Dynamic Models for File Sizes and Double Pareto Distributions

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

Abstract

In this paper, we introduce and analyze a new generative user model to explain the behavior of file size distributions. Our Recursive Forest File model combines ideas from recent work by Downey with ideas from recent work on random graph models for the Web. Unlike similar previous work, our Recursive Forest File model allows new files to be created and old files to be deleted over time, and our analysis covers problematic issues such as correlation among file sizes. Moreover, our model allows natural variations where files that are copied or modified are more likely to be copied or modified subsequently.

Previous empirical work suggests that file sizes tend to have a lognormal body but a Pareto tail. The Recursive Forest File model explains this behavior, yielding a double Pareto distribution, which has a Pareto tail but close to a lognormal body. We believe the Recursive Forest model may be useful for describing other power law phenomena in computer systems as well as other fields.

1 Introduction

In this paper, we attempt to provide a simple generative user model that provides a good approximation for file size distributions. Accurate models for file size distributions are important for both our current understanding of and simulation of the Internet. For example, many studies have shown that traffic patterns in the Internet appear to have self-similarity (see, e.g.,[4, 5, 8, 9, 20]). This is naturally explained if the underlying distribution of file sizes obeys an appropriate power law. Understanding why a power law distribution for files might or might not arise naturally is therefore important. In a similar vein, tools for simulating Web servers such as SURGE [4] may require a suitable model for simulating file size distributions. Simple generative models may enhance such tools by providing accurate and flexible mechanisms for simulations.

We provide a model that combines ideas from recent work by Downey [12] and recent work on models for the Web graph [3, 13, 6, 17, 18, 19]. Underlying Downey's model is the following idea: one way that users create new files is by taking old files and performing modifications on them, including possibly editing, copying, translating, or filtering. The size of such a new file can be modeled by taking the size of an old file and multiplying it by a random variable. Downey suggests that his model yields a lognormal distribution for file sizes, which arguably counters other previous work that has suggested file size distributions have a lognormal body, but a heavy tail [4, 5].[1]

*Supported in part by an Alfred P. Sloan Research Fellowship and NSF grants CCR-9983832, CCR-0118701, and CCR-0121154.

[1]We believe there are minor problems with Downey's analysis. We therefore provide our own analysis of his model in Section 4.

Downey's model suffers from the weakness that all files derive from a single initial file. Files not derived from extant files cannot enter the file system, and old files are not deleted. We expand to a *dynamic* model; that is, we allow additions and deletions in a natural way. As a result we obtain a family of models, which we refer to generally as the Recursive Forest File model. What is most interesting is that our changes have a dramatic effect in the analysis. The resulting distribution of file sizes is a double Pareto distribution (or the slightly more general double Pareto-lognormal distribution). Double Pareto distributions have recently been suggested to describe income distributions and other power law phenomena [24, 25]. These distributions have Pareto tails for large files as well as Pareto tails for small files, in a manner we describe in Section 3. As we show, such distributions match well with previously used hybrid lognormal-Pareto distributions. We believe that such distributions may be useful for modeling other power law phenomena in computer systems, and we believe our generative model may prove useful for other applications.

We provide a detailed analysis of the Recursive Forest File model that is interesting in its own right. In particular, we find several connections to the theory of random graphs that we expect will provide a useful framework for future work. We also show how to cope with the effects of correlation that are implicit in a file system model where new files are derived from existing files, using a martingale analysis.

In other prior work, the highly optimized tolerance (HOT) model provides another generative model for file size distributions which uses an optimization framework [7, 27]. Downey suggests (and we concur) that applying this framework to Web file systems requires strong assumptions about how Web sites are designed, and does not explain why local file systems have similar file size distributions [12]. Downey's simpler framework appears more intuitively appealing, and therefore we have focused on improving it. We caution, however, that any simple user model is necessarily only approximate, and certainly various models may apply in different situations.

It is also worth noting that this potential confusion between whether file size distributions obey a power law or follow a lognormal distribution is not surprising. Similar discussions have arisen in many fields over several decades. Indeed, there is a rich history of models that generate power law and lognormal distributions, and many models that have been recently proposed to explain such distributions in computer systems have historical antecedents in other fields. Moreover, there are extremely close connections between generative models for power law distributions and lognormal distributions. Rather than dwell on these issues here, we refer the reader to a historically oriented survey [21] that can be seen as a companion to this paper.

To summarize, our primary contribution is a simple, general generative Recursive Forest File model for file size distributions and a corresponding analysis. We also demonstrate how the resulting double Pareto distribution fits well with previous studies that suggest file sizes have a lognormal body and a Pareto tail. While empirical validation of the underlying assumptions of our model would clearly strengthen this work, we suggest that such validation is extremely non-trivial and leave it as a subject for future work. We believe that our model is interesting in its own right and expect that it will find uses explaining other phenomena besides file size distribution.

2 Review of Definitions

2.1 Power Law Distributions

For our purposes, a non-negative random variable X is said to have a power law distribution if the *complementary cumulative distribution function* (ccdf), or $\Pr[X \geq x]$, satisfies

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

for constants $c > 0$ and $\alpha > 0$. Here $f(x) \sim g(x)$ denotes that the limit of the ratios goes to 1 as x grows large. 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 $k > 0$. Note the Pareto distribution requires $X \geq k$. If α falls in the range $0 < \alpha < 2$, then X has infinite variance. If $\alpha \leq 1$, then X also has an 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 the ccdf, asymptotically the behavior is a straight line. This is the basis for many tests for power law behavior. The same is true for the density function, which we find easier to work with mathematically. For example, for the Pareto distribution, the log of the density function is exactly linear:

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

2.2 Lognormal Distributions

A random variable X has a lognormal distribution if the random variable $Y = \ln X$ has a normal (i.e., Gaussian) distribution. The density function for a lognormal distribution satisfies

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

where μ is the mean and σ is the standard deviation of the associated normal distribution. We will say that X has parameters (μ, 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 $\mathrm{e}^{\mu - \sigma^2}$. Although the lognormal distribution, in contrast to the Pareto distribution, has finite moments, it is extremely similar in shape to power law distributions, in that a large portion of the body of the density function and the ccdf can appear linear [21, 22]. Specifically, for a lognormal distribution 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 is small for a large range of x values, and hence the logarithm of the density function will appear linear for a large range of values. (The same is also therefore true for the ccdf.)

