New Models and Methods for File Size Distributions

Michael Mitzenmacher
Harvard University
Email: michaelm@eecs.harvard.edu

Brent Tworetzky
Harvard University
Email: tworetzky@post.harvard.edu

Abstract

We introduce several ideas into the debate on how to model file size distributions, including proposing the log-t distribution, suggesting a suitable metric for distribution fit, and studying mixture models based on file types.

1 Introduction

Understanding the distribution of file sizes being downloaded on the World Wide Web is a long-standing problem. In particular, the distribution of requested file sizes has been of primary study, both because requested file size sets are larger and more easily available than other size sets, and because many believe that requested file sizes heavily influence Web traffic [3, 12]. Furthermore, other Internet properties, such as Web file transfer times and related FTP distributions, may be closely related to this distribution [2]. Understanding requested file sizes, therefore, has potentially important consequences, especially for optimizing Web traffic processes, such as routing, caching, queuing, and scheduling [6, 7].

It is generally agreed that the body of the distribution of requested file sizes roughly follows a lognormal distribution. There has been more debate regarding the behavior of the right tail of the distribution, as file sizes get large [10]. Understanding the tail behavior is important if using a distribution for simulation purposes; underestimating or overestimating the largest files may have a large impact on the accuracy of the simulation. Similarly, if one plans to extrapolate the distribution to future situations, where there may be more files and in particular more files of larger size, it is important to model the tail accurately.

One possibility is to simply use a lognormal distribution, or a lognormal distribution with additional parameters [4, 5]. A more common approach is to use a lognormal distribution with a Pareto tail [1]. As another alternative, Mitzenmacher has recently suggested using the double Pareto distribution [11]. Discussions regarding the best fit have been hampered by the lack of a commonly accepted systematic framework for measuring the fit of a distribution to a data set.

In this paper, we make the following contributions to this line of research:

- We suggest a previously untested distribution, the log-t distribution, for file size distributions. The log-t distribution appears superior in several ways to previous distributions.
- We provide new and more robust metrics that allow for better fits and better comparisons of fits between empirical data and proposed distributions.
- We analyze the behavior of different file types, demonstrating that different file types exhibit different distribution shapes. Based on these findings, we consider mixtures of distributions based on file types using the above contributions.

All of our contributions therefore improve the ability to fit empirical data to an underlying distribution model.

The rest of the paper proceeds as follows. In section 2 we introduce the log-t distribution and compare it to previously suggested distributions for modeling file sizes. We introduce our

methodology for comparing distribution fits in Section 3, and present results in Section 4, focusing on comparing the results for log-t distributions to results for other distributions. In Section 5 we discuss the potential for mixture models based on allowing multiple distributions, one for each file type, and present results quantifying the effectiveness of mixture models. Discussion of relevant prior work will appear as necessary in the relevant sections.

2 Distributions

We review the distributions used in this work; for more background, see [10]. The lognormal and Pareto distributions are well known in the study of file size distributions. We also describe the double Pareto distribution suggested by Mitzenmacher [10], and the log-t distribution, which has not (as far as we know) previously been used in the study of file sizes.

2.1 The Pareto Distribution

A random variable X exhibits a power law distribution if, for some constants k, $\alpha > 0$, the complementary cumulative distribution function (ccdf) $\bar{F}(x)$ satisfies

$$\bar{F}(x) \sim kx^{-\alpha},$$

where $\sim$ represents that the ratio of the left and right hand sides approaches 1 as $x \to \infty$.

The Pareto distribution is the most commonly-used power law distribution, and, for some constants k, $\alpha > 0$, has a ccdf given by

$$\bar{F}(x) = \left(\frac{x}{k}\right)^{-\alpha}, \quad x \geq k$$

where k represents the minimum value possible. The corresponding probability density function (pdf) is

$$f(x) = \alpha k^{\alpha} x^{-(\alpha+1)}, \quad x \geq k$$

Taking the natural logarithm of the ccdf

$$\log \bar{F}(x) = \alpha \log k - \alpha \log x$$

which is linear in terms of $\log(x)$. Therefore in a log-log plot the Pareto ccdf should appear as a straight line; this is the standard eyeball test of whether data fits a Pareto distribution. Similarly, for a general power law distribution, once checks if the ccdf eventually becomes a straight line for large x.

