hydrocarbon Field Size Distributions: A Case Study in Mixed Integer Nonlinear Programming. The Journal of the Operational Research Society Vol. 47, No. 12, 1433-1442 (1996)
44. Liu, X., Jin, Z., Chen, S., Liu, L.: Generalized Pareto distribution model and its application to hydrocarbon resource structure prediction of the Huanghua depression. Petroleum Science, Vol. 3, No. 2, 22-27 (2006)
45. Michel, B.: Oil production: A probabilistic model of the Hubbert curve. Applied Stochastic Models in Business and Industry, DOI: 10.1002/asmb.851 (2010)
46. Stauffer, D., Aharony, A.: Introduction to percolation theory, Second Edition. London: Taylor and Francis (1992)
47. Reed, W., McKelvey, K.S.: Power-law behaviour and parametric models for the size-distribution of forest fires. Ecological Modelling, Vol. 150, 239-254 (2002)
48. Wildfire Suppression. Wikipedia http://en.wikipedia.org/wiki/Wildfire_suppression
49. Schoenberg, F.P., Peng R., Woods J.: On the distribution of wildfire sizes. Environmetrics, Vol. 14, No. 6, 583-592 (2003)
50. Alvarado, E., Sandberg D., Pickford S.: Modeling large forest res as extreme events. Northwest Science Vol. 72, 6675 (1998)
51. Cumming, S.G.: A Parametric models of the fire-size distribution. Forest Research Vol. 31, No. 8, 12971303 (2001)
52. Holmes, T.P., Huggett, R.J., Westerling, A.L.: Statistical Analysis of Large Wildfires. Forestry Sciences, Vol. 79, II, 59-77 (2008)
53. Cui, W., Perera A.H.: What do we know about forest fire size distribution, and why is this knowledge useful for forest management? International Journal of Wildland Fire, Vol. 17, 234-244 (2008)