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.

Lognormal distributions can be naturally generated by *multiplicative processes*. We start with a biological example. Suppose we start with an organism of size X_0. At each step j, the organism may grow or shrink by a certain percentage, according to a random variable F_j, so that

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

If the $F_k, 1 \leq k \leq j$, are all governed by independent lognormal distributions, then so is each X_j, inductively, since the product of lognormal random variables is again a lognormal random variable. More generally, lognormal distributions may be obtained even if the F_j are not themselves lognormal. Specifically, consider

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

Assuming the random variables $\ln F_k$ satisfy appropriate conditions, the Central Limit Theorem says that $\sum_{k=1}^{j} \ln F_k$ converges to a normal distribution, and hence for sufficiently large j, X_j is well approximated by a lognormal distribution. In particular, if the $\ln F_k$ are independent and identically distributed variables with finite mean and variance, then asymptotically X_j will approach a lognormal distribution. Lognormal distributions are natural for describing growth of organisms, growth in options prices, and any process where over a time step the underlying growth is a random factor independent of the current size [10, 21].

3 From Lognormal to Power Law: Double Pareto Distributions

Before presenting our model, we explain how a natural mixture of lognormal distributions yields a power law distribution. This result provides the foundation for much of our later analysis, and is interesting in its own right. We therefore present it first and show how it arises in the context of the model subsequently.

Suppose we have a system $X_t = F_t X_{t-1}$, where $X_0 = 1$ and F_t is a lognormal distribution with parameters (μ, σ^2). We think of the index t as referring to time. If we let the system run and stop it at some fixed time k, we obtain a random variable from the lognormal distribution with parameters $(k\mu, k\sigma^2)$. Suppose instead we run the process until some random time k. Then we obtain a random variable that comes from a mixture of lognormal distributions. Specifically consider the case where we have a geometric mixture of lognormal distributions: we stop the process at time k with probability $\gamma(1-\gamma)^k$, where γ is the parameter for the geometric distribution. Hence with probability $\gamma(1-\gamma)^k$ we obtain random number from lognormal distribution with parameters $(k\mu, k\sigma^2)$. We claim that the resulting distribution from this mixture will have a power law.

To see this, we present a result of Reed [24, 25] for the continuous analogue where the mixture of an exponentially distributed number of lognormal distributions is considered.[2] Suppose that we choose a random number X from a lognormal distribution with parameters $(k\mu, k\sigma^2)$, where k itself is a random variable with an exponential distribution with mean $1/\lambda$. The resulting density function is

$$f(x) = \int_{k=0}^{\infty} \lambda \mathrm{e}^{-\lambda k} \frac{1}{\sqrt{2\pi k}\sigma x} \mathrm{e}^{-(\ln x - k\mu)^2/2k\sigma^2} dk. \tag{2}$$

[2] Huberman and Adamic [15, 16] also examine this distribution and conclude that it has a power law distribution. Their earlier work, however, fails to note the behavior of the distribution goes through a phase shift, which Reed clarifies.

Using the substitution $k = u^2$ gives

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

An integral table gives us the identity

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

which allows us to solve for the resulting form. Note that in the exponent $2\sqrt{ab}$ of the identity we have $b = (\ln x)^2/2\sigma^2$. Because of this, there are two different behaviors, depending on whether $x \geq 1$ or $x \leq 1$. Let $C_1 = \lambda/\left(\sigma\left(\sqrt{(\mu/\sigma)^2+2\lambda}\right)\right)$ and let $C_2 = \left(\sqrt{(\mu/\sigma)^2+2\lambda}\right)/\sigma$. For $x \geq 1$, $f(x) = C_1 x^{-1+\mu/\sigma^2-C_2}$, so the result is a power law distribution. For $x \leq 1$, $f(x) = C_1 x^{-1+\mu/\sigma^2+C_2}$. In particular, a case we use later is when $\mu = 0$ and $\sigma = 1$. In this case, for $x \geq 1$, $f(x) = \left(\sqrt{\lambda/2}\right) x^{-1-\sqrt{2\lambda}}$, and for $x \leq 1$, $f(x) = \left(\sqrt{\lambda/2}\right) x^{-1+\sqrt{2\lambda}}$.

A key characteristic of the double Pareto distribution is that it has a power law at both tails. That is, if we look at the cumulative distribution function (cdf) on a log-log plot, it will also have a linear tail (for the small files). This provides a test for seeing whether a distribution has a double Pareto distribution; look at both the ccdf and the cdf on log-log plots for linear tails.

When we have the discrete geometric mixture instead of the continuous exponential mixture, the proper equation for the density function is

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

The summation is well approximated when $\ln x$ is very large or very small by the corresponding integral

$$f(x) \approx \int_{k=0}^{\infty} \frac{\gamma}{\sqrt{2\pi k}x\sigma} e^{k\ln(1-\gamma)-(\ln x-k\mu)^2/2k\sigma^2} dk. \tag{3}$$

Comparing (2) and (3), we get essentially the same tail behaviors from the geometric mixture as the exponential mixture (although we do not obtain such a nice closed form).

The double Pareto distribution falls nicely between the lognormal distribution and the Pareto distribution. Like the Pareto distribution, it is a power law distribution. But while the log-log plot of the density of the Pareto distribution is a single straight line, for the double Pareto distribution the log-log plot of the density consists of two straight line segments that meet at a transition point. This is similar to the lognormal distribution, which has a transition point around its median e^{μ} due to the quadratic term, as shown in equation (1). Hence an appropriate double Pareto distribution can closely match the body of a lognormal distribution and the tail of a Pareto distribution. For example, Figure 1 shows the complementary cumulative distribution function for a lognormal, double Pareto, and Pareto distribution. (These graphs have only been minimally tuned to give a reasonable pictorial match; they could be made to match more closely.) The lognormal and double Pareto distributions match quite well with a standard scale for probabilities, but on the log-log scale in Figure 2 one can see the difference in the tail behavior, where the double Pareto more closely matches the Pareto.

