Extremal Dependence: Internet Traffic Applications

Felix Hernandez-Campos
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599-3175

J. S. Marron
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, New York 14850
and Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Sidney I. Resnick
School of Operations Research and Industrial Engineering
and Department of Statistical Science
Cornell University
Ithaca, New York 14850

Cheolwoo Park
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Kevin Jeffay
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599-3175

June 30, 2004

Abstract

For bivariate heavy tailed data, the extremes may carry distinctive dependence information not seen from moderate values. For example a large value in one component may help cause a large value in the other. This

is the idea behind the notion of *extremal dependence*. We discuss ways to detect and measure extremal dependence. We apply the techniques discussed to internet data and conclude that for files transferred, file size and throughput (the inferred rate at which the file is transferred) exhibit extremal independence.

1 Introduction

Internet file transfers are frequently subject to delays of various types. Intuitively, download times on the Internet depend on the sizes of the files being transferred, so the larger the file, the longer its takes to download it. In this paper it is seen that for very large transfers this notion is overly simplistic, using some novel ideas from statistics and probability. In particular, in the context of HTTP (HyperText Transfer Protocol, i.e. web browsing) responses, the joint behavior of large values of three variables, size of response, time duration of response, and throughput (rate = size / time) are considered. In Section 3 it is seen that for the largest responses, throughput tends to be more closely related to duration, and essentially independent of response size. This result is consistent with that of Maulik, Resnick and Rootzén (2002). See also Resnick (2003), (2004a).

The identification of the tendency of large values of object size and throughput to be independent has important ramifications for networking researchers. While the very large file transfers considered here are comparatively rare, measurements of Internet web traffic demonstrate that the transfer of these files comprises a significant fraction of all the bytes transferred on the Internet (e.g., the data in Hernandez-Campos, Jeffay and Smith (2003) showed that only 0.38% of the web files had a size of 100 KB or more, but these files represented 48.69% of the total bytes in the web traffic measured in a gigabit link)"). Hence understanding the dynamics of these transfers is critical to understanding the impact of diverse networking technologies such as routing, congestion control, and server design on end-user performance measures. For example, during file transfers, Internet servers typically maintain state for each transfer separately and the maintenance of this state is a significant factor affecting the scalability of servers (i.e., the ability of servers to service increasing numbers of connections or clients). Conventional wisdom leads one to eschew large file transfers (and hence the hosting of large files) as their (supposed) extreme transfer times may reduce the servers request throughput (request completion rate). The analysis presented here suggests these concerns may be unfounded. This agrees with the network-centric intuition that the rate of communication depends mostly on the state and speed of the network and not so much on the amount of data transferred. It is interesting, however, to confirm that the network does behave in the expected way as this implies that no extra bottlenecks are experienced by long file transfers. Viewed another way, the indications of extremal dependence of inverse throughput and duration seem to suggest that a significant percentage of the longest durations are due to low throughput rates rather than large file

transfers. This supports a hypothesis that persistent network pathologies (e.g., persistent network congestion or misconfigured routers/servers), or the heterogeneity of the network technologies used to connect end-systems to the Internet (from slow modems to gigabit Ethernet), distort the natural dependency between size and duration.

An almost opposite conclusion is presented in Zhang, Breslau, Paxson and Shenker (2002). This difference will be analyzed and reconciled with the above ideas, using a very careful analysis of different thresholding methods, in an upcoming paper. In this new paper it is seen that the differing conclusions are due to a surprisingly large systematic impact on the log-log correlation coefficient, of the choice of variable (i.e. size or duration) that is thresholded, (i.e., thresholding can create strong artificial correlations). A major advantage of the methods developed here is that they work in terms of threshold methods that are unaffected in this way.

We base the statistical analysis in this paper on the notion of "extremal dependence", a concept motivated by multivariate extreme value theory. The idea is that dependence between large values of a bivariate vector can be of different strength than dependence between moderate values. For joint bivariate probability distributions having heavy tailed marginal distributions (applicable to the data discussed above), the large values carry a useful type of dependence information. An example is "asymptotic independence", see Chapter 5 of Resnick (1987) for a formal introduction, and see Resnick (2002) for an overview of recent work in this area. Section 2 surveys this background. Asymptotic independence is the concept that for a bivariate random vector with heavy tailed distribution, the probability of both variables being large simultaneously is negligible in comparison to the probability of one of them being large. Extreme values of the two variables tend to occur separately, not simultaneously.

Figure 1 shows some examples that illustrate this concept for HTTP response size variables. The displays are "scatterplots", where pairs of data are plotted in the Cartesian plane as a graphical device for studying the structure of each joint bivariate distribution.

The data shown in Figure 1 are based on HTTP responses, gathered from the UNC main link during April of 2001. An HTTP "response" is defined here to be the set of packets associated with a single HTTP data transfer, and "duration" is taken as the time between the first and last packets. To allow study of diurnal effects, packets were gathered over 21 four hour blocks, over each of the 7 days of the week, and for "morning" (8:00AM-12:00AM), "afternoon" (1:00PM-5:00PM) and "evening" (7:30PM-11:30PM) periods on each day. The total number of HTTP responses over the four hour blocks ranged from ~ 1 million (weekend mornings) to ~ 7 million (weekday afternoons). Here we only consider "large responses", defined to mean those with more than 100 kilobytes (with numbers ranging from about 3,500 to more than 20,000). It is important to eliminate responses smaller than 100 KB from this analysis, because their durations are very sensitive to measurement inaccuracies and, more importantly, these responses are frequently transferred using only one or a few TCP (Transmission Control Protocol) windows using TCP's slow start mode.

This mode represents a regime that is quite different (and slower) from the one seen for larger transfers (that are allowed to fully utilize their bandwidth/delay product, depending on loss and maximum window sizes). Ignoring small file transfers enables useful application of the asymptotic lessons in our data analysis, which are not visible if the slower transfer regime of small files is included in the data. Furthermore, it also enables us to focus on the true effect of network characteristics on the size/duration dependency, since durations that are skewed by end-point artifacts are eliminated. The HTTP responses are analyzed separately for each of these 21 time blocks. To save space, only graphics for Wednesday afternoon are shown at most points in this paper. This time block was chosen as being frequently representative, and important differences for other time blocks are noted in the text. Similar analyses for the other time blocks can be seen on the web site
http://www.cs.unc.edu/Research/dirt/proj/marron/ExtremalDependence/.
Throughput vs. size analyses, similar to the left hand panel of Figure 1 appear on the page and in the file UNC2001RS1SPRScombine.pdf, and inverse throughput vs. duration also both appears on the page and is available in UNC2001RS1SPIRT.pdf. Similar scatterplots for the third combination of duration vs. size also appear and can be found in the file UNC2001RS1SPTScombine.pdf.

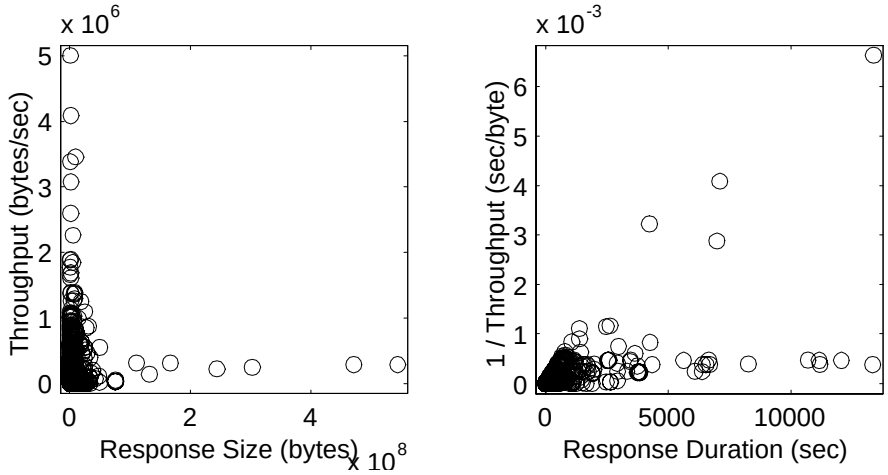

FIGURE 1: *Scatterplots of HTTP response throughput vs. size (left) and inverse throughput vs. duration (right). The "axis hugging" characteristic of extremal independence is visible on the left (suggesting throughput and size are independent for large values). Much different behavior exhibited in the right hand panel suggests strong extremal dependence for throughput and duration.*

The left panel of Figure 1 illustrates a situation where potentially the variable pairs are extremally independent. There are some HTTP responses (circles in this plot), with a very large size (horizontal coordinate), but not unusually large throughput (vertical coordinate). There are a number of responses with very large throughput (vertical coordinate), but not unusually large sizes (horizontal

coordinate). Thus the large values of throughput and size do *not* tend to occur together. The data tend to hug the axes and there is a very large empty box in the upper right corner of the plot. Extremal independence is expected here because larger files are expected to encounter more network delays, and thus are not likely to have large throughput. Note this information could be quite different than what would be contained in, say, a sample correlation coefficient which represents an average of very many values closer to the mean, which can be insensitive to a very few relatively large values. Another relatively weak point of the sample correlation is that it essentially measures how close the data lie to a slanted line, which is not a useful notion for heavy-tailed bivariate distributions.

The right hand panel of Figure 1 shows the opposite case. In particular, large values of inverse throughput (i.e. small values of throughput) and long durations do tend to happen simultaneously, as expected since inverse throughput = 1 / rate = time / size. Here the data do not hug the axes, and there is no large empty box in the upper right corner. Furthermore, the largest observation of each variable occurs simultaneously. Thus these variables have large values which seem quite dependent, which again is expected since lower throughput (typically caused by long delays) should be associated with longer durations.

While the above two scatterplots suggest expected bivariate dependence structure, a different visual impression appears in Figure 2. This time the comparison is duration vs. size of HTTP response. One might expect larger size to mean longer durations, resulting in a pattern similar to the right panel of Figure 1 (large values tend to occur simultaneously) but it is difficult to draw this conclusion from the scatterplot.

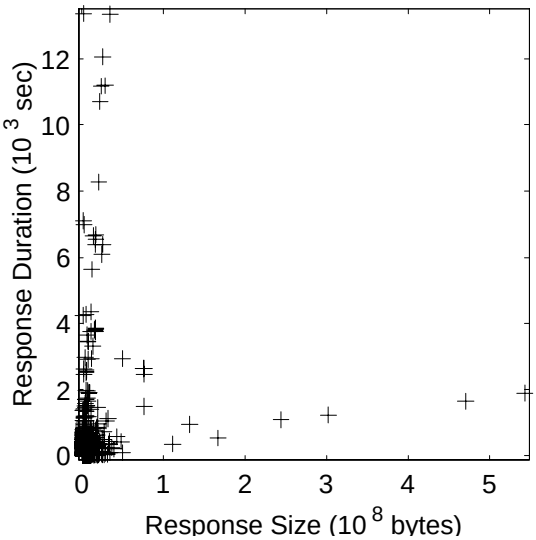

FIGURE 2: *Scatterplot of HTTP response duration vs. size. Note the surprising "axis hugging" characteristic of extremal independence.*

An important goal of this paper is to formalize, to the extent possible, a statistical basis for these visual and heuristic ideas. The data analysis is exploratory in nature, and is based on both a visual SiZer analysis and the quantitative Extremal Dependence Measure. Classical multivariate extreme value theory provides a context for development of procedures. However it has assumptions (e.g. multivariate regular variation) that are difficult or impossible to verify. The procedures developed here are motivated by extreme value theory, but are intended to be useful in a broader domain.