If $\alpha \in (0, 1]$, the Pareto distribution has infinite mean, and if $\alpha \in (0, 2]$, the Pareto distribution has infinite variance.

2.2 The Lognormal Distribution

A random variable X is lognormally distributed if the random variable $Y = \ln(X)$ is normally distributed. The pdf of the lognormal distribution is given by

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(\frac{-(\log x - \mu)^2}{2\sigma^2}\right), \quad x > 0,$$

where μ and σ^2 are the mean and variance of the associated normal distribution. The lognormal distribution has a ccdf given by

$$\bar{F}(x) = 1 - \Phi\left(\frac{\log(x)}{\sigma}\right), \quad x > 0.$$

where $\Phi(x)$ is the cdf of the standard normal distribution. It is somewhat simpler to deduce the lognormal ccdf's behavior from the behavior of the lognormal pdf than to attach the ccdf directly. Taking the natural logarithm of the pdf yields

$$\log f(x) = -\log x - \log \sigma - \frac{\log(2\pi)}{2} - \frac{(\log x - \mu)^2}{2\sigma^2}. \quad (1)$$

Equation (1) is quadratic in terms of $\log(x)$. As the quadratic term, $(\log x - \mu)^2/2\sigma^2$, increases as the distance between $\log(x)$ and μ increases, the log-log plot of the lognormal pdf curves out from the mean to the tails. The same is true of the ccdf.

2.3 The Lognormal-Pareto Distribution

Crovella *et al.* noticed that file size distributions appear to have a lognormal body and a power law right tail, and they therefore proposed a hybrid lognormal-Pareto distribution. The hybrid distribution consists of a lognormal distribution, a Pareto distribution, and a split point. The distribution follows the ccdf of the lognormal distribution for values below the split point, and follows the distribution of the Pareto distribution after the split point. (Technically, there may be a discontinuity at the split point.)

2.4 The Double Pareto Distribution

Mitzenmacher [10] presents a file system model using a mixture of lognormal distributions that combine into a double Pareto distribution. A random variable X exhibits a double Pareto distribution if X is drawn from a lognormal distribution with parameters $(k\mu + \log\tau, k\sigma^2)$, where μ and σ^2 are standard lognormal parameters, k is a random variable drawn from an exponential distribution with mean $1/\lambda$, and τ is an additional location parameter. Letting $C_1 = \lambda \Big/ \left(\sigma\sqrt{(\mu/\sigma)^2 + 2\lambda}\right)$, $C_2 = \left(\sqrt{(\mu/\sigma)^2 + 2\lambda}\right) \Big/ \sigma$, and $C_3 = \mu/\sigma^2$, the pdf of the double Pareto is given by

$$f(x) = \begin{cases} C_1 (\frac{x}{\tau})^{-1+C_2+C_3} & x \le \tau \\ C_1 (\frac{x}{\tau})^{-1-C_2+C_3} & x \ge \tau \end{cases}$$

The double Pareto pdf can be viewed as two power laws joined at transition point τ.

The corresponding cdf is given by

$$F(x) = \begin{cases} \frac{C_1}{C_2+C_3} (\frac{x}{\tau})^{C_3+C_2} & x \le \tau \\ \frac{C_1}{C_2+C_3} + \frac{C_1}{C_2-C_3} \left(1 - (\frac{x}{\tau})^{C_3-C_2}\right) & x \ge \tau \end{cases}$$

The ccdf follows as $\bar{F}(x) = 1 - F(x)$.

Like the lognormal, the double Pareto has a transition point where its curve changes behavior. Like the Pareto, the double Pareto is a power law distribution and has a heavy tail. The double Pareto therefore exhibits the approximate body of a lognormal and the approximate right tail of a Pareto.