Reed also suggests a generalization of the above called a double Pareto-lognormal distribution with similar properties [25]. The double Pareto-lognormal distribution has more parameters, but might allow closer matches with empirical distributions.

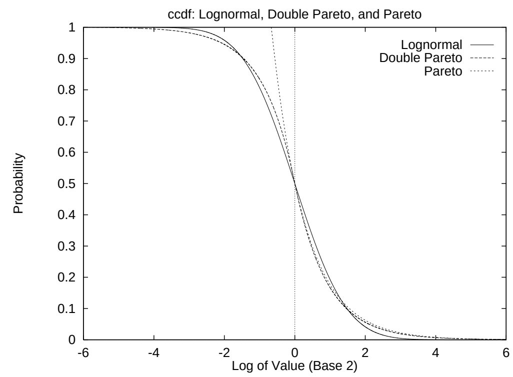

Figure 1: Shapes of lognormal, double Pareto, and Pareto distributions.

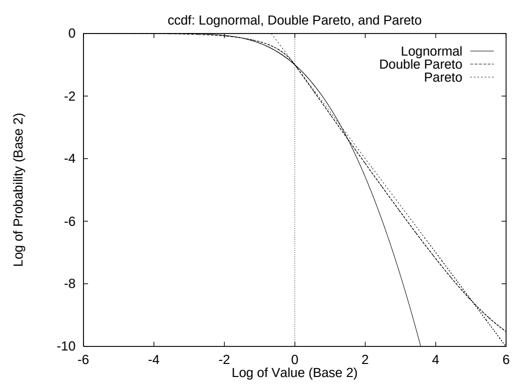

Figure 2: Shapes of lognormal, double Pareto, and Pareto distributions – log-log plot.

4 Downey's Multiplicative File Size Model

We now present Downey's model to provide appropriate background. In particular, we point out weaknesses in Downey's model that we ameliorate and introduce features of analysis that prove useful subsequently for our dynamic model.

Downey's model for file sizes is based on the following idea: users tend to create old files from new files, by copying, editing, or filtering in some way. Downey therefore suggests the following model. The system begins with a single file S_0, and the user repeatedly performs the following actions.

- Select a file S to modify uniformly at random. Let the size of S be s.
- Choose a multiplicative factor f from a given distribution $\mathcal{D}$.
- Create a new file S' with size fs.

The assumption behind this model is that creating a new file from a template file from processes such as copying, editing, translating, or filtering yields a file whose size differs from the template file by a factor that is independent of the size of the template. With filtering, for example, a fraction of the input may be recorded. For editing, if the amount of changes made is proportional to the size of the file (three edits per page), then this assumption appears reasonable. (Arguably, in many cases edits are additive rather than multiplicative; a constant number of changes are made. This can be modeled in a way reasonably consistent with the assumption by giving the distribution $\mathcal{D}$ a strong mode around 1.)

Looking at any individual file, there is a history of j steps that created all the previous versions, or predecessors, of that file. That is, a file S_j was created from a file S_{j-1} and so on back to the root S_0. Let X_0 represent the size of S_0 and let F_k represent the random multiplicative factor chosen from $\mathcal{D}$ in the creation of S_k. Then $\ln X_j = \ln X_0 + \sum_{k=1}^{j} \ln F_k$, and hence if $\mathcal{D}$ is lognormal each individual file has a lognormal distribution. Alternatively, even if $\mathcal{D}$ is not lognormal, X_j will be approximately lognormal if j is sufficiently large. Downey therefore suggests that the entire file size distribution resulting from this process is lognormal. This is not entirely accurate, as we explain below.

We note that preliminary empirical studies by Downey suggest that the right distribution for $\mathcal{D}$ is roughly lognormal, although it is more leptokurtotic; that is, there are more values near the mode, which is close to 1 since the most common operation on a file is a copy or a small change [12]. Downey finds that this has little effect on the overall results.

4.1 Random Tree Models

We provide an alternative view of the generative file process above by embedding it into a tree structure. Initially, we start with a root node, corresponding to the initial file. For convenience let us here take the size of the original file to be 1.

At each step, a random node of the current tree is chosen, and a new child of that node is created. Each node therefore corresponds to a new file that was created from the file corresponding to its parent, and the path from the root to the node corresponds to the file history. From here on, we use the terms node and file interchangeably. If we think of each edge as being labeled by a multiplicative factor, then by multiplying the numbers on the path from the root to a node we obtain its size (relative to the root node). Alternatively, we consider each edge as being labeled with the log of the multiplicative factor; then summing the weight along each edge gives the logarithm of the file size.

This tree model emphasizes that files have varying depths. While nodes at the same depth have the same size distribution, the size distribution will vary for nodes at different depths. Assuming that the distribution of the growth factor is lognormally distributed, a node at depth $k \geq 1$ has a lognormally distributed size with parameters $(k\mu, k\sigma^2)$. Hence, if the file sizes were independent, the distribution would be a mixture of lognormal distributions, derived from weighing the distribution for each depth with the proportion of nodes at each depth.

The tree developed under this model is well-studied in the combinatorial literature. It is known as a *uniform random recursive tree*, since the process looks the same to each node in the tree. Results regarding the height of tree, the distribution of depths of nodes, and so on are known. We provide a brief summary, based on [26].

An exact formula for the average number of nodes of depth k in a tree with n nodes is $\frac{1}{(n-1)!}\left[\begin{array}{c} n \\ k+1 \end{array}\right]$, where $\left[\begin{array}{c} n \\ k \end{array}\right]$ is the Stirling number of the first kind, or the number of ways to arrange n objects into k non-empty cycles. From this result one may prove that asymptotically the distribution of the depths of the nodes is sharply concentrated around $\ln n$; indeed, the distribution is asymptotically normal with mean $\ln n$ and variance $\ln n$. This explains why empirically Downey's model yields close to a lognormal distribution for file sizes; most nodes are at approximately the same depth and therefore have almost the same lognormal distribution, with additional symmetry to smooth out the effects of deep and shallow nodes.

This result suggests another possible problem with this model. It is not clear that in real systems the average depth of a file should be dependent on n, the number of files in the system. (Arguing that the maximum depth depends on n is more clear; perhaps some file, such as a script file, is used and modified occasionally as new files arise.)