Our methods are based on two types of non-parametric transformations. The Inverse Complementary Rank Transform (Huang, 1992; de Haan and de Ronde, 1998; Einmahl, de Haan and Piterbarg, 2001), is first defined in Section 2, used for analysis in Section 3.1, with details on implementation given in Section 4.1. An advantage of this approach is that there is some sound probabilistic basis for the methodology, even though the asymptotic theory is not completely worked out. A possible limitation is that this theory is asymptotic in nature, and relies on the assumption of regular variation. The angular rank method, based on a different non-parametric transformation, is used for data analysis in Section 3.2, with implementation details described in Section 4.2. The advantage of this different transformation is a useful alternative approach to scaling issues. In Section 4.3 we begin an investigation of the mathematical properties of this method under classical heavy tailed assumptions but, presently, the approach is somewhat heuristic and has some unsolved normalization problems.

Although the two approaches are different, they point to the same major conclusions, which are consistent with those of Maulik, Resnick and Rootzén (2002), that large values of HTTP response size and throughput tend to be independent of each other.

While the analyses of this paper have been motivated by a particular problem in the area of Internet traffic, we believe the methods will also be useful for tackling other problems. For example in finance, an important issue is whether large changes in exchange rate returns for different currencies tend to occur together or separately; see Stărică (1999), Coles, Heffernan and Tawn (1999), Stărică (2000), Poon, Rockinger and Tawn (2003), Resnick (2004a). Environmental statistics, including the study of extrema of sea and wind conditions, is another area where such methods are likely to be useful, see Ledford and Tawn (1996, 1997) and de Haan and Ronde (1998).

In Section 2, we provide probability theory needed to develop our statistical methodology. The data analysis demonstrating the claims made here is described in Section 3. Some methodological details appear in Section 4. Concluding remarks are in Section 5.

2 Probability Background

In extreme value theory, the concept of asymptotic independence is designed to make the asymptotic, limiting distribution of extremes a product distribution. To be more concrete, suppose $\{(X_n, Y_n), n \geq 1\}$ are iid random vectors in a

domain of attraction of an extreme value distribution. The common distribution $F(x,y) = P[X_1 \le x, Y_1 \le y]$ possesses asymptotic independence if there exist normalizing constants $a_n > 0, b_n \in \mathbb{R}, c_n > 0, d_n \in \mathbb{R}$ such that as $n \to \infty$

$$P[\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \le x, \frac{\bigvee_{i=1}^n Y_i - d_n}{c_n} \le y] \to G_1(x)G_2(y),$$

where each $G_i(\cdot)$ is an extreme value distribution. For positive X's and Y's such as we consider for response sizes, durations and throughputs (rates), we focus on the heavy tailed case where both X and Y are in a domain of attraction of a heavy tailed Frechet extreme value distribution.

The background is best understood by first assuming that components of the vector can be normalized by just n. This is called the *standard case* so assume $X_1 \ge 0$, $Y_1 \ge 0$ and $x > 0$, $y > 0$ and

$$P[\bigvee_{i=1}^n \frac{X_i}{n} \le x, \bigvee_{i=1}^n \frac{Y_i}{n} \le y] \to G_*(x,y), \tag{1}$$

where $G_*(x,y)$ is a standard multivariate extreme value distribution. It is convenient to use vector notation $\boldsymbol{x} = (x,y)$. Relation (1) is equivalent to multivariate regular variation of the multivariate tail function $1 - F(\boldsymbol{x})$, that is, for $\boldsymbol{x} > \mathbf{0}$, $\boldsymbol{x} \ne \mathbf{0}$

$$\lim_{s\to\infty} \frac{1 - F(s\boldsymbol{x})}{1 - F((s,s))} = -\log G_*(\boldsymbol{x}). \tag{2}$$

A scaling argument shows that

$$G_*^t(t\boldsymbol{x}) = G_*(\boldsymbol{x}), \tag{3}$$

where G_*^t denotes G_* raised to the power t, and that there is a positive measure ν_* defined on subsets of the punctured first quadrant $\mathbb{E} := [\mathbf{0}, \boldsymbol{\infty}] \setminus \{\mathbf{0}\}$ such that

$$G_*(\boldsymbol{x}) = \exp\{-\nu_*([\mathbf{0}, \boldsymbol{x}]^c)\}.$$

The measure ν_* is fundamental and is called the *exponent measure*.

Asymptotic independence means G_* is the product distribution

$$G_*(x,y) = e^{-x^{-1}-y^{-1}},$$

and then ν_* concentrates on the axes through $\mathbf{0}$:

$$\nu_*(\boldsymbol{x}, \boldsymbol{\infty}) = 0$$

for $\boldsymbol{x} > \mathbf{0}$. This is equivalent to

$$\frac{P[X_1 > t, Y_1 > t]}{P[X_1 > t]} \to 0, \tag{4}$$

which is obtained from (1) by taking logarithms or directly from (2) when one assumes the limit distribution is a product. (See Resnick (1987), Chapter 5, for

details.) This gives rise to the interpretation, that when the two components have distributions which are asymptotically equivalent (hence the same scaling by n works for both components), given one component is large, it is unlikely the other component is large. Hence a scatterplot of data which is scaled the same in each component should have points hugging the axes.

The scaling property (3) translates to the measure ν_* and yields for Borel subsets of $\mathbb{E}$

$$\nu_*(tB) = t^{-1}\nu_*(B), \quad t > 0. \tag{5}$$

Make a polar coordinate transformation. Pick a norm $\|\cdot\|$ on $\mathbb{R}^2$ and from (5) we get

$$\begin{aligned}\nu_*\{\boldsymbol{x} \in \mathbb{E} : \|\boldsymbol{x}\| > t, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in \Lambda\} =& t^{-1}\nu_*\{\boldsymbol{x} \in \mathbb{E} : \|\boldsymbol{x}\| > 1, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in \Lambda\} \\ =& : t^{-1}S_*(\Lambda),\end{aligned} \tag{6}$$

where Λ is a subset of the unit sphere

$$\aleph := \{\boldsymbol{x} \in \mathbb{E} : \|\boldsymbol{x}\| = 1\}.$$

The measure S_* is called the *spectral measure*. It is customary, but not obligatory, to use the Euclidean L_2 norm and to parameterize the unit sphere by angles in $[0, \pi/2]$ and to think of S_* as a distribution on subsets of $[0, \pi/2]$. S_* is always a finite measure and at the expense of writing the limit in (6) with a constant as $ct^{-1}S_*(\Lambda)$, we can, and do, assume S_* is a probability measure.

Note that asymptotic independence means ν_* has empty interior and concentrates on the axes through $\mathbf{0}$. This translates into S_* concentrating on $\{0\}$ and $\{\pi/2\}$. So density estimates of an S_*, in case of asymptotic independence, should pile mass at the extremes of the interval $[0, \pi/2]$ and have little left for the interior.

Assuming the standard case (1) holds means we are assuming that the marginal distribution tails of both X_1 and Y_1 are asymptotically equivalent and regularly varying with index -1. This is at best a very crude approximation in practice. See Hernandez-Campos, Marron, Samorodnitsky and Smith (2002), Resnick (2003, 2004a). Thus it is important to consider the general case of heavy tailed components which is handled by functional transformation. We replace assumption (1) with

$$P[\bigvee_{i=1}^{n} \frac{X_i}{b_1(n)} \le x, \bigvee_{i=1}^{n} \frac{Y_i}{b_2(n)} \le y] \to G(x, y) = e^{-\nu([\mathbf{0}, \boldsymbol{x}]^c)}, \tag{7}$$

where

$$\nu([\mathbf{0}, \boldsymbol{x}]^c) = -\log G(\boldsymbol{x})$$

is the exponent measure corresponding to G. If $F_{(i)}(x)$, $i = 1, 2$ represent the marginal distributions of X_1 and Y_1 respectively, we can take the scaling functions $b_i(n)$ as the quantile functions

$$b_i(n) = \left(\frac{1}{1 - F_{(i)}}\right)^{\leftarrow}(n),$$

where for a non-decreasing function U, we denote by $U^{\leftarrow}$ the left continuous inverse. We then have that (7) holds iff

$$P[\bigvee_{i=1}^{n} \frac{b_1^{\leftarrow}(X_i)}{n} \leq x, \bigvee_{i=1}^{n} \frac{b_2^{\leftarrow}(Y_i)}{n} \leq y] \to G_*(x,y), \tag{8}$$

satisfies the standard case.

Note that for the general case, (X_1, Y_1) are asymptotically independent iff the transformed variables for the standard case $(b_1(X_1), b_2(Y_1))$ are asymptotically independent which translates to

$$\frac{P[b_1^{\leftarrow}(X_1) > t, b_2^{\leftarrow}(Y_1) > t]}{P[b_1^{\leftarrow}(X_1) > t]} \to 0. \tag{9}$$

This emphasizes the importance of having the components on the proper scale before inquiring about asymptotic independence.

2.1 Estimation in the Standard Case.

For what follows we use the notation for random element X and set A

$$\epsilon_X(A) = \begin{cases} 1, & \text{if } X \in A, \\ 0, & \text{if } X \notin A, \end{cases}$$

to denote the indicator that $X \in A$. This is a convenient device for denoting counting measures.

For the standard case, we can estimate ν_* with the empirical measure

$$\frac{1}{k} \sum_{i=1}^{n} \epsilon_{(\frac{k}{n}X_i, \frac{k}{n}Y_i)}(\cdot)$$

de Haan and Resnick (1993) where k is sometimes chosen based on a scaling plot (cf. Stărică (1999, 2000), Resnick (2004b)). The spectral measure S_* can then be estimated by

$$\hat{S}_* = \frac{\sum_{i=1}^{n} 1_{[r_i > n/k]} \epsilon_{\mathbf{\Theta}}(\cdot)}{\sum_{i=1}^{n} 1_{[r_i > n/k]}}, \tag{10}$$

where $(r_i, \mathbf{\Theta}_i)$ are the polar coordinates of (X_i, Y_i). This is the empirical distribution of $\mathbf{\Theta}$'s corresponding to big radius vectors which means we threshold the data according to radius vector and then look at the empirical distribution of resulting $\mathbf{\Theta}$'s. Asymptotic independence can be tested based on a statistic

$$\hat{v}_n = \int_0^{\pi/2} (\theta - \pi/4)^2 \hat{S}_*(d\theta), \tag{11}$$

which is extreme for the case of asymptotic independence, see Resnick (2004b).

Provided $n \to \infty$ and $k = k(n) \to \infty$ with $k/n \to 0$, all estimators are consistent. When $\hat{v}_n$ is based on polar coordinate angles whose empirical measure consistently estimates S_*, $\hat{v}_n$ consistently estimates

$$\int_0^{\pi/2} (\theta - \pi/4)^2 S_*(d\theta). \tag{12}$$

2.2 Estimation in the Non-Standard Case.

For the non-standard case, there are (at least) two ways to proceed.

1. One can, somewhat crudely, hope each marginal tail is asymptotically Pareto and use a power transformation to bring the Pareto parameter to 1. (See Maulik, Resnick and Rootzén (2002) for an example.) The sample

$$\{(X_i^{\hat{\alpha}_1}, Y_i^{\hat{\alpha}_2}), i = 1, \ldots, n\} \tag{13}$$

   where $\hat{\alpha}_i$ is the estimated α value for $1 - F_{(i)}$, $i = 1, 2$ should be approximately from the standard case. This has the obvious disadvantage of requiring estimation of the two α's which introduces much uncertainty. This uncertainty can be avoided (at a price) by using the next method.