2.5 The log-t Distribution

The log-t distribution actually refers to a family of distributions that includes and generalizes the lognormal distribution. The normal distribution belongs to a family of distributions, called t distributions, which vary in body and tail behavior. The t distributions have mean and variance parameters, like the normal distribution, but also have an additional degree of freedom parameter, ν, which ranges between integral values from 1 to ∞. The t distribution has the following pdf:

$$f(x) = \int N(x; \mu, \sigma^2) \left[\frac{\chi_\nu^2}{\nu}\right]^{-1/2} d\chi_\nu^2$$

where $N(x; \mu, \sigma^2)$ represents the normal distribution pdf $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ and χ^2_ν represents a chi-square distribution with ν degrees of freedom[1]. Varying this third parameter modifies normal-generated values by a multiplicative factor, which varies by cumulative percentile. Therefore the ν parameter affects the body differently than the tails. Specifically, as ν is reduced from ∞ to 1, the tails of a t distribution are lengthened and the body is compressed inward [9], as shown in Figure 1.

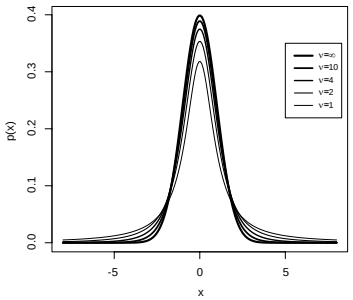

Figure 1: Examples of t distributions with various degrees of freedom.

The two endpoints of the ν range from ∞ to 1 define the normal distribution and the Cauchy distribution, respectively. We can therefore widen the tails of a normal distribution by using a t distribution with finite ν value.

From the t distribution we define the log-t distribution in the natural way; a random variable X has a log-t distribution if the random variable $Y = \ln(X)$ has a t distribution. Using the log-t distribution can overcome the lognormal distribution's shorter tail, producing a more nearly log-linear tail. The log-t distribution includes the lognormal distribution, as the case where $\nu = \infty$.

As far as we know the log-t distribution has not been proposed in the file size debate. The distribution has, however, been implemented successfully in economics as an approximation model with a fixed number of degrees of freedom [8]. The related t distribution has been more widely studied and used [9].

3 Evaluating Models

In this section, we develop our criteria for determining how well a distribution fits a given set of data and for comparing fits across distributions. While the visual representation of the log-log plot quickly relays the fit of model to data, we need metrics to quantify the goodness-of-fit of our models to the data and to compare models against each other. As we describe below, we believe previous work in this field has not presented a suitable discrepancy measure for model comparison. We therefore suggest what we believe is an appropriate test.

All of the distributions above have parameters that must be determined. We begin by pointing out that there are (at least) two natural ways to determine appropriate parameters before testing how well the data and model fit. We adopt the approach historically used for this problem, based on *discrepancy functions*. A discrepancy function takes the empirical data and the parametrized

[1] The pdf of the t distribution can also be characterized in terms of a Beta function:

$$f(x) = \frac{\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}}{\sqrt{\nu} B(\frac{1}{2}, \frac{1}{2}\nu)}.$$

This formulation allows for a non-integral number of degrees of freedom. Here we take the more traditional representation and use only integral degrees of freedom.

distribution as input and outputs a score measuring the goodness of the fit. We provide several examples of discrepancy functions below. Given a discrepancy function, we can simply determine the best parameters to minimize the score of the distribution against the empirical data. This approach is appropriate for achieving the best match with the given data, but it runs the risk of overfitting when trying to extrapolate future trends. Also, such parameter searches generally must be done exhaustively and/or approximately, and such searches can be computationally expensive, especially as the number of parameters grows. This highlights an advantage of using simple models with few parameters.

A second approach that we do not evaluate here is to find the parameters for the distribution that maximize the likelihood of the given empirical data appearing. For example, for a given set of data points $X_1, X_2 \ldots, X_n$, the maximum likelihood estimators for a normal distribution for the data are $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ and $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}}$. One could then measure the fit of the resulting parametrized distribution to the observed data with a discrepancy function. This approach is appealing if the goal is to extrapolate future trends in the data, such as predicting the tail behavior as the number of files grows larger in the future. While for some distributions there are simple formula for maximum-likelihood estimators of parameters, this is not the case for all of the distributions we consider; they could again be found by an expensive parameter search.