One obvious way to generalize the file model is to use a different recursive tree model, such as plane-oriented recursive trees [11, 26]. In this model, the probability that a new node is the child of a node x is proportional to $c(x) + 1$, where $c(x)$ is the number of existing children of x. (Adding one avoids problems at the leaves and root.) This model is entirely similar to current models for the Web graph, which use this sort of preferential attachment in order to obtain power law distributions [3, 13, 17, 19]. Specifically, in this tree model, the fraction of nodes with k children is roughly proportional to $1/k^3$, a power law distribution. Such a model could apply if a user is more likely to modify versions of files that have already been modified several times. This may be quite possible – a useful shell script, for instance, may be more likely to be modified multiple times for various situations. In this case, in a tree with n nodes the depth of the nodes are sharply concentrated (again with an asymptotically normal distribution) around $\frac{1}{2} \ln n$.

One can generalize this model by having the probability that a new node is the child of a node x be proportional to $b \cdot c(x) + 1$ for some constant $b > 0$. A larger constant b strengthens the effect that nodes with children get more children; as b approaches 0, the model becomes more like the uniform random recursive tree. We revisit this possibility in the context of our Recursive Forest File model. We also note that further variations can be created by using different probabilities for the generation of children at each depth; however, such models seem excessively complex to be useful, and we avoid them here.

4.2 Correlations

The tree model also clarifies that file sizes are necessarily correlated: a child is clearly correlated to the size of its parent. Because of this, it is not clear what the resulting overall distribution of file sizes will be in this model. For example, one large multiplicative factor close to the root will affect several nodes, changing the overall distribution for an entire subtree. We emphasize that while the distribution of individual nodes is not affected by correlation, because of correlation it is difficult to make statements about the resulting joint distribution of the entire file system determined by the model.

We attempt to highlight the problem of correlation with a simple experiment. We simulated Downey's model, placing weights chosen from a normal distribution with mean 0 and variance 1 on each edge. Recall the logarithm of the ratio of the file size at a node to the initial file size is the sum of the weights on the edges along the path from the root; using this distribution, the average of these values, or the *average log ratio*, should be 0. Over 1,000 different runs generating 10,000 files, we found the average log ratio varied significantly, between -4.2 and 5.2. The absolute value of the average log ratio was greater than 2 over 150 times. These high average log ratios occur even though the sample variance is small; it is generally between 5 and 10. Moreover, similar experiments generating 100,000 and 1,000,000 files yield the same high average log ratios; over 1,000 trials, the range was roughly the same, and about 15% of the trials have average log ratio with absolute value at least 2. This effect is entirely due to the fact that a single large edge near the root can affect many nodes, moving the entire average log ratio. For a comparison, we performed 1,000 trials of taking the average of 10,000 independent normal random variables with mean 0 and an extremely large variance of 100. The distribution of the average is a random variable with mean 0 and a standard deviation of 0.1; over 1,000 trials, the averages ranged between -0.33 and 0.32.

4.3 Minimum File Sizes

A further potential argument against the multiplicative model is that it allows files to grow arbitrarily small as well as arbitrarily big. In practice, there is generally a natural lower bound to a file size (for instance, one byte). It is therefore worth asking how the multiplicative process behaves when there is a lower bound on the minimum size. That is, suppose that we have a (near) multiplicative process

$$X_j = \max\{F_j X_{j-1}, \epsilon\}$$

for some constant ϵ. In this case, the limiting distribution of X_j is not lognormal, but instead a power law [14]. This close connection between the lognormal and power law distributions is discussed more fully in [21], but it suggests that attempting to distinguish strictly between file size models that yield lognormal distributions from models that yield power law distributions may be a futile exercise. We avoid further focus on this issue in the analysis, however; generally we believe the effect on the model is relatively minor, although the issue arises in Section 6 when we examine real file size data.

5 The Recursive Forest File Model

5.1 Insertions

We now suggest a new class of dynamic models, based on similar dynamic models for modeling Web graphs. We call our models dynamic because they allow the introduction of new files into the system as well as the deletion of old files. We begin by handling insertion of new files only. We also temporarily ignore the problems of correlation until Section 5.3.

Our model begins with a collection of one or more files, whose sizes are drawn from a distribution $\mathcal{D}_1$. Repeatedly new files are generated as follows:

- With probability γ, add a new file with size chosen from a given distribution $\mathcal{D}_1$.

- With probability $1-\gamma$: select a file S (with size denoted by s) uniformly at random, choose a multiplicative factor f from a given distribution $\mathcal{D}_2$, and create a new file S' with size fs.

This generalizes the uniform random recursive tree model, so that the model produces a random recursive forest [2]. This explains why we refer to our class of models as Recursive Forest File models. Also, we have given each file an initial size. Implicitly, we may think of an edge to each root giving its initial size.

We first ask in this model how many nodes of each depth k there are when n files are in the system. Note that we could write exact recurrences for the expected value of these variables. Instead, we adopt a more intuitive limiting argument. Let $X_{t,j}$ be the number of nodes at depth j at time t. Since new nodes of depth 0, or roots, enter the system with probability γ, it is clear that $X_{t,0} \to \gamma t$, where $\to$ signifies convergence with probability 1 in the limit as t goes to infinity. Now for $X_{t,1}$ to increase, a new node that is the child of an existing node must enter; this happens with probability $1-\gamma$. Its parent must be a root; if $X_{t,0}$ is γt, this occurs with probability γ. Hence nodes of depth 1 arise at a rate of $\gamma(1-\gamma)$, so $X_{t,1} \to \gamma(1-\gamma)t$. Continuing inductively, we find (asymptotically) that $X_{t,j}$ approaches $\gamma(1-\gamma)^k t$; that is, node depths have a geometric distribution.

This model has several appealing implications. The average depth of a node is constant, which seems more reasonable than the logarithmic average depth in Downey's model. The maximum depth still depends on the number of nodes. The most likely depth of a file is 0, which means it is not derived from other files. The forests themselves demonstrate preferential attachment: a forest with several nodes is more likely to produce new children. Hence the forest sizes obey a power law, and in particular a constant fraction of the nodes are roots that have no children. These features appear realistic.