2. A simple scaling argument, see de Haan and de Ronde (1998), Einmahl, de Haan and Piterbarg (2001) and Huang (1992), for the tail empirical measure shows that

$$\tilde{\nu}_* = \frac{1}{k} \sum_{i=1}^{n} \epsilon_{(\frac{k}{\overline{R}_i^{(X)}}, \frac{k}{\overline{R}_i^{(Y)}})} \tag{14}$$

   is a consistent estimator of ν_* from the standard case, where

$$\overline{R}_i^{(X)} = \sum_{l=1}^{n} 1_{[X_l \geq X_i]} = \#\{j : X_j \geq X_i\} \tag{15}$$

   is the complementary rank of X_i; that is, the number of observations at least as large as X_i. The ranks $\overline{R}_i^{(X)}$ are called "complementary ranks" because they relate to conventional ranks (the index of the ordered data starting at the minimum) in the same way as the cumulative distribution function relates to the complementary cumulative distribution function. From (14), we can derive an estimator of S_* and can compute $\hat{v}_n$ from (11) and the EDM (described next).

2.3 The Extremal Dependence Measure.

We can base a simple quantitative measure on $\hat{v}_n$ which reflects ideas about extremal dependence and axis hugging and which is more widely applicable than

the classical context in which $\hat{v}_n$ was introduced. We thus define the Extremal Dependence Measure (EDM), based on a set of angles $\theta_1, ..., \theta_k \in [0, \pi/2]$ by:

$$EDM = 1 - \left(\frac{4}{\pi}\right)^2 \frac{1}{k} \sum_{i=1}^{k} \left(\theta_i - \frac{\pi}{4}\right)^2. \tag{16}$$

EDM is well-defined for any set of angles and will also be used for the angular rank method used in Sections 3.2 and 4.2. In the context of Subsection 2.1 and equation (11) we have

$$EDM = 1 - \frac{\hat{v}_n}{(\pi/4)^2}.$$

Typically, the parameter k will be the number of multivariate exceedences above a threshold; for example, as in the definition of $\hat{v}_n$, k is the number of observations whose modulus r is greater than some threshold value. The basis of the EDM is the mean squared distance from the data angles to $\frac{\pi}{4}$, the center of the range of possible values, but it is linearly adjusted so that its values correspond to familiar values for the usual correlation. In particular, when the data points hug the axes (essentially extremal independence), most of the angles are near 0 or $\frac{\pi}{2}$, so $\frac{1}{k}\sum_{i=1}^{k}\left(\theta_i - \frac{\pi}{4}\right)^2 \approx \left(\frac{\pi}{4}\right)^2$, and $EDM \approx 0$ (thus working like the usual notion of correlation). When the data points lie near the 45 degree line, so $\frac{1}{k}\sum_{i=1}^{k}\left(\theta_i - \frac{\pi}{4}\right)^2 \approx 0$, and $EDM \approx 1$ (again working like conventional correlation because such data are nearly linearly related). Note that, assuming the axes are made properly "comparable" by one of the transformations outlined in Section 2.2, this also includes linear dependence with other slopes, since appropriate rescaling will change that line to the 45 degree line. Furthermore, these cases are the extremes, in the sense that $EDM \in [0, 1]$. One more landmark for interpretation of EDM comes from the fact that when the data have angles that are nearly uniformly distributed on $[0, \pi/2]$, a simple calculation shows that $\frac{1}{k}\sum_{i=1}^{k}\left(\theta_i - \frac{\pi}{4}\right)^2 \approx \frac{1}{3}\left(\frac{\pi}{4}\right)^2$, where the approximation holds in several possible senses, including the Weak and Strong Law of Large Numbers: $\frac{1}{k}\sum_{i=1}^{k}\left(\theta_i - \frac{\pi}{4}\right)^2 \underset{k\to\infty}{\to} \frac{1}{3}\left(\frac{\pi}{4}\right)^2$, in probability and almost surely. Similarly $EDM \approx \frac{2}{3}$, in the same senses. These approximations reflect that (11) consistently estimates (12) under a variety of hypotheses. We use the uniform distribution to interpret the EDM here, only because it is clearly in between the cases of extremal dependence and independence.

3 Data Analysis

A serious practical hurdle to applying the concept of extremal dependence is the requirement, seen from (7), (8) (9), that the two variables be on similar scales, or transformed to the standard case. Figures 1 and 2 create the suspicion that comparable scaling is not present since the variables are of far different orders of magnitude. Furthermore, there are many "large values" in some directions, and apparently fewer in others.

As outlined in (13), one can proceed by considering power transformations of the components of the data of the form

$$(x, y) \mapsto (x^{\alpha_x}, y^{\alpha_y}) . \tag{17}$$

These will drastically change visual impressions of the type illustrated in Figure 1, and thus are critical to a precise data based formalization of extremal dependence. Maulik, Resnick and Rootzén (2002) basically use this approach and estimate the marginal tail indices, and then renormalize with appropriate power transformations.

However, estimation of tail parameters is fraught with difficulties. For the problem of HTTP response behavior considered here (indeed for some of the same data sets), a strong case is made by Hernandez-Campos, Marron, Samorodnitsky and Smith (2002) that the classical tail index, assuming it exists, is not estimable from the data and that estimates of the tail index vary over a substantial range in various parts of the tail of the distribution. This casts doubt upon the viability of tail index power transformation as an analysis tool for HTTP responses. Tail index estimation might have more success in other contexts, perhaps including both Internet and non-Internet data. An advantage of the extremal dependence approaches considered here is that we can avoid the needed consideration of this issue on a case by case basis.

Two different approaches to this problem are presented in this paper. In Section 3.1, an extremal dependence data analysis of the HTTP response data is presented through the use of a nonparametric rank based transformation, the Inverse Complementary Rank Transformation (ICRT) which avoids the need for tail index normalization. Section 3.2 gives a parallel extremal dependence analysis of the HTTP response data, using the much different angular rank method. This uses different nonparametric rank ideas based on polar coordinates. Details of the implementation, illustrated using some toy examples, are developed in Section 4.2. These two methods are carefully compared, in the context of all of the HTTP response data sets, in Section 3.3.

The EDM, defined and discussed in Sections 2 (see (16)) and 4.1 is useful for both our analysis methods in that it is a simple quantitative measure, reflecting the above ideas about extremal dependence and axis hugging. An interesting open problem is the development of a null distribution, so that EDM could be used as the bases of formal statistical hypothesis tests about extremal dependence. Presumably, the null distribution would be an asymptotic normality statement where the asymptotic mean and variance can be estimated from the data, even in cases where the marginal distributions are unknown and not equal. An initial result has been proven under assumptions too restrictive for widespread statistical use, see Resnick (2004b). In this paper, we will only use EDM for comparison of "levels of extremal dependence" across cases.

3.1 ICRT Extremal Dependence Analysis

A fundamental part of our extremal dependence analysis is attempted reduction to the standard case using the ICRT, described in Section 2.2, equation 14 and

Section 4.1. The transformed data are represented in polar coordinates and then thresholded to the subset with largest radius components. Then the distribution of the angles corresponding to the exceedences is studied for indications of extremal independence. In Figures 3-5, the data are thresholded so that only the largest 2000 remain, but it is useful to look at a range of different thresholds. A number of other thresholds are considered, in the context of all of the HTTP response data sets (all 21 four hour time blocks), in Section 3.3. There it is seen that while the threshold can have a substantial effect, general lessons and comparisons are fairly insensitive to the precise choices. The threshold of 2000 is used here, because it provides interesting contrasts between the extremal dependence properties of the variables under consideration. Alternatively, the scaling technique of Stărică (1999) could be used to help decide on a threshold.

Figure 3 shows this analysis for the HTTP response throughput vs. size. The raw data are shown in the left scatterplot in Figure 1, which suggested extremal independence in this case, as intuitively expected.

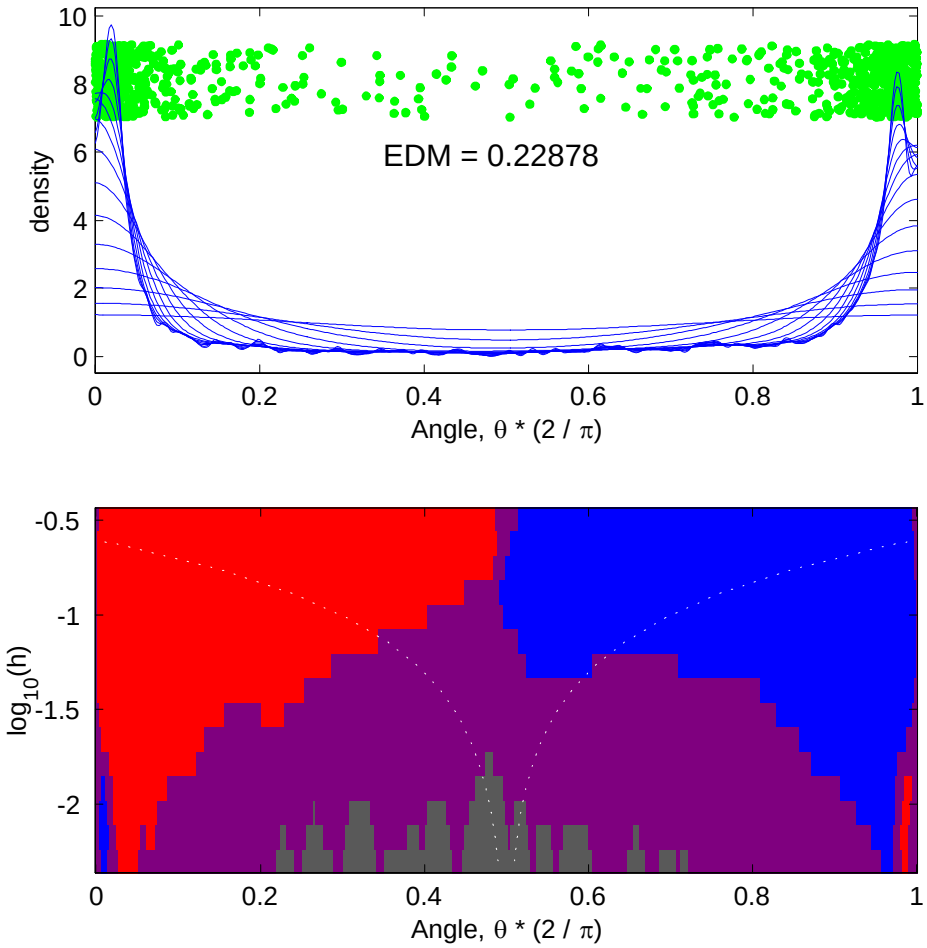

FIGURE 3: *ICRT extremal dependence analysis of throughput vs. size for HTTP response data. Note the apparent extremal independence.*

The top panel of Figure 3 shows the distribution of the angles for the top (with respect to the radius r) 2000 HTTP responses, in two ways. First there is a "jitter plot" (see Tukey and Tukey (1990) or pages 121-122 of Cleveland (1993)), shown as green dots, where the horizontal coordinate of each dot represents the angle, and a random vertical coordinate is used for visual separation of the dots. Second there is a family of smoothed histograms, shown in blue. The differing blue curves correspond to different histogram "binwidths", representing a range of different levels of smoothing, from grossly oversmoothed, to clearly undersmoothed. The version of "smoothed histogram" used here is the kernel density estimate, see Wand and Jones (1995) for a good introduction. Finally note that angles are shown on the scale of $\theta \times (2/\pi)$, because the range of $[0, 1]$ is more easily interpreted than the range of $[0, \pi/2] \approx [0, 1.57]$.