3.1 Discrepancy Functions

There are several standard discrepancy functions used in statistics; we briefly describe the most relevant ones below. Unfortunately, these tests appear unsuitable for the task of coping with heavy-tailed distributions. We require a test that provides a good match to the body but also presents a good match to the rare data points at the tail. Most standard tests emphasize the body, where most of the data points lie, excessively for our purposes.

The chi-square test is possibly the most commonly-used discrepancy test. This test separates the data value range into consecutive bins, and counts the number of elements in each bin (with each bin containing at least five data values [9]). The discrepancy is $\chi^2 = \sum_{i=1}^{N} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$, where N is the number of bins, Observed_i is the observed number of data values in the ith bin according to the data, and Expected_i is the expected number of data values in the ith bin according to the proposed distribution. The chi-square test can be used to gauge the statistical significance of the association between the model and the data set.

The chi-square test does a poor job of distinguishing at the right tail. The bins at the right tail tend to contain very few values, and large deviations in values at the right tail are not properly distinguished by the test. The test we suggest can be seen as a variation of the chi-square test that provides better comparisons for the right tail.

The Kolmogorov-Smirnov (K-S) test is a second commonly-used discrepancy function. The K-S test finds the largest difference between the proposed and empirical cumulative distribution functions: $D = \max_{1 \le i \le N} \left| F(x_i) - \frac{i}{N} \right|$, where F is the cdf of the model, N is the number of data points. and the x_i are the data points in sorted order. The Anderson-Darling (A^2) test is another well-known test similar to the K-S test.

Since the K-S and A^2 tests return the largest discrepancy between actual and estimated ccdfs at any one point, these discrepancy values do not reflect overall goodness-of-fit of an estimated model. One bad sample point, which is not unusual for heavy-tailed distribution, can have a large effect on the score.

A third natural discrepancy function (which we have not seen in the literature) is a sum-of-squares (SS) measure. This function is commonly expressed in the form $\sum_{i=1}^{N} (g(x_i) - g(\hat{x}_i))^2$, where $g(x)$ is a function chosen for the application, $\hat{x}_i = F^{-1}(i/(N+1))$ for the model cdf F, and x_i represents the ith data value sorted in ascending order. Simple datasets often compare

data to their estimated values using $g(x) = x$. For heavy-tailed distributions, using $g(x) = x$ results in the SS measure being dominated by the right tail of file sets. A natural choice would be $SS = \sum_{i=1}^{N} \left(\log(x_{(i)}) - \log(\hat{x}_{(i)})\right)^2$, as usually the curves are considered in log-scale. We have found SS to be a reasonable measure for fit, although it still tends to underweight the right tail.

Our suggested discrepancy function is based on the chi-square test, which we call a *chi-square fit*, or simply FIT. We define $FIT = \sum_{i=1}^{N} \frac{(x_i - \hat{x}_i)^2}{\hat{x}_i}$, where again the x_i are in sorted order. This metric is entirely similar to the formula for the chi-square formula, without the binning, and carries much of the same intuition as to why it is suitable for measuring fits. It is also quite similar to the standard SS measure, but scaled so as to avoid extreme dominance by terms in the right tail. In particular, while large data values can affect the score dramatically (as with the standard sum-of-square metric), this effect is lessened by dividing through by the value $\hat{x}_i$.

We have found by testing many graphs visually that the FIT measure serves as a suitable discrepancy function for our purposes. In particular, minimizing the FIT measure appears to yield better visual fits to data on log-log ccdf plots than the SS measure. For example, Figure 2 shows SS- and FIT-optimized fits to one of our file traces. (This is from the Boeing data set; it admittedly emphasizes our point strongly.) The FIT-optimized model has a much closer right tail fit than the SS-optimized model, which is overly focused on fitting the large number of points in the body as well as possible. We therefore propose FIT as a file size set discrepancy measure.

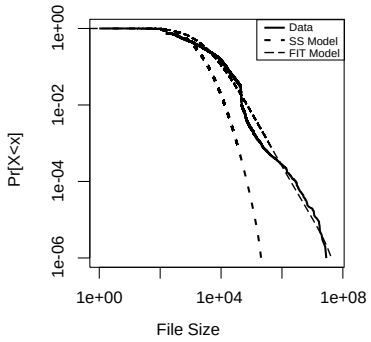

Figure 2: Behavior of SS against FIT on a sample file trace.