The geometric distribution of the depths is also appealing considering our results of Section 3. If $\mathcal{D}_1$ has all weight on size 1 and $\mathcal{D}_2$ is a lognormal distribution with parameters (μ, σ^2), then the results of Section 3 imply that the resulting distribution of file sizes is (approximately) double Pareto, since a node of depth k is lognormal with parameters $(k\mu, k\sigma^2)$.

It is clear that in this model the choice of distributions for $\mathcal{D}_1$ and $\mathcal{D}_2$ can have an important effect. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are both lognormal, the resulting distribution is what Reed calls a double lognormal-Pareto distribution, which has properties similar to the double Pareto distribution [25]. Similarly, if $\mathcal{D}_1$ is double Pareto or double lognormal-Pareto and $\mathcal{D}_2$ is lognormal, we still expect a distribution similar to the double Pareto (with Pareto tails and an approximately lognormal body). Overall, it is clear that the effect of $\mathcal{D}_1$ is small, as long as $\mathcal{D}_1$ is not a highly skewed or otherwise unusual distribution.

If $\mathcal{D}_2$ is not lognormal, then nodes with sufficiently large depth will appear approximately lognormal (by the Central Limit Theorem argument of Section 2.2), but shallow nodes will not. The resulting distribution may therefore depend on how deep the nodes are and how quickly the product of random variables chosen from $\mathcal{D}_2$ converges to a lognormal distribution; however, we emphasize that $\mathcal{D}_2$ does not strictly need to be lognormal for our results to hold. As mentioned previously, Downey's preliminary results suggest that $\mathcal{D}_2$ appear to have the property that it quickly converges to an almost lognormal distribution after a small number of multiplicative steps, which is favorable for our analysis. Further experimental analysis and understanding of both the initial file size distribution and the multiplicative growth distribution would be an excellent starting point for future work. Also, a stronger result demonstrating the robustness of our model to deviations in the distribution of $\mathcal{D}_2$ would be useful, but outside the scope of this work.

5.2 Deletions

We now consider the addition of deletions to the Recursive Forest File model. Suppose at each step that a new root enters with probability γ, a file chosen uniformly at random is deleted with probability η, and a new child node is introduced as before with probability $1-\gamma-\eta$. The introduction of deletions into the model has a surprisingly small overall effect on our previous analysis. We again give an intuitive argument for the limiting distribution, using a mean-field limit approach. Let $X_{t,j}$ be the number of nodes at depth j at time t, and $n(t)$ be the number of nodes at time t. Then

$$\frac{dX_{t,0}}{dt} = \gamma - \eta \frac{X_{t,0}}{n(t)},$$

and for $j \geq 1$

$$\frac{dX_{t,j}}{dt} = (1-\gamma-\eta)\frac{X_{t,j-1}}{n(t)} - \eta\frac{X_{t,j}}{n(t)}.$$

Now clearly $n(t) = (1-2\eta)t$ in the limit as t goes large, and inductively we can solve for the limiting values of $X_{t,j}$. The fraction of nodes at time t with depth j is then $X_{t,j}/n(t)$, and a simple induction yields $X_{t,j}/n(t) \rightarrow \gamma(1-\gamma-\eta)^j/(1-\eta)^{j+1}$. Hence the final distribution is again a geometric mixture of lognormal distributions, with the parameters slightly changed to account for deletions. As a result, the incorporation of deletions into the model does not disrupt the resulting power law distribution of file sizes.

More complex models can naturally be introduced in this framework. For example, in some situations it might be reasonable to suppose that the probability of a file being deleted is related to its depth; shallower (older) nodes may be more likely to disappear. This approach can be generalized to handle such situations, although it will affect the distribution of node depths, and again such models may be too complex to be useful.

5.3 Correlations

In our model, the file system is represented by a forest, instead of single tree. There are still correlations between file sizes; a file is still related to the size of its parent. However, the effect of these correlations is smaller, since the number of files descended from a single node is generally small, compared to the size of the file system.

To make this statement rigorous, we provide a martingale argument. For convenience, throughout we consider the case where there are no deletions; the argument generalizes naturally. Choose any fixed constant z. Let Z_j be the expected number of nodes with size greater than z once the first j nodes and their corresponding edges are revealed. (Recall that we may think of the root node of a tree as having an edge providing the size of the node.) Then $Z_0, Z_1, Z_2, \ldots, Z_n$ is a martingale, with Z_0 being the expected number of nodes with value at least z before any information is revealed, and Z_n being the actual number of nodes with value at least z. Let ν_j be the expected number of nodes in the subtree rooted at the jth node, where the nodes are numbered in the order of arrival (initial nodes may be ordered arbitrarily). Since the jth node and its corresponding edge only affect the nodes in this subtree, ν_j gives an upper bound on $|Z_j - Z_{j-1}|$.

Using Azuma's inequality (see, e.g., [23]), we have

$$\Pr[|Z_n - Z_0| \geq \epsilon n] \leq 2\mathrm{e}^{-\epsilon^2 n^2 / \left(2\sum_{j=1}^{n} \nu_j^2\right)}.$$

Suppose we can show that $\sum_{j=1}^{n} \nu_j^2$ is $O(n^{2-\epsilon})$ for some $\epsilon > 0$. Then we have that for any value z, the fraction of nodes with value greater than z is within ϵ of its expectation with very high probability; specifically, the probability is exponential in n^{ϵ}. This would demonstrate that the effect of correlation is very small when looking at the ccdf.

Hence we need an upper bound on $\sum_{j=1}^{n} \nu_j^2$. One approach is to simply use ν_1 as an upper bound on ν_j, so $\sum_{j=1}^{n} \nu_j^2 \leq n\nu_1^2$. To bound ν_1, let $\nu_{1,k}$ be the expected number of nodes in the tree of the initial root when there are k total nodes. If we begin with a sinle root node, then $\nu_{1,1} = 1$ and $\nu_{1,k} = \nu_{1,k-1}\left(1 + \frac{1-\gamma}{k-1}\right)$. Using $1 + x \leq \mathrm{e}^x$ we obtain

$$\begin{aligned}
\nu_{1,n} &= \prod_{j=1}^{n-1}\left(1 + \frac{1-\gamma}{j}\right) \\
&\leq \mathrm{e}^{(1-\gamma)\sum_{j=1}^{n-1} 1/j} \\
&= \mathrm{e}^{(1-\gamma)(\ln n + O(1))}.
\end{aligned}$$

This gives us that ν_1 is $O(n^{1-\gamma})$. This is only sufficient for Azuma's inequality if $\gamma > 1/2$, which is fairly limiting.