Both the green jitter plot, and the family of blue smooth histograms show

that the distribution of angles piles mass near the ends of the interval as is characteristic of extremal independence. While evidence of extremal independence is strong in this case, in other situations, it is not. The bottom panel of Figure 3 is a SiZer map (introduced by Chaudhuri and Marron (1998), but see also the web site:
http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer_Intro.html
for a useful introduction), which indicates which distributional features that are observable in the smooth histograms represent important underlying structure, and which are driven by spurious sampling variability. Each row of the SiZer map corresponds to one of the blue curves (representing a view of the data at a single "scale", i.e. smooth histogram with different binwidth). The horizontal axis is the same as in the top panel, thus representing the angle $\theta\frac{2}{\pi}$. Significance of structure is assessed via confidence intervals for the slope of the corresponding blue curve. When the confidence interval is completely above 0, the slope is significantly positive, and the color blue is used. In the bottom panel of Figure 3, the large amount of blue on the right shows that the large upward slope on that side is statistically significant. When the confidence interval is completely below 0, the slope is significantly negative, and the color red is used. In the bottom panel of Figure 3, the large amount of red on the left side reflects the general downward trend in those regions. When the confidence interval contains 0, it is unclear whether the slope is up or down, and the intermediate color of purple is used. The purple regions in Figure 3 appear where the blue curves are quite flat, and where they wiggle in apparently random ways (the purple in the SiZer map confirms that those small scale wiggles are indeed sampling fluctuations). The final SiZer color is gray, which is used in regions where the data are too sparse (this happens when the histogram binwidth is so small there is not enough data in each bin) for reliable statistical inference of the type done by SiZer.

While the visual analysis of Figure 3 presents a strong case for extremal independence, it has the possible drawback that one must understand the full distribution of angles (which is time consuming). For some purposes, for example analyzing many such data sets as done in Section 3.3, it is convenient to have a simple numerical measure. We suggest the EDM, defined in (16), for this purpose. The value $EDM = 0.23$ is shown in the top panel of Figure 3, and will be used for comparison below.

Figure 4 shows the same analysis for the HTTP response inverse throughput vs. duration. These raw data are shown in the right scatterplot in Figure 1, where it was seen that large values tended to appear simultaneously, as expected.

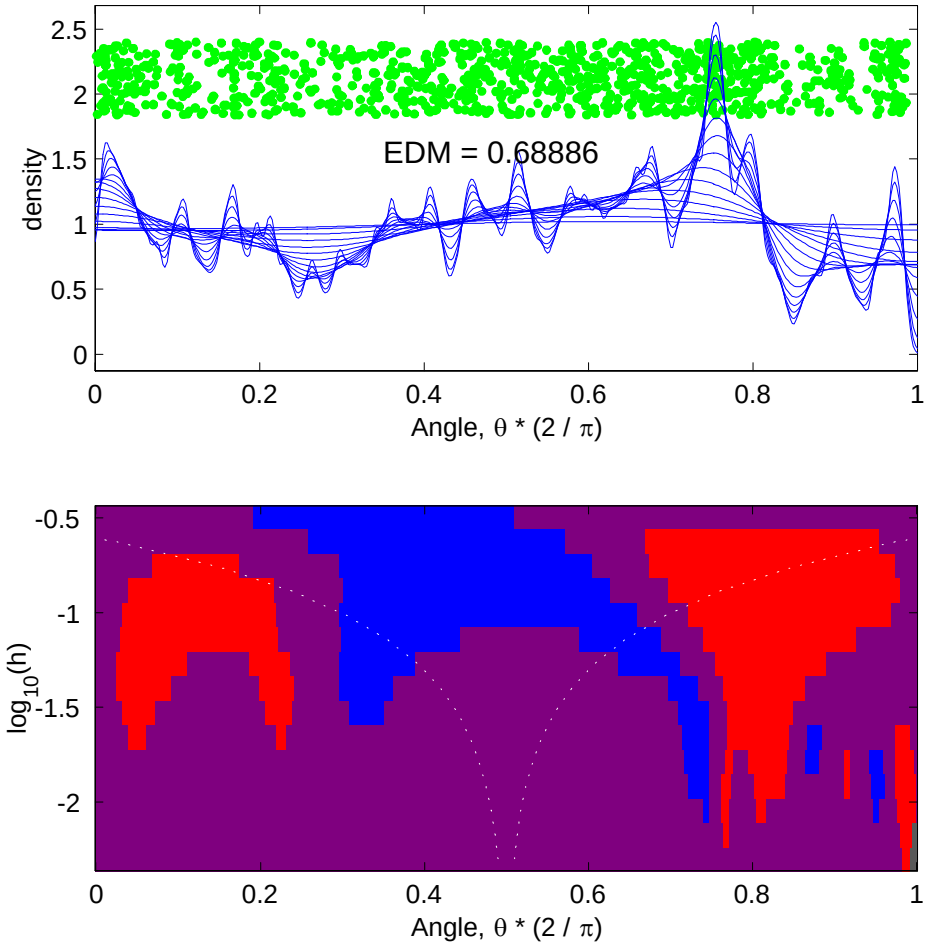

FIGURE 4: *ICRT extremal dependence analysis of inverse throughput vs. duration for HTTP response data. Note that large values occur simultaneously.*

The distribution of angles shown in Figure 4 is far different from that of Figure 3. The green dots of the jitter plot appear roughly homogeneous (corresponding to a uniform density). The family of blue smoothed histograms gives a more precise indication of the distributional shape showing a distinct peak near the angle $\theta \times (2/\pi) \approx 0.75$, and perhaps an important valley near the angle $\theta \times (2/\pi) \approx 0.3$. The smaller binwidths suggest a number of other possible peaks and valleys. Here the SiZer map in the bottom panel is very useful for understanding which features are statistically significant, in particular revealing that the indicated big peak and big valley are important, but most of the other wiggles cannot be distinguished from the background sampling variability, except for a few appearing in the lower right. The significant peak in the angle distribution corresponds to more than usual data points lying near a particular

line through the origin in the right panel in Figure 1.

The main conclusion from Figure 4 is that there is clearly no tendency for the angles to pile up at the ends of the interval, i.e. for the scatterplot data to hug the axes. This is reflected numerically by $EDM = 0.69$. This value is larger than $EDM = 2/3$, which corresponds to the uniform distribution, suggesting "less piling at the ends than for the uniform".

Figure 5 shows this same analysis for the HTTP response duration vs. size. Recall this was the surprising case, illustrated in Figure 2, where it was seen that the variables appear to exhibit some extremal independence, despite the dependence one might expect.

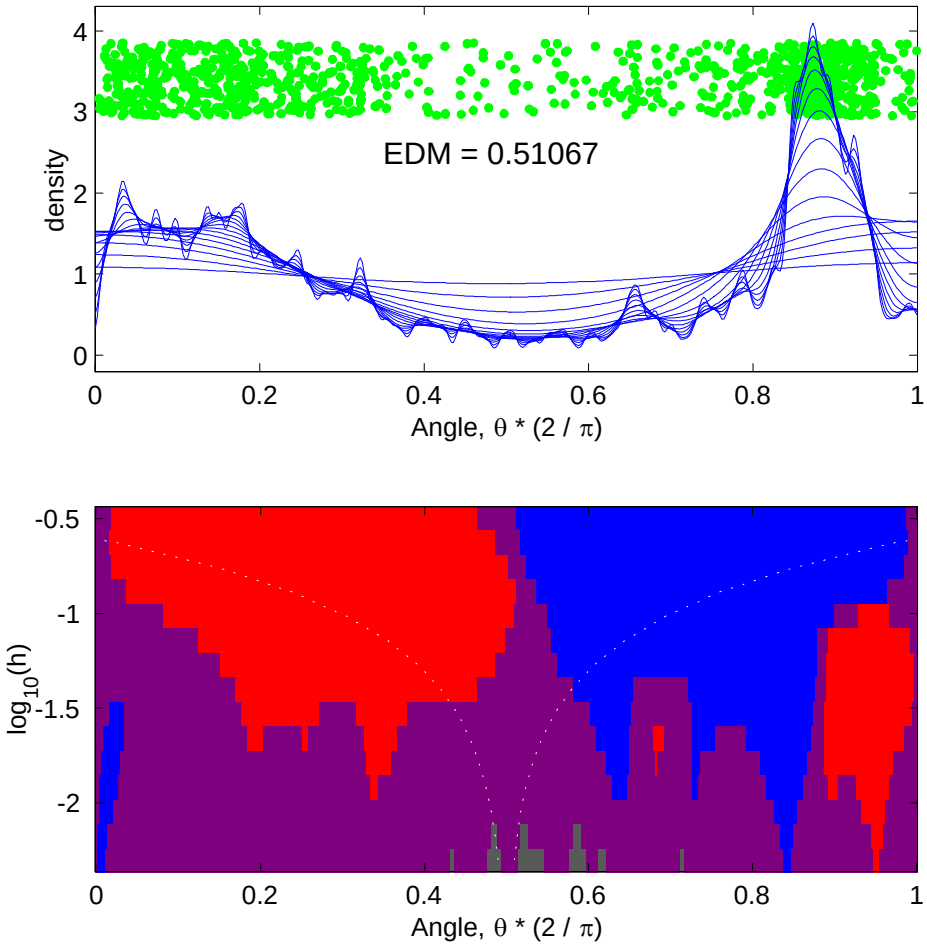

FIGURE 5: *ICRT extremal dependence analysis of duration vs. size for HTTP response data. Note the tendency towards extremal independence.*

The amount of extremal dependence exhibited in Figure 5 is between Figure 3 (extremal independent) and Figure 4 (extremal dependent), in terms of angles

piling up at the end points. While the blue curves appear to pile up somewhat near the endpoints, the piles are now more towards the middle of the distribution (see the SiZer blue at the left end, and red at the right end). This picture is less conclusive than seen in Figure 3, although it is tempting to identify a trend towards a "bathtub shape", i.e. a mild tendency towards extremal independence. A related impression comes from the value of $EDM = 0.51$, which is smaller than the uniform reference value of $EDM = 2/3$. This conclusion is confirmed from a different viewpoint in Section 3.2.

3.2 Angular Rank Extremal Dependence Analysis

The angular rank method also starts with transformation to make the axes (marginal distributions) comparable. Then, there is a transformation based on "angular equal spacing", and "end equal rescaling" ideas, described in Section 4.2. Next the analysis proceeds by thresholding the largest values, and using SiZer on the distribution of angles, as in the previous section. Again the largest 2000 values are kept for the analysis. These results are also sensitive to this choice of threshold, and as in Section 3.1, 2000 gave interesting contrasts between the variables. Again, more thresholds are considered, in the context of all of the HTTP response data sets, in Section 3.3.

Figure 6 shows the angular rank analysis for throughput vs. size. Recall the scatterplot on the left of Figure 1, and the ICRT analysis of Figure 3 both indicated extremal independence of these variables.