Finally, we note that as a sanity check, besides goodness-of-fit measures, we also compare the medians and the 99.99% midmeans of the data. The 99.99% midmean is the mean of the central 99.99% of the data; using the midmean instead of the mean mitigates the extreme variability of the mean when dealing with heavy-tailed distributions. Considering the median and midmean provides a rough guide as to the fit of the model on the body of the curve (which can of course also be tested simply by visual inspection).

4 Comparing Models

4.1 Empirical Data

For our test data, we use several publicly available traces; these traces, although generally somewhat out of date, have been used substantially in previous work, and appear to be among the most up to date available. Having publicly available newer traces would be helpful for continuing work in this area; in particular, it would be useful to have snapshots of the downloads from a WWW server months or years apart to examine how distributions change over time. Gathering new traces,

however, is outside the scope of this work.[2]

Name	Period Collected	Type	Valid Requests (1000s)
Boeing	1999	Web Proxy	1,010
Boston University 95	1994-5	Client	281
EPA	1995	Web Server	35
NASA	1995	Web Server	1,729
University of Calgary	1994-5	Web Server	567
University of California, Berkeley	1996	Web Proxy	1,975
University of Saskatchewan	1995	Web Server	2,164

Table 1: Traces and Logs

Our primary results appear in Table 2, where we find the best FIT score after optimizing the distribution parameters by exhaustive search. (The exhaustive search is done by successive refinement, focusing in on the most promising areas of the parameter space. Because the parameter space is multi-dimensional, it is possible that this approach misses optimal solutions; however, the space is generally well-behaved, and we expect our results are near optimal.) We emphasize that in examining these results we must keep in mind the complexity of the models; the lognormal distribution requires just two parameters, the log-t three, the double Pareto four, and the lognormal-Pareto hybrid five. Moreover, the log-t distribution generalizes the lognormal distribution; hence, at its worst the result should be at least as good as for the lognormal distribution. We also provide data on the 99.99% midmean and the median in Table 3.

The log-t distribution appears to be a significant enhancement, yielding substantially improved results over the lognormal distribution, and nearing the performance of the lognormal-Pareto hybrid. While the lognormal-Pareto distribution does obtain the best fit in the majority of the data sets, because the hybrid model requires more parameters, the performance of the log-t model is surprisingly good. Indeed, the performance of the lognormal model is itself very good, given the consideration that it only requires two parameters.

Trace	Lognormal	Log-t	Lognormal-Pareto	Double Pareto
Berkeley	31.5	8.8	5.4	12.9
Boeing	16.1	8.3	6.6	17.4
BU95	3.69	1.51	2.20	6.30
Calgary	8.78	5.88	7.31	16.5
EPA	0.58	0.37	0.43	0.43
NASA	19.6	19.6	17.6	148
Saskatchewan	29.1	10.6	8.13	24.6

Table 2: FIT scores of the best-fitting distribution of various types (in multiples of 10^8). The best score was found by an exhaustive search of the parameter space.

For lack of space, we include only one pictorial example in Figure 3, using the Saskatchewan data set.

[2]Web traces are available at the Internet Traffic Archive (<*http://ita.ee.lbl.gov/html/traces.html*>) and at Brian D. Davison's Web Caching site (<*http://www.web-caching.com/traces-logs.html*>). We received all traces unprocessed, and removed size zero file requests during formatting. For the Berkeley data set, we use the first of the trace set's four time-separated sections.

Trace	Lognormal		Log-t		Lognormal-Pareto		Double Pareto	
	Midmean	Median	Midmean	Median	Midmean	Median	Midmean	Median
Berkeley	1.18	0.70	1.10	0.94	0.96	0.94	1.02	1.36
Boeing	1.24	1.11	1.05	2.03	1.03	1.32	1.10	3.71
BU95	1.14	0.75	1.01	0.91	0.93	1.11	1.04	2.19
Calgary	1.09	0.75	1.06	1.01	0.98	1.07	1.09	2.14
EPA	1.12	0.48	1.06	0.72	0.86	0.89	0.93	1.28
NASA	0.97	1.06	0.97	1.06	0.98	1.07	1.05	2.64
Saskatchewan	1.17	0.85	1.05	1.15	0.85	1.35	1.05	1.71

Table 3: Ratios of the 99.99% midmean and the median between the results of the best-fitting models and actual data.