One way to cope with this problem is to use more initial nodes at the beginning of the process. For example, suppose that we begin with $\sqrt{n}$ root nodes in the file system originally. The expected size of the tree rooted at any of these nodes follows the same recurrence, but now the initial condition is $\nu_{1,\sqrt{n}} = 1$. Hence

$$\begin{aligned}
\nu_{1,n} &= \prod_{j=\sqrt{n}}^{n-1}\left(1 + \frac{1-\gamma}{j}\right) \\
&\leq \mathrm{e}^{(1-\gamma)\sum_{j=\sqrt{n}}^{n-1} 1/j} \\
&= \mathrm{e}^{(1-\gamma)(\ln n - \ln\sqrt{n} + O(1))} \\
&= \mathrm{e}^{(1-\gamma)(\ln n)/2 + O(1)}.
\end{aligned}$$

Now for any $\gamma > 0$, ν_1 is $O(n^{(1-\gamma)/2})$, and Azuma's inequality applies.

Using the above analysis, however, we can obtain a tighter bound on ν_j, even if we begin with a single root node. Let $\nu_{j,k}$ be the expected number of nodes in the subtree of the jth node when there are k total nodes. Then $\nu_{j,j} = 1$, and $\nu_{j,n} = \nu_j$ satisfies

$$\begin{aligned}
\nu_{j,n} &= \prod_{k=j}^{n-1}\left(1 + \frac{1-\gamma}{k}\right) \\
&\leq \mathrm{e}^{(1-\gamma)\sum_{k=j}^{n-1} 1/k} \\
&= \mathrm{e}^{(1-\gamma)\ln(n/j) + O(1)}.
\end{aligned}$$

In the above the $O(1)$ term can be taken to be independent of j for sufficiently large n. Hence ν_j^2 is $O((n/j)^{2(1-\gamma)})$. Algebra now yields that $\gamma < 1/2$, $\sum_{j=1}^{n} \nu_j^2$ is $O(n^{2(1-\gamma)})$; for $\gamma = 1/2$, $\sum_{j=1}^{n} \nu_j^2$ is $O(n \ln n)$; and for $\gamma > 1/2$, $\sum_{j=1}^{n} \nu_j^2$ is $O(n)$. In all cases Azuma's inequality gives strong probabilistic bounds.

We may conclude that the fraction of node values greater than any particular value is very close to its expectation with high probability. In broader terms, the effects of correlation are small for large enough systems with small enough trees. Note this argument demonstrates that correlation can be substantially reduced if we have more initial nodes to start the process.

Experiments using the average log ratio demonstrate that the unusual effects of correlation evident in Downey's original model do not occur in the Recursive Forest File model.

5.4 Variations of the Model

In our dynamic Recursive Forest File model, it is again possible to consider variations on how new nodes derive from old nodes, just as it was in the recursive tree model. The variety of possibilities is rather broad, so we content ourselves here to variations of the plane-oriented recursive forest. In this setting a new root is introduced at each step with probability γ; otherwise, a new child node is introduced, and the probability that the new node is the child of a node x is proportional to $b \cdot c(x) + 1$, where $b > 0$ is a constant and $c(x)$ is the number of children of x.

Again we may begin by looking at the number of children of each depth. As before let $X_{t,j}$ be the number of nodes of depth j at time t. Let $w(t)$, or the *weight* at time t, be the sum of $b \cdot c(x) + 1$ over all nodes. In the mean field limit, with one node added per unit time,

$$\frac{dX_{t,0}}{dt} = \gamma,$$

so that $X_{t,0} \to \gamma t$ asymptotically. The case for $j \geq 1$ simplifies once we use the fact that the total number of children of nodes of depth j equals the number of nodes of depth $j+1$. Hence the probability of creating a child at depth j is proportional to $bX_{t,j} + X_{t,j-1}$, since this is the sum of $b \cdot c(x) + 1$ over all nodes of depth $j - 1$. Hence

$$\frac{dX_{t,j}}{dt} = (1-\gamma)\frac{bX_{t,j} + X_{t,j-1}}{w(t)}.$$

In the limit for large t, $w(t)$ grows to $((1 + b)(1 - \gamma) + \gamma)t$, since every new root node contributes 1 to the weight and every other node contributes $1 + b$. Now if $X_{t,j}$ approaches $x_j t$ asymptotically, we find from the above that

$$x_j = (1-\gamma)\frac{bx_j + x_{j-1}}{((1+b)(1-\gamma)+\gamma)}.$$

Simplifying the above yields

$$x_j = (1-\gamma)x_{j-1},$$

so a simple induction again yields $X_{t,j} \to \gamma(1-\gamma)^j t$. Surprisingly, this is the same result as in the random recursive forest model, regardless of the value of b!

The value of b therefore does not affect the resulting geometric distribution of the depths of the nodes, and hence the double Pareto analysis still applies. We believe this demonstrates substantial robustness for this model in the face of changes.

The value of b does affect the model, however, in how the nodes are distributed among the trees in the forest. As a concrete example, comparing the uniform case ($b = 0$) with the plane-oriented recursive forest model ($b = 1$), we find for the larger b value there are a substantially greater number of trees consisting of just a single vertex and there is greater variance in the number of offspring from a root node. Hence the choice of b might be used to fine tune the underlying model to various file systems.