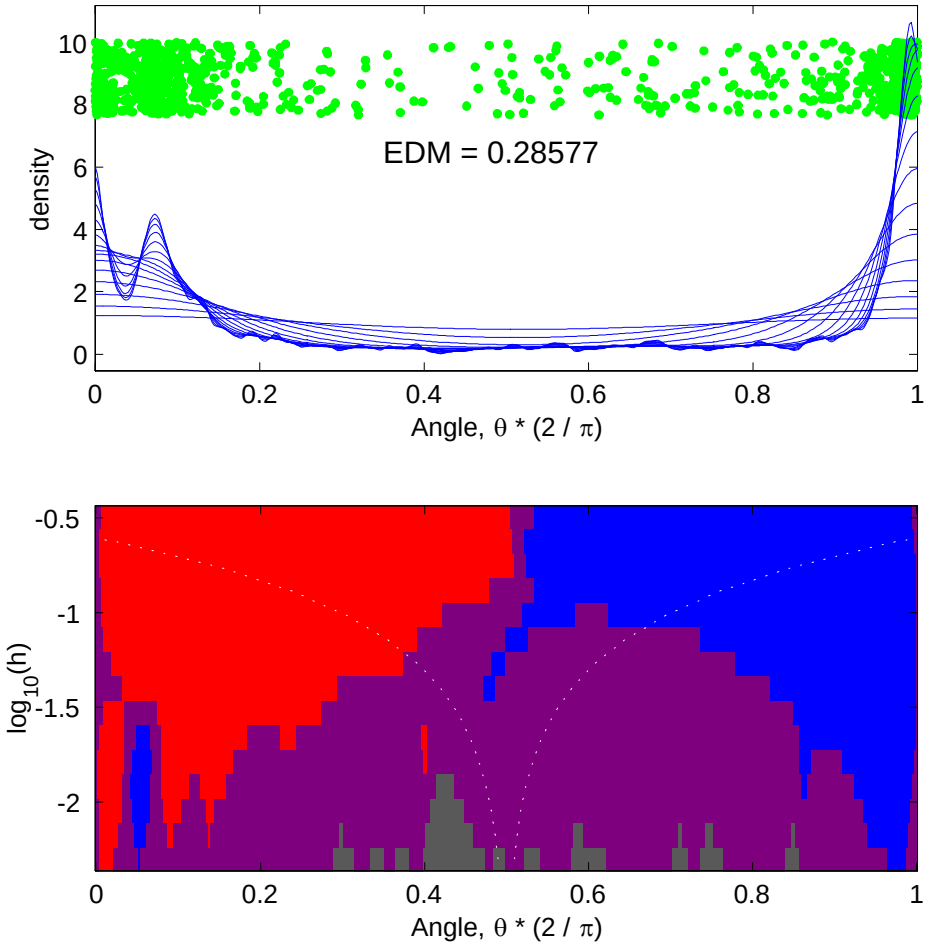

FIGURE 6: *Angular rank extremal dependence analysis of throughput vs. size for HTTP response data. Note the very strong extremal independence.*

Figure 6 provides confirmation of the earlier impression of extremal independence, indicating a bathtub shaped distribution of the angles. This is verified by the SiZer map, and by the relatively small value of $EDM = 0.29$.

Figure 7 shows the angular rank analysis for inverse throughput vs. duration. Recall the scatterplot on the right of Figure 1, and the ICRT analysis of Figure 4 both indicated extremal dependence of these variables, i.e. a tendency for variables to be large simultaneously.

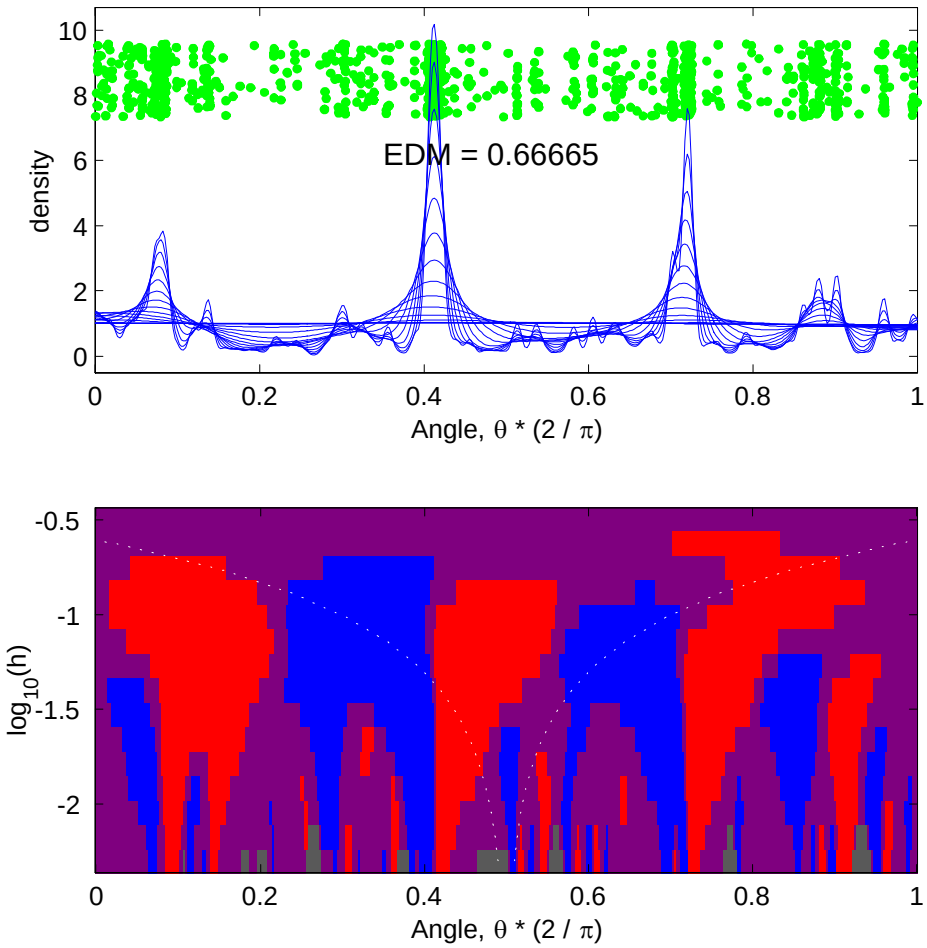

FIGURE 7: *Angular rank extremal dependence analysis of inverse throughput vs. duration for HTTP response data. Note the extremal dependence.*

Figure 7 shows an angular distribution with some large and statistically significant peaks, near $\theta \times (\pi/2) = 0.42$ and 0.73. These indicate that the largest values of inverse throughput tend to be one out of just a few multiples of duration time. This time the $EDM = 0.67$, which is close to the value of $2/3$ that would appear for the uniform. This is caused by the chance location of the peaks, and show that EDM alone has limited ability to distinguish between different forms of extremal dependence.

Note, Figure 7 indicates density plots which seem to indicate that S_* is discrete with several atoms. In such cases, a scatter plot of thresholded vectors, culled from the sample by dint of having large radius vectors, should show exceedences tending to follow rays at angles corresponding to the peaks in Figure 7. Such an S_* could be realized as a mixture model. The extreme value background is given on page 276 of Resnick (1987) and models exhibiting this type

of behavior can be realized by taking p (=number of atoms of S_*) max-linear combinations of the form

$$\bigvee_{j=1}^{p} a_j(p_j X_j \vee q_j X_j)$$

where $X_1, \ldots, X_p$ are iid Frechet random variables and for $1 \leq j \leq p$ we have $p_j + q_j = 1$.

Figure 8 shows the angular rank analysis for duration vs. size. Recall from the ICRT analysis that this case was between the two previous cases in terms of extent of extremal dependence.

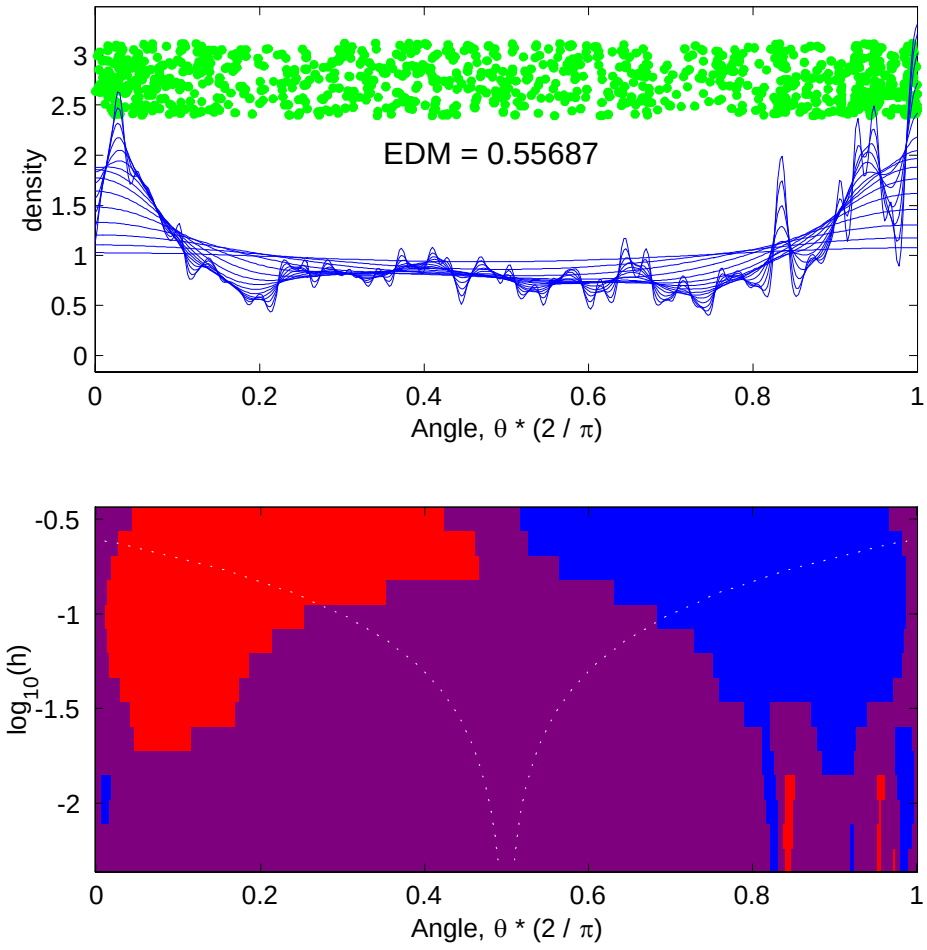

FIGURE 8: *Angular rank extremal dependence analysis of duration vs. size for HTTP response data. Note the slight tendency towards extremal independence.*

Once again, the conclusion is between that of Figures 6 and 7. There is an impression of a bathtub shape (indicating extremal independence) which is

confirmed by large blue and red regions in the SiZer map. But there is also a statistically significant, although rather small spike near $\theta \times (\pi/2) = 0.83$ (indicating that duration is occasionally a particular multiple of size) of the type in Figure 7. This suggests a complicated distributional structure, which is a mixture of components having both extremal independence and dependence. The value of $EDM = 0.56$ is also between that for Figures 6 and 7, and again is similar to that from Figure 5.

3.3 Comparison of Methods

In Sections 3.1 and 3.2, the ICRT and angular rank methods of extremal dependence were studied in the context of a single real data set, using a single threshold. Here we further compare these methods, using a wider range of data sets and thresholds.

First, we consider other values of the threshold, that was set at 2000 in Sections 3.1 and 3.2. Movie versions of Figure 3-8, showing a range of thresholds, from 100 to 2000, are available on the same web site:
http://www.cs.unc.edu/Research/dirt/proj/marron/ExtremalDependence/.

We have also done similar analyses, for all 7 days of the week, and all 3 times of day. The resulting values of EDM are summarized in the spreadsheet summary_100K.xls, which is available from the same web site. Figure 9 shows a graphical display for and even broader set of threshold values from 100 to 10,000, using a parallel coordinates plot, see Inselberg (1985). The horizontal axis in Figure 9 is the threshold level. The vertical axis is the value of the EDM, at that threshold. Each curve represents a single date and time. Colors are used to code the pairs of variables, with the comparisons of: Throughput (Rate) vs. Size shown as red, Duration vs. Size shown as blue, and Inverse Throughput vs. Duration shown as green. Finally, because weekends were noticeably different from weekdays, weekdays are shown as dashed curves, while weekends are solid. Results based on the ICRT method appear in the top panel, and those using the angular rank method appear in the bottom.