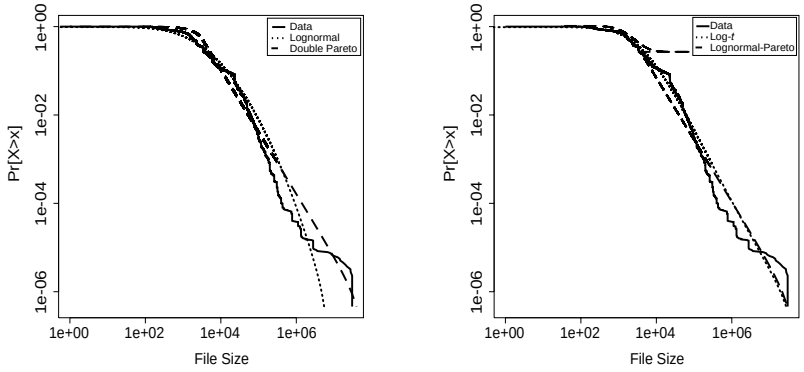

Figure 3: Best-fitting graphs under various distributions for the Saskatchewan data set.

5 Mixture Models

5.1 Different File Types have Different Distributions

In all previous work file size distributions have been fit by considering the entire data set of all files simultaneously. Different file types clearly have different behaviors: their storage formats, minimum or maximum size, or their likelihood of being modified may differ. It is therefore not unreasonable to posit that different file types may be governed by different distributions. In such a case, one would expect to obtain a better fit by classifying files according to type and using a different distribution for each type.

To test this hypothesis, we consider GIF and HTML file sets, as these file types routinely compose 80 percent of Web files, and also seem intuitively different from each other (being different sensory types of media). To contrast differences in right tail behavior, we use the best fitting log-t distribution as a distinguisher, in the following manner. We find the best-fitting log-t distributions (minimizing the FIT discrepancy measure), and consider the parameter ν. Recall that ν provides insight into the tail behavior of the distribution; when ν is infinite, the behavior is that of a lognormal, and smaller numbers correspond to heavier tails. Values of ν above 30, however, yield distributions quite close to lognormal distributions.

Table 4 shows the results[3]. In almost all file size sets, GIF file sets have larger ν values than HTML file sets; the one exception is the Boeing fiel trace. These results give evidence that GIF

[3]When we say the best fitting ν is infinity, we mean that we have tested larger and larger values of ν, beyond 10^7, and keep improving the FIT score. Further, it can be clarified by visual inspection of the appropriate plots, one can see that GIF files are best fitted by the lightest tail behavior, corresponding to $\nu = \infty$.

Trace	GIF ν	HTML ν
Berkeley	∞	13
Boeing	37	∞
BU95	∞	7
Calgary	∞	27
EPA	700	13
NASA	∞	36
Saskatchewan	∞	25

Table 4: GIF and HTML Shape Parameters

and HTML files exhibit different size distribution shapes. While there is clearly variance within the columns of Table 4, we still observe consistent behavior that would allow us to distinguish GIF and HTML file types.

We use the difference between file types to create mixture models. In what follows, we focus only on the log-t and lognormal-Pareto distributions, as these are the best performing and most interesting to compare. We break files up into four types: HTML, GIF, JPEG, and OTHER. If the percentage of JPEG files is less than 5%, files of this type are folded into the OTHER type. For each file type and distribution, we find the best-fitting parameters. Our log-t mixture model then consists of four (or three) distinct log-t distributions, one for each type, and the proportion of files of each type. Similarly we can construct a lognormal-Pareto mixture model.

Table 5 shows the potential benefit of mixture models. In most cases, there is a noticeable improvement, although it is worth noting that a mixture model can actually give a worse FIT score, as is the case for the Boeing data set. The improvement is somewhat larger for the lognormal-Pareto mixture models, but this is not surprising, as each lognormal-Pareto distribution has more parameters available to achieve a better fit. Again, we provide a pictorial example with the Saskatchewan data set in Figure 4.