To see how b affects the distribution of the size of trees in the forest, we again describe an asymptotic mean field argument. Let $Y_{t,j}$ be the number of trees with j nodes at time t. Note that the total

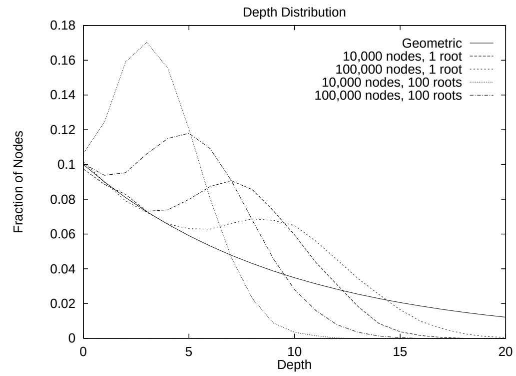

Figure 3: Depth distributions from simulations.

weight corresponding to a tree with j nodes is $(j-1)b+j$, since every node contributes $1+b$ to the weight except for the root. Hence we obtain the following equations:

$$\frac{dY_{t,1}}{dt} = \gamma - (1-\gamma)\frac{Y_{t,1}}{w(t)},$$

and for $j \geq 2$,

$$\frac{dY_{t,j}}{dt} = (1-\gamma)\frac{Y_{t,j-1}((j-2)b+j-1) - Y_{t,j}((j-1)b+j)}{w(t)}.$$

The asymptotic behavior of this system is easy to solve for, and the distribution of tree sizes in the forest follow a power law with the exponent in the power law depending on b [13, 18].

We also note that a similar derivation shows that the distribution of the depths of the nodes remains geometric under these variations when random deletions occur as in Section 5.2.

6 Simulations and Comparison with Data Sets

In this section, we examine simulations using the Random Recursive Forest model to compare it to the theory. In particular, we examine the issues of correlation and convergence to the limiting depth distribution. Overall, we find that simulations match the theory well.

We also examine data sets of file sizes to show that they reasonably match our model. Besides looking at the general shapes of the distributions, we attempt to see whether previously studied data sets have the double Pareto tail behavior predicted by our model. We emphasize that in this paper we explicitly do not provide great detail on empirical distributions derived from actual file systems, as we feel this would be redundant with the prior work on this subject [4, 5, 12]. Also, we do not strive for exact or fine-tuned matches because at this point we do not have a design tool that can match data sets to double Pareto distributions. Such a tool could be useful for future work.

6.1 Simulations of the Model

Consider first the problem of correlation. Recall we simulated Downey's model by placing weights chosen from a normal distribution with mean 0 and variance 1 on each edge. In 1,000 runs of generating

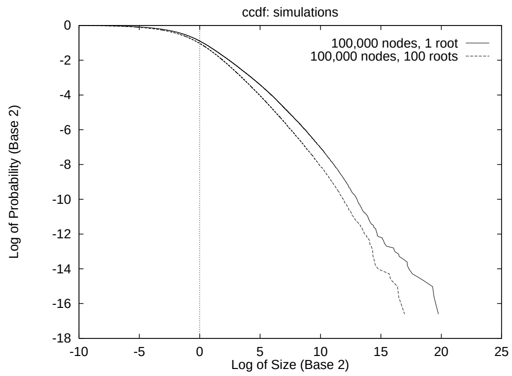

Figure 4: ccdfs for the simulations.

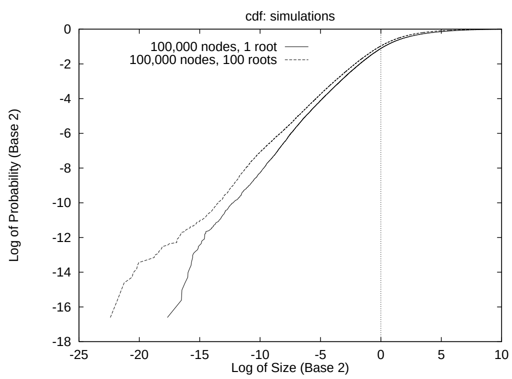

Figure 5: cdfs for the simulations.

10,000 files, the average log ratio varied between -4.5 and 4.0. We repeated the experiment using our dynamic model with $\gamma = 0.1$. Starting initially with 1 root node, the average log ratio varied between -2.26 and 2.52; starting with 10 root nodes, it varied between -0.95 and 1.22; and starting with 100 root nodes, it varied between -0.38 and 0.49. While it is clear that there are still correlations in the file sizes, they are dramatically reduced over Downey's model. Similarly, increasing the number of files leads to sharper concentration of the average log ratio, as our analysis would predict.

A second issue is convergence in the depth distribution. While asymptotically the depths will converge to a geometric distribution, it is not clear how many files are necessary for this to occur, especially if one starts with multiple roots at the beginning. Indeed, we find that the convergence in the depth distribution is slow, but it does not dramatically change the characteristics of the distribution shapes produced.

A representative example is instructive. We generated sets with 10,000 and 100,000 nodes, using $\gamma = 0.1$ and beginning with 1 and 100 initial roots. The results are presented in Figure 3. The resulting distribution does not match the theoretical geometric distribution; there is a bump in the distribution depending on the number of nodes generated and the number of initial roots. The more nodes generated, the closer to equilibrium.

Despite this deviation from the theory, examining plots on a log-log scale reveals that the cdf and the ccdf of the file sizes generated by the Recursive Forest File model still have linear tails, as

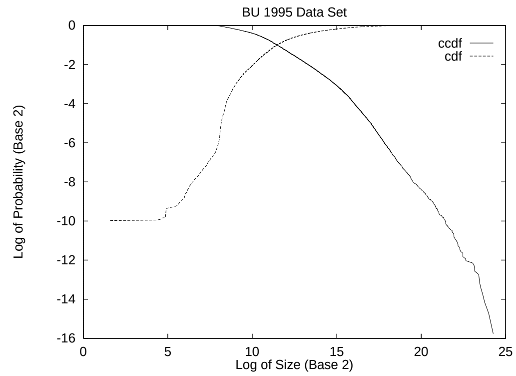

Figure 6: ccdf and cdf for the W95 data set.