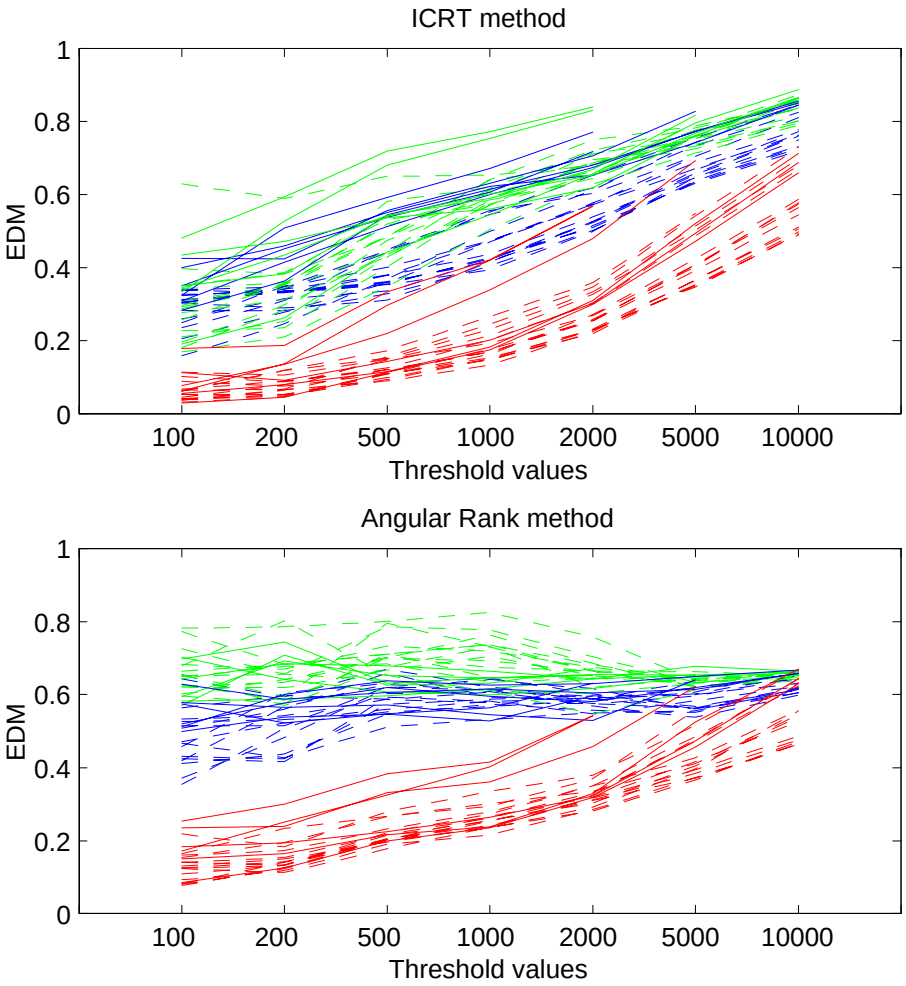

FIGURE 9: *Parallel coordinate plot summaries of EDM, for different thresholds (horizontal axis), different time periods (different curves, weekdays are dashed weekends are solid), different variable pair (T vs. S in red, D vs. S in blue, and IT vs. D in green), and different transformations (ICRT in the top panel, angular rank in the bottom panel).*

Figure 9 shows, for both methods, the main ideas from Sections 3.1 and 3.2 hold quite generally. In particular, Inverse Throughput vs. Duration (green) shows clearly the most dependence among the large values, while Throughput vs. Size (red) is the closest to independent, with Duration vs. Size (blue) lying in between.

Figure 9 also allows other types of insights. First note that the ICRT values of EDM steadily increase with the threshold. For the angular rank method, the EDM values are rather constant, except for Throughput vs. Size (red). This gives an impression that the angular rank values of EDM are less threshold

dependent than for the ICRT.

Also observe that on the weekends (shown as the solid curves), the values of EDM are generally larger than on weekdays. For Duration vs. Size (blue) this is expected, because on weekends traffic is lighter resulting in less congestion and packet loss, so Duration is more likely to be a multiple of size. For Throughput vs. Size (red) this also makes sense, because many large recreational files (music, movies, etc.) are likely being downloaded, and throughputs should be relatively fast because of lack of congestion.

This view also reveals the limits of what can be done using this analysis on these data. Some of the curves in Figure 9 do not extend over all of the threshold values. This is because of our restriction of attention to responses with size larger than 100 kilobtyes, as discussed in the introduction. Note that the curves tend to turn upwards at the right ends, as the asymptotic underpinnings of this analysis become increasingly invalid. In particular, our motivating mathematics are driven by tail considerations, and do not consider the body of the distribution. As this threshold increases, the analysis moves further out of the tail region and into the body of the distribution, which will often have quite different characteristics, as observed here. However, it is interesting that the main lesson, of Throughput vs. Size showing somewhat more independence than the others, still continues to hold, even in this gray region.

4 Methodological Details

This section gives methodological details for the ICRT method, used in Section 3.1, and for the angular rank method, used in Section 3.2.

4.1 Inverse Complementary Rank Transform

The key to making axes comparable, before studying the polar coordinates in the extremal independence analysis of Section 3.1, is the ICRT (Inverse Complementary Rank Transformation) defined in Section 2. This transformation is based on the notion of complementary ranks, as defined at (15).

The ICRT essentially uses the complementary ranks for both marginal distributions simultaneously, but in a way that preserves the critical bivariate structure. In particular, map the set of pairs

$$\{(X_i, Y_i) : i = 1, \ldots, n\}$$

to the pairs

$$\left\{ \left(\overline{R}_i^{(X)}, \overline{R}_i^{(Y)} \right) : i = 1, ..., n \right\}.$$

That is, we replace each data value by its corresponding complementary rank. This bivariate complementary rank transformation is essentially the "Copula Transformation" that is used in the area of rank based non-parametric statistics to study dependence between variables in a distribution free manner.

While the Copula indeed maintains the dependence structure in the data, for purposes of studying extremal dependence it has the serious drawback of giving essentially uniform marginal distributions (ruling out strong potential axis hugging in the scatterplot). The ICRT remedies that, essentially by turning the uniform marginals into Pareto marginals, by working with the *inverse* of the $\overline{R}_i^{(X)}$. In particular the data are mapped to

$$\left\{\left(1/\overline{R}_i^{(X)}, 1/\overline{R}_i^{(Y)}\right) : i = 1, ..., n\right\}.$$

This transformation allows one to recover the exponent and angular measure associated with the standard case. See de Haan and de Ronde (1998), Huang (1992), and pages 207-208 of Resnick (2004a). These are the raw data used in the ICRT extremal dependence analyses of Section 3.1.

4.2 Angular Rank Method

As opposed to the ICRT analysis which applies a non-parametric transformation to the marginal distributions, the angular rank method applies the non-parametric transformation directly to the angles θ.

The Angular Probability Integral Transform (APIT), applied to the full data set, makes the angles equally spaced on $[0, \pi/2]$. One replaces the angles by their ranks (scaled to fill the interval $[0, \pi/2]$). More precisely, given a set of angles, $\theta_1, ..., \theta_n$, define the rank of θ_i to be the number of angles in the set that are less than or equal to θ_i:

$$R_i^{(\theta)} = \sum_{j=1}^{n} 1_{[\theta_j \leq \theta_i]} = \#\left\{j : \theta_j \leq \theta_i\right\}. \tag{18}$$

Then to make the data equally spaced on $[0, \pi/2]$, the set of angles is replaced by $\widetilde{\theta}_1, ..., \widetilde{\theta}_n$ where

$$\widetilde{\theta}_i = \left(\frac{R_i^{(\theta)} - 1/2}{n}\right)\frac{\pi}{2}.$$

After the full data are transformed by the APIT, the large values are again considered (by thresholding on the radius r), so piling up of the angles at the endpoints, with statistical significance assessed by SiZer, provides a clearly interpretable sense in which large values tend to occur together or separately. This time the comparison is done with non-thresholded angles in contrast to comparison with the uniform distribution in Section 3.1. The main ideas are illustrated in Figure 10, in the context of $n = 50,000$ simulated data points from the independent bivariate Pareto (1.5) distribution.

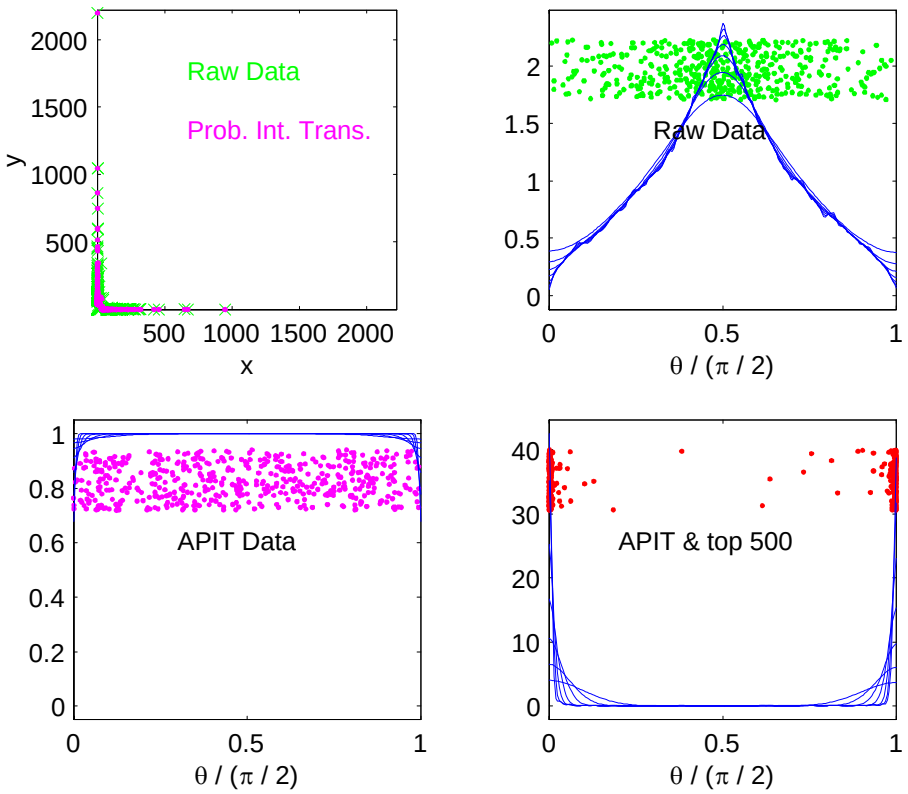

FIGURE 10: *Simulated independent Pareto (1.5) data (shown using x symbols in the upper left), illustrating the Angular Probability Integral Transformation (thick dots in the upper left), and effect of large value thresholding (lower right).*

A scatterplot of the data is shown as x symbols in the upper left panel. The corresponding angular distribution is shown in the upper right panel, as both the jitter plot, and also the family of smooth histograms. This distribution has a distinct peaked shape, showing that the scatterplot does not give much insight into the angular distribution (e.g. one might expect more of a bathtub shaped distribution). The angular rank visualization investigates specifically whether larger values are dependent by first transforming the data from Cartesian to polar coordinates and then replacing angles by their ranks resulting in equally spaced angles, using the APIT. After transforming back to Cartesian coordinates, this is shown as dots in the upper left, and the angles themselves and corresponding estimated angular density shown in the lower left. The dots in the upper left tend to follow very closely the x symbols of the raw data (in particular each dot lies virtually at the center of the corresponding x), suggesting that the APIT does not make large changes in the data for this example. This is because the angular distribution shown in the top right is not so far from the uniform distribution. Of course the APIT data in the lower left look

uniformly distributed.