Trace	Log-t	Log-t Mix	Lognormal-Pareto	Lognormal-Pareto Mix
Berkeley	8.8	3.7	5.4	2.2
Boeing	8.3	11.1	6.6	7.9
BU95	1.51	0.92	6.30	1.23
Calgary	5.88	3.86	7.31	1.97
EPA	0.37	0.25	0.43	0.07
NASA	19.6	28.7	17.6	10.3
Saskatchewan	10.6	9.2	8.13	3.2

Table 5: FIT scores of the best-fitting distribution of various types (in multiples of 10^8). The best score was found by an exhaustive search of the parameter space.

6 Conclusions

We have introduced the log-t distribution as a model for file size distributions. Our results suggest that it is nearly as good as the hybrid lognormal-Pareto distribution while using fewer parameters, and that it is significantly better than the lognormal distribution. We have based our results on the discrepancy function FIT, which we have suggested is a useful discrepancy function for capturing the desired fit between the data and specific distribution functions. We have also presented initial results on mixture models. Our results suggest that modeling different file types with different

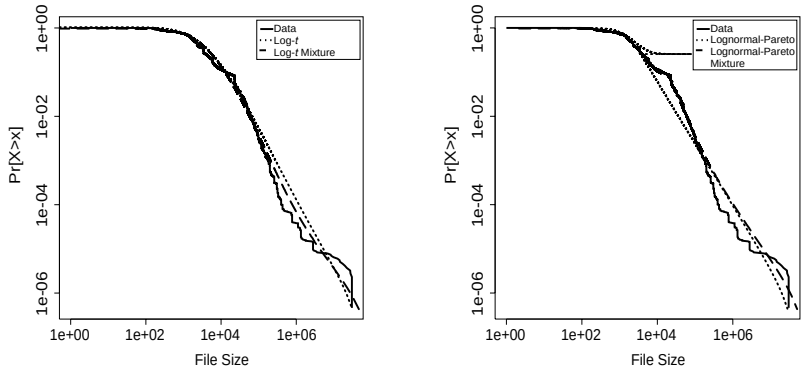

Figure 4: Comparing fits from the original model to the mixture model.

distributions can yield substantial improvements.

References

[1] M. Crovella and A. Bestavros. Self-similarity in World-Wide Web traffic: Evidence and Possible Causes. In *ACM SIGMETRICS'96*, pp. 160-169, May 1996.

[2] M. Crovella, M. Taqqu, and A. Bestavros. "Heavy-Tailed Probability Distributions in the World Wide Web". In *A Practical Guide to Heavy Tails*, ed. R. Adler, R. Feldman, M. Taqqu. Chapter 1, pp. 3-26, Chapman and Hall, 1998.

[3] M. Crovella and M. Taqqu. Estimating the Heavy Tail Index from Scaling Properties. In *Methodology and Computing in Applied Probability*, 1(1), pp. 55-79, 1999.

[4] A. Downey. The Structural Causes of File Size Distributions. In *Proceedings of IEEE MASCOTS '01*, August 2001.

[5] A. Downey. Evidence for Long-Tailed Distributions in the Internet. *ACM SIGCOMM Internet Measurement Workshop*, November 2001.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM 1999 Conference*, pp. 251-261, 1999.

[7] W. Gong, Y. Liu, V. Misra, and D. Towsley. On the Tails of Web File Size Distributions. In *Proceedings of 39th Allerton Conference on Communication, Control, and Computing*, October 2001.

[8] R. Hogg, and S. Klugman. On the estimation of long-tailed distributions with actuarial data. In *Journal of Econometrics*. 23, pp. 91-102, September 1983.

[9] N. Johnson, S. Kotz, and N. Balakrishnan. **Continuous Univariate Distributions**. Vol I, II(2). New York: John Wiley & Sons, Inc., 1994

[10] 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*, pp. 182-191, 2001. To appear in *Internet Mathematics*.

[11] M. Mitzenmacher. Dynamic Models for File Sizes and Double Pareto Distributions. 2002. To appear in *Internet Mathematics*.

[12] S. Manley and M. Seltzer. Web Facts and Fantasy. In *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems*, pp.125-133, 1997.