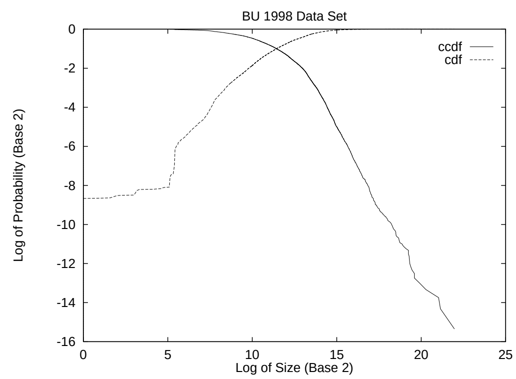

Figure 7: ccdf and cdf for the W98 data set.

shown in Figures 4 and 5. The deviation of the model from the theoretical double Pareto distribution appears to the body of the distribution a slight curve. Also, the linear tails break down somewhat at the extremes, because of the small number of samples and because the distribution has not reached the theoretical equilibrium. These features are also apparent in real data sets, as we see below, further lending credence to our model.

6.2 Comparing to Data Sets

We now examine whether our model reasonably matches file size distributions for known data sets. We begin with the unique file size data from file the W95 and W98 data sets of [5].[3] Log-log plots of the cdf and ccdf are given in Figures 6 and 7. The ccdfs in both cases closely match the shape of the double Pareto plot in Figure 1. This stands to reason, as the authors in [5] use a hybrid lognormal body and Pareto tail distribution to match these distributions. As we have already noted, the double Pareto distribution has a similar form. Deviations occur at the extremes of the tail, and the body does appear to have a slight curvature. As we have noted, however, similar behaviors also appear in simulations based on our model.

[3]We thank the authors for giving us access to the processed data.

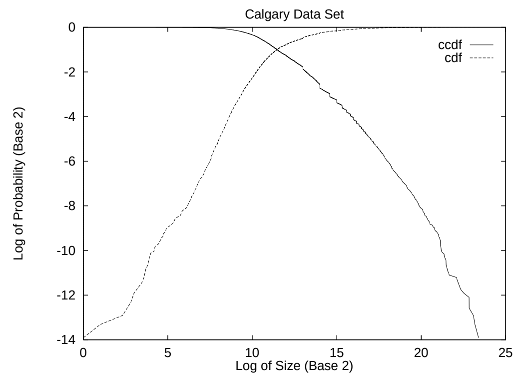

Figure 8: ccdf and cdf for the Calgary data set.

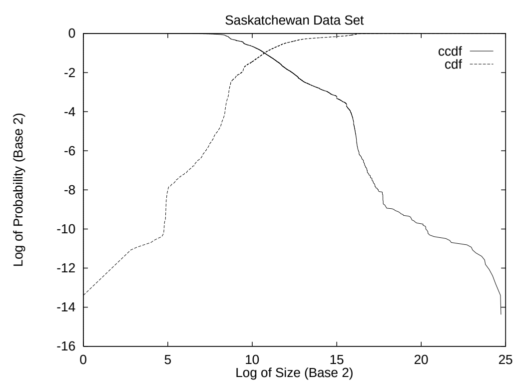

Figure 9: ccdf and cdf for the Saskatchewan data set.

We also examine the cdfs. Recall that a double Pareto distribution should also have a linear tail (for the small files) in a log-log plot. The W95 data set follows a roughly linear trend, and the W98 data set clearly does down to a certain size. The linear trends in the data sets break down for very small files (around 30-40 bytes). This might match the problem in the model discussed in Section 4.3 regarding a lower boundary for file sizes, as there seem to be a very large number of small files in these data sets.

For completeness, we also include two data sets (available on the Internet traffic archive) derived from traces of Arlitt and Williamson for the University of Calgary and the University of Saskatchewan. They are shown in Figures 8 and 9. We pulled out the sizes of all the unique files in the log. For the Calgary data set, the linear trend of the tails of the cdf and ccdf are quite strong. This is less true for the Saskatchewan data; however, this data has proven problematic in other studies. For example, Downey suggests using a two-mode lognormal to match this distribution [12]. On the whole, we feel that the Recursive Forest File model and the theoretical double Pareto distribution appear to fit empirical file size data well. A clear point for future work is to design tools to determine the best matching paramters for a double Pareto distribution given a data set.

7 Conclusions

We have provided and analyzed a new generative user model, the Recursive Forest File model, for file size distributions. Understanding the behavior of file size distributions is an important building block for understanding Internet behavior. Our model is extremely simple and well suited for simulation tools.

The underlying idea behind the model is to combine a multiplicative generating process with a dynamic insertion and deletion process reminiscent of recent Web graph models. A fundamental point in the analysis is to connect the file size model with corresponding random tree and forest models. We have shown the depth distributions are asymptotically geometrically distributed, and this in turn yields a double Pareto distribution for the file sizes.

From a practical standpoint, this model explains why file size distributions may appear to have a lognormal body and a Pareto tail. (In fairness, we point out that the shape of these distributions is still a subject of debate.) While previous work has suggested using specific hybrid distributions to model file sizes, our generative model appears sufficiently accurate, and has the advantage that it can be used to simulate dynamic systems where files may change over times. An open question for future work is how to design tools to fit properly parametrized double Pareto (or double Pareto-lognormal) distributions to empirically observed distributions.

From a theoretical standpoint, a Recursive Forest model provides a general mechanism for producing power law distributions that may apply to other natural systems. The robustness of the model to deletions and to changes in how elements produce offspring appear to be extremely appealing features. The flexibility and simplicity of the random graph framework should allow for further variations worthy of study.

8 Acknowledgments

The author would like to thank Laura Serban for help coding simulations; Steve Lumetta for help processing HTTP logs; Mark Crovella and Paul Barford for making their data available; and Mark Crovella, John Byers, Steve Lumetta, and Allen Downey for helpful comments.

References

[1] M. F. Arlitt and C. L. Williamson. Web server workload characterization: the search for invariants. In *Proceedings of the 1996 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, 1996.

[2] K. T. Balinska, L. V. Quintasem, and J. Szymański. Random recursive forests. *Random Structures and Algorithms*, 5:3-12, 1994.

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In **P**robabilistic Methods for Algorithmic Discrete Mathematics, ed. M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, pp. 249-314, Springer-Verlag, Berlin, 1998.

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

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

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. In *Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing*, 2001.

[22] 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.

[23] R. Motwani and P. Raghavan. **R**andomized Algorithms, Cambridge University Press, 1995.

[24] 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.

[25] 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.

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

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