The angular rank method next focuses on large values, by considering radii r, and restricting attention to only the largest values. In the lower right of Figure 10 the data have been thresholded to the largest 500. The lower right shows that the independent Pareto exhibits strong extremal independence, because the large value thresholded angular distribution is once again bathtub shaped, i.e. the scatterplot data are "axis hugging". This viewpoint is a simple and natural way of studying whether the larger values are associated with each other, but it steps outside the classical ideas of asymptotic independence, because of the APIT, hence the different name.

SiZer maps are not included in Figure 10 to save space, and because they show only the expected results for this simulated example. As seen in Section 3.1, SiZer maps are useful for EDM, and also very useful for the angular rank method because the APIT provides a null uniform distribution, which will show up entirely purple in the SiZer map. Thus, structure which is found in the SiZer map has immediate implications in terms of extremal dependence. As with the ICRT extremal dependence, it remains critical to make the axes comparable before taking polar coordinates. Figure 11 provides an investigation of this issue, for the Wednesday afternoon UNC response size data set considered above. This time the variables are taken to be duration time vs. size, as studied in Figures 2, 5 and 8 above. The various parts of Figure 11 show the results of experimentation, using the simple multiplicative rescaling (19), of the effect of various choices of the scales s_x and s_y. in the scale transformation

$$(x/s_x, y/s_y)\,. \tag{19}$$

In each case the results of a full extremal dependence analysis are shown. This means that after rescaling, the APIT is applied, then the data are thresholded to the largest 200, and finally, the angular distribution is shown. Again the SiZer analysis is not shown to save space.

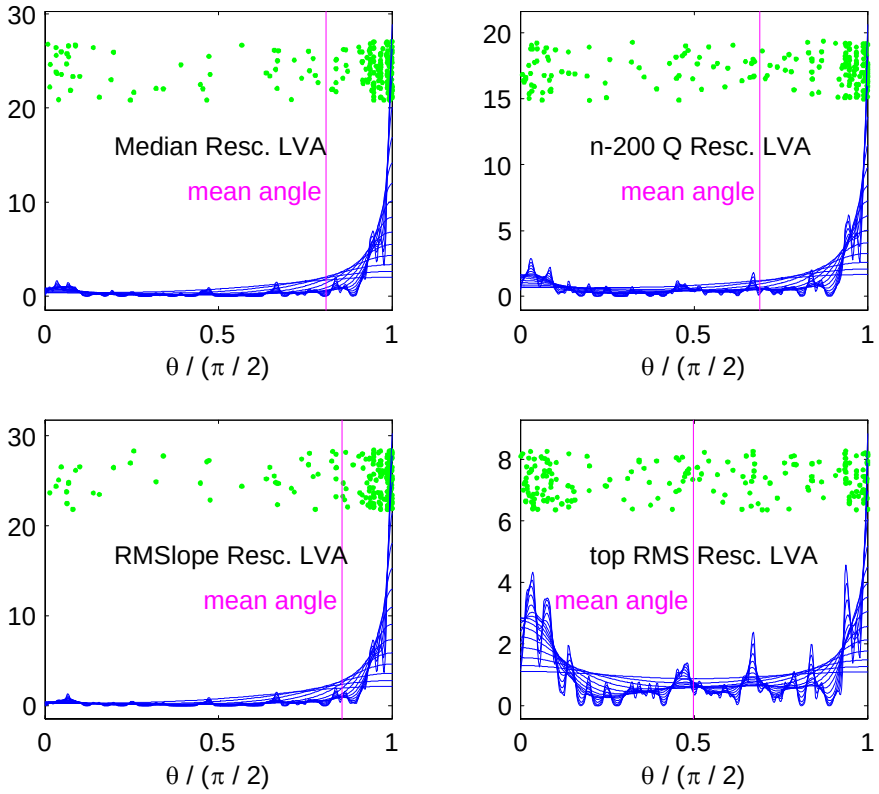

FIGURE 11: *Angular rank distributions, for various rescalings: simple median (upper left), upper tail quantile (upper right), Root Median Slope (lower left), upper tail Root Median Slope (lower right)*

While median rescaling, shown at the upper left of Figure 11, is a simple approach to making axes comparable, it is not helpful for the EDM, because the angles tend to pile up on the right side. This means that in the original scatterplot, there is a very strong tendency for the large values to hug the vertical axis, with very few near the horizontal axis. This scale is not useful for studying extremal dependence, because only large y values are visible in the scatterplot, yet it is the interaction between the large y values and the large x values that is being studied. While this point is clear in this case, precise quantification can be done by computing the mean of the 200 angles shown. This is shown in each plot as the vertical magenta line.

A simple solution is to recognize that all data appear on the right, because the y distribution has a much greater percentage of values that are very large in comparison to the median, which suggests that a more reliable result can be obtained by replacing median rescaling by a larger quantile rescaling. Since only the largest 200 data points are ultimately considered, some improvement can be expected from replacing the medians $\overline{R}^{(x)}_{n/2}$ and $\overline{R}^{(y)}_{n/2}$, by quantiles that

are "200 from the end", i.e. $\overline{R}_{200}^{(x)}$ and $\overline{R}_{200}^{(y)}$. The result of the rescalings $s_x = \overline{R}_{200}^{(x)}$ and $s_y = \overline{R}_{200}^{(y)}$, is shown in the top right panel of Figure 11. There is some improvement over median rescaling, in that the blue spike on the right is thinner, there are more green dots on the left, and the mean angle has moved towards the center. However, the result is still not adequate for useful extremal dependence analysis, because again the angles only pile up at one end.

While neither of these marginal quantiles (nor others that were tried) solved this problem, there must be some transformation that will do so. For example, there should be some scale transformation, of the form (19), that will move the mean angle shown by the vertical lines in Figure 11 to the center.

Another approach is the Median Slope idea. This is a methodology for choosing "aspect ratio", see Cleveland (1993). The aspect ratio in a conventional two dimensional graphic (such as the scatterplots in Figures 1 and 2) quantifies the relationship between the scales of the x and y axes. When curves are displayed, typically aspect ratios are chosen to "maximally utilize graph space", i.e. so that extreme x and y coordinates of the curve fit within the allowed area. However, this view can provide a quite deceptive impression of the data. The Median Slope approach to this problem is to choose the aspect ratio so the median slope of the line segments (the microscopic piecewise lines which make up the apparently smooth curves used in computer graphics) is 1. This same idea is applied in the present context by considering the line segment that connects each data point (x, y) to the origin $(0, 0)$. These line segments have slopes x/y. For a given set of data $(x_1, y_1), .., (x_n, y_n)$, define the Root Median Slope:

$$RMS = \sqrt{\underset{i=1,...,n}{median}\left(\frac{x_i}{y_i}\right)}.$$

When the data are transformed using the scale transformation (19), with

$$s_x = RMS,\ s_y = 1/RMS, \tag{20}$$

note that

$$\underset{i=1,...,n}{median}\left(\frac{x_i/s_x}{y_i/s_y}\right) = \underset{i=1,...,n}{median}\left(\frac{x_i}{y_i}\right)\left(\frac{s_y}{s_x}\right) = RMS^2 RMS^{-2} = 1,$$

i.e. the median slope is one.

The extremal dependence analysis, that results from the RMS scale transformation, using (20) in (19), is shown in the lower left panel of Figure 11. This result is disappointing, being no better than the median rescaling directly above. The reason is the same as discussed above: the median behavior is driven by the center of the distribution, but the extremal dependence analysis feels only the large values.

This suggests applying the extremal dependence analysis to only angles corresponding to larger values. The challenge here is to define "larger". The polar coordinate representation won't work, because that depends on first finding a suitable scaling. Several approaches were tried, with the most success coming

from taking the data points with the largest 100 x coordinates, together with the data points with the largest 100 y coordinates (with duplicates counted twice). The extremal dependence analysis resulting from this rescaling is shown in the lower right panel of Figure 11. Now the result is quite impressive, with the blue smoothed histograms sloping up at both ends, and the mean angle being very close to the center.

While this rescaling gives good performance for these data, the performance was unfortunately much worse in other cases, including the comparisons of throughput vs. size and inverse throughput vs. duration. This suggests that no single rescaling will work uniformly well in all cases. However, as noted above, there is a rescaling that will work well in the sense of moving the mean angle to the center, so we propose finding this by an iterative approach.

To understand our iterative approach to rescaling, consider the rescalings s_x and s_y to be free parameters to be chosen later. These are initially taken to be the best rescaling above, and then will be adjusted to achieve "balance" of the thresholded set of angles. For a given choice of s_x and s_y, the data are transformed to polar coordinates. Next the APIT is applied, and finally the data are thresholded in terms of the radius, keeping only the angles with largest corresponding radii.

The initial rescaling was very effective when these angles hugged both axes to about the same degree, i.e. if the thresholded angles "pile up evenly" on the right and the left of the interval $[0, \pi/2]$. As noted above, a simple notion of "piling up evenly" is that the sample mean is equal to the angular interval centerpoint of $\pi/4$. The next step is iterative improvement of the initial rescaling, with the goal of moving the sample mean $\bar{\theta}$ of the thresholded angles towards $\pi/4$. This is done by adjusting the rescaling to

$$RMS \times \sqrt{\tan \overline{\theta}},$$

where $\overline{\theta}$ is the mean angle. The iterative procedure was ended either when $\left|\overline{\theta} - \frac{\pi}{4}\right| < 0.01$ or else when 100 steps had been taken. The convergence was usually quite fast, but there were a few cases where the convergence was slow, and some where there seemed to be "oscillation" between local solutions, without convergence. However, overall this gave reasonable answers, as shown in the extremal dependence analyses in Section 3.2.

The lower right part of Figure 11, Figures 6 and Figure 8 suggest that while a reasonable job of "putting the axes on the same scale" is done by the iterative end equal rescaling, some improvement may be possible. For example the left endpoint pile in the lower right of Figure 11 is "short and wide", while the right endpoint pile is "tall and narrow". This asymmetry is somewhat unpleasant, and cannot be improved by modifying the multiplicative rescaling (19), because that will instead only shift the angular mean $\overline{\theta}$. A non linear transformation of the axes is needed to handle this effect, and the power transformation (17), seems ideal for this. An interesting open problem is to find a method (perhaps iterative?) for choosing the powers α_x and α_y, to make the thresholded APIT distributions "more symmetric" in some meaningful sense.

It is tempting to combine the ICRT, that was the major basis of the extremal dependence analysis done in Sections 3.1 and 4.1, with the angular rank based analysis of Sections 3.2 and 4.2. This can be done by applying the angular rank analysis to the ICRT data. This was tried, but the answer was not particularly useful because all three pairs of variables then gave strong bathtub shaped distributions, i.e. they all showed extremal independence. The views actually shown in this paper seem to be more useful, because they give a more clear view of "which variables exhibit relatively more extremal dependence".

4.3 Angular Rank Method Asymptotics

The probabilistic background surveyed in Section 2, used to offer some justification of the ICRT method, can also be used to understand large sample properties of the angular rank method in the standard case when (1) or (2) hold.

Let $\{(X_i, Y_i), 1 \leq i \leq n\}$ be the original data in Cartesian coordinates, whose common joint distribution satisfies (1) or (2). Transform the data to polar coordinates $\{(r_i, \Theta_i), 1 \leq i \leq n\}$ and (2) is equivalent to (see, for example, Resnick (1987); Resnick (2002); Resnick (2004a); Basrak (2000); Basrak, Davis and Mikosch (2002)) existence of $b(n) \uparrow \infty$ such that

$$nP[\big(\frac{r_1}{b(n)}, \Theta_1\big) \in \cdot] \xrightarrow{v} \nu_\alpha \times S_*, \tag{21}$$

vaguely on $(0, \infty] \times [0, \pi/2]$, where ν_α is the measure with tail

$$\nu_\alpha(w, \infty] = w^{-\alpha}, \quad w > 0,$$

and S_* is a probability measure on $[0, \pi/2]$. It is also the case that for $k = k(n) \to \infty$ such that $k/n \to 0$

$$\frac{1}{k} \sum_{i=1}^{n} \epsilon_{\big(\frac{r_i}{b(n/k)}, \Theta_i\big)} \Rightarrow \nu_\alpha \times S_*, \tag{22}$$

weakly, in the space of Radon measures on $(0, \infty] \times [0, \pi/2]$.

Now we create the ranks of the Θ's. For the n angles $\Theta_1, \ldots, \Theta_n$, write the order statistics in increasing order as

$$\Theta_{(1:n)} \leq \cdots \leq \Theta_{(n:n)}$$

and define the rank of Θ_i as at (18), so that again $R_i^{(\Theta)}$ is the number of Θ's which are no bigger than Θ_i. It follows that for $1 \leq \ell \leq n$

$$R_i^{(\Theta)} \leq \ell \text{ iff } \Theta_i \leq \Theta_{(\ell:n)}. \tag{23}$$

Having ranked the angles, we now take the subset corresponding to the original data which exceeds a threshold according to radius vectors of the data.

We retain the angle ranks with indices i corresponding to $r_i > b(\frac{n}{k})$. The empirical measure of these ranks is

$$\begin{aligned}\hat{T}_n(\cdot) =& \frac{\sum_{i=1}^n 1_{[r_i > b(n/k)]} \epsilon_{R_i^{(\Theta)}/n}}{\sum_{i=1}^n 1_{[r_i > b(n/k)]}} \\ =& \frac{\sum_{i=1}^n \epsilon_{\left(\frac{r_i}{b(n/k)}, \frac{R_i^{(\Theta)}}{n}\right)} \left((1,\infty] \times \cdot\right)}{\sum_{i=1}^n \epsilon_{\left(\frac{r_i}{b(n/k)}, \frac{R_i^{(\Theta)}}{n}\right)} \left((1,\infty] \times [0,1]\right)}. \end{aligned} \tag{24}$$

For the empirical measure of the ranks, we have the following result.

Proposition 1 *Suppose the joint distribution F of the original data $\{(X_i, Y_i), 1 \leq i \leq n\}$ satisfies (1) (or equivalently (2) or (21)) so that the standard case assumptions are in force. Define for $0 \leq \theta \leq \pi/2$*

$$H(\theta) = P[\Theta_1 \leq \theta] = P[\arctan\left(\frac{Y_1}{X_1}\right) \leq \theta]$$

for the marginal distribution of Θ_1. Then

$$\hat{T}_n \Rightarrow S_* \circ H^{\leftarrow} \tag{25}$$

weakly in the space of Radon measures on $[0,1]$. Here $H^{\leftarrow}$ is the left continuous inverse of H.

Remark. Thus, the empirical measure of ranked angles, after pruning the ranks by thresholding, approximates $S_* \circ H^{\leftarrow}$. In cases where (X_1, Y_1) has a positive density on $\mathbb{R}_+^2$, H is continuous and strictly increasing and thus $H^{\leftarrow} : [0,1] \mapsto [0, \pi/2]$ is also continuous and strictly increasing. If asymptotic independence holds so that S_* is the two-point distribution concentrating on $\{0, \pi/2\}$, then $S_* \circ H^{\leftarrow}$ is the two-point distribution concentrating on $\{0,1\}$. In such cases, even in the absence of knowledge of H, $\hat{T}_n$ should be effective in discerning extremal independence.

Proof. For $0 \leq t \leq 1$

$$\hat{T}_n([0,t]) = \frac{\sum_{i=1}^n \epsilon_{\left(\frac{r_i}{b(n/k)}, \frac{R_i^{(\Theta)}}{n}\right)} \left((1,\infty] \times [0,t]\right)}{\sum_{i=1}^n \epsilon_{\left(\frac{r_i}{b(n/k)}, \frac{R_i^{(\Theta)}}{n}\right)} \left((1,\infty] \times [0,1]\right)}.$$

Now

$$\frac{R_i^{\mathbf{\Theta}}}{n} \leq t \text{ iff } \Theta_i \leq \Theta_{([nt],n)}.$$

As a process, $\{\Theta_{([nt],n)}, 0 \leq t \leq 1\}$ is the inverse of the empirical cdf of the Θ's. Since the empirical cdf of the Θ's converges almost surely to H weakly, the same is true of the inverses and hence almost surely

$$\Theta_{([n\cdot],n)} \to H^{\leftarrow}(\cdot)$$

at points of continuity of the limit. We have, therefore,

$$\hat{T}_n\left([0,t]\right) = \frac{\frac{1}{k}\sum_{i=1}^{n} \epsilon_{\left(\frac{r_i}{b(n/k)},\Theta_i\right)}\left((1,\infty] \times [0,\Theta_{([nt]:n)}]\right)}{\frac{1}{k}\sum_{i=1}^{n} \epsilon_{\left(\frac{r_i}{b(n/k)},\Theta_i\right)}\left((1,\infty] \times [0,\pi/2]\right)} \Rightarrow S_* \circ H^{\leftarrow}(t)/1 = S_* \circ H^{\leftarrow}(t),$$

provided t is a continuity point of $H^{\leftarrow}$. ■

5 Concluding Remarks

These results should encourage additional measurements to confirm the observations at a variety of network locations (access, edge, and backbone links), and to determine the range of file sizes for which the independence result holds.

Another interesting issue is the opposition of our results with those of Zhang, Breslau, Paxson and Shenker (2002). In an upcoming paper, we will show that this is driven by biases created by choice of variable on which to threshold, that have been used to select "large values", and will conclude that the methods in the present paper are sensible approaches.

The angular rank method seems interesting and promising and we intend to continue investigating mathematical bases for the method. In particular, we would like to understand under what conditions extremal dependence analysis based on either the ICRT or the angular rank method give the same conclusions and whether it is possible to conclude that one method is superior to the other.

6 Acknowledgements

The collaboration of this paper is a result of the course OR778 at Cornell University, during the Fall of 2001. J. S. Marron is grateful to the Cornell University School of Operations Research and Industrial Engineering for hospitality, support, and an exciting research environment. The research of J. S. Marron was supported by Cornell University's College of Engineering Mary Upson Fund and NSF Grants DMS-9971649 and DMS-0308331. That of Sidney Resnick was supported by NSF grant DMS-0303493 and NSA grant MSPF- 02G-183 at Cornell University.

References

[1] Basrak, B. (2000) *The sample autocorrelation function of non-linear time series*, PhD thesis, Rijksuniversiteit Groningen, Groningen, Netherlands.

[2] B. Basrak, R. A. Davis, and T. Mikosch. (2002) A characterization of multivariate regular variation, *The Annals of Applied Probability*, 12, 908–920.

[3] Chaudhuri. P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807–823.

[4] Cleveland, W. S. (1993) *Visualizing Data*, Hobart Press, Summit, New Jersey, U.S.A.

[5] Coles, S. Heffernan, J. and Tawn, J. (1999) Dependence measures for extreme value analyses, *Extremes*, 2, 339–365.

[6] de Haan, L. and Resnick, S. (1993) Estimating the limit distribution of multivariate extremes. *Comm.Statist. Stochastic Models*, 9(2):275–309.

[7] de Haan, L. and de Ronde, J. (1998) Sea and wind: multivariate extremes at work, *Extremes*, 1, 7–46.

[8] Einmahl,J., de Haan, L. and Piterbarg, V. (2001) Nonparametric estimation of the spectral measure of an extreme value distribution, *Annals of Statistics*, 29(5):1401–1423.

[9] Hannig, J., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2001) Log-normal durations can give long range dependence, unpublished manuscript, web available at http://www.unc.edu/depts/statistics/postscript/papers /marron/NetworkData/LogNorm2LRD/.

[10] Heath, D., Resnick, S. and Samorodnitsky , G. (1998) Heavy tails and long range dependence in on/off processes and associated fluid models, *Mathematics of Operations Research*, 23, 145–165.

[11] Hernandez-Campos, F., Jeffay, K. and Smith, F. D. (2003) Tracing the Evolution of Web Traffic: 1995-2003, ACM MASCOTS 2003, Orlando, FL.

[12] Hernandez-Campos, F., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2002) Variable Heavy Tailed Durations in Internet Traffic, unpublished manuscript, web available at http://www-dirt.cs.unc.edu/marron/VarHeavyTails/.

[13] Huang, X. (1992) *Statistics of Bivariate Extreme values*, PhD thesis, Tinbergen Institute Research Series 22, Erasmus University Rotterdam, Postbus 1735, 3000DR, Rotterdam, The Netherlands.

[14] Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer*, 1, 69–91.

[15] Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values, *Biometrika*, 83, 169–187.

[16] Ledford, A. W. and Tawn, J. A. (1997) Modelling dependence within joint tail regions, *Journal of the Royal Statistical Society, Series B*, 59, 475–499.

[17] Marron, J. S., Hernandez-Campos, F. and Smith, F. D. (2001) A SiZer analysis of IP Flow start times, unpublished manuscript.

[18] Maulik, K. and Resnick, S. and Rootzén, H. (2002) Asymptotic independence and a network traffic model, *Journal of Applied Probability*, 39, 671–699.

[19] Poon, S.- H., Rockinger, M. and Tawn, J. (2003) Modelling extreme-value dependence in international stock markets, *Statistica Sinica*, 13, 929–953.

[20] Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*, Springer-Verlag, New York.

[21] Resnick, S. and Samorodnitsky, G. (1999) Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues, *Queueing Systems*, 33, 43–71.

[22] Resnick, S. (2002) Hidden regular variation, second order regular variation and asymptotic independence, *Extremes*, 5, 303–336.

[23] Resnick, S. (2003) Modeling Data Networks, in *SEMSTAT, Seminaire Europeen de Statistique, Exteme Values in Finance, Telecommunications, and the Environment*, eds. B. Finkenstadt and H. Rootzén, Chapman and Hall. London.

[24] Resnick, S. (2004) On the foundations of multivariate heavy tail analysis. in *Stochastic Methods and their Applications, J. Applied Probability Special Volume 41A; Papers in honour of C.C. Heyde*, eds. J. Gani and E. Seneta, 191–212.

[25] Resnick, S. I. (2004b) The Extremal Dependence Measure and Asymptotic Independence, *Stochastic Models*, 20, 205–227.

[26] Smith, F. D., Hernandez, F., Jeffay, K. and Ott, D. (2001) "What TCP/IP Protocol Headers Can Tell Us About the Web", *Proceedings of ACM SIGMETRICS 2001/Performance 2001*, Cambridge MA, June 2001, pp. 245–256.

[27] Stărică, C. (1999) Multivariate extremes for models with constant conditional correlations, *Journal of Empirical Finance*, 6, 515–553.

[28] Stărică, C. (2000) Multivariate extremes for models with constant conditional correlations, in P. Embrechts, Ed., *Extremes and Integrated Risk Management*, Risk Books, London

[29] Tukey, J., and Tukey, P. (1990). Strips Displaying Empirical Distributions: Textured Dot Strips. Bellcore Technical Memorandum.

[30] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, New York.

[31] Zhang, Y., Breslau, L., Paxson, V. and Shenker, S. (2002) On the characteristics and origins of internet flow rates, in *Proceedings of SIGCOMM02*, ACM, Pittsburgh.