Thresholded Log-Log Correlation Analyses of HTTP Response Characteristics

Cheolwoo Park
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

Felix Hernandez Campos
Department of Computer Science
University of North Carolina
Chapel Hill, NC 25799-3175

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 25799-3260

F. D. Smith
Department of Computer Science
University of North Carolina
Chapel Hill, NC 25799-3175

June 5, 2003

Abstract

Dependencies between variables characterizing HTTP responses are studied. Earlier results, based on thresholded log-log correlations and other methods are contradictory. The contradiction is explained using a more general thresholded analysis. The analysis reveals that thresholded log-log correlation is an especially treacherous way of understanding the large value dependence of distributions. Hence the more recent Extremal Dependence Measure is recommended.

1 Introduction

HTTP responses are the data (components of a web page) that are sent from a web server, in response to a request from a web browser. Both protocol

researchers and web page developers are interested in the size and time characteristics of these responses, with a focus on the responses that are larger and/or take a long time.

A straightforward tool for studying the dependence between say the variables of size and duration (time required for the transfer) is to compute the standard Pearson correlation coefficient, of the variables on the log-log scale. It is important to take logs, because these distributions tend to be heavy tailed (see Hernández-Campos, Marron, Samorodnitsky and Smith (2002) for a particularly deep analysis of this type) in a way that drastically impacts quadratic based measures such as variance and correlation. Because of the interest in the larger values of these variables, and to avoid the influence of TCP effects including slow start, it is natural to threshold the data before computing the correlation.

This type of analysis was done by Zhang, Breslau, Paxson and Shenker (2002), who analyzed several data sets, and reported that flow rates (throughput) were strongly correlated with size, and that duration (time of transfer) had a weak or non-existent correlation with both rate and size. These results are contradictory to those of Hernández Campos, Marron, Resnick, Park and Jeffay (2003), who used the completely different Extremal Dependence Analysis, and reported that size and rate were not correlated, while duration and inverse rate had a strong correlation. Likely explanations of the different results include the different data sets studied, and the different analysis methods used. These issues are studied in detail in Section 1.1.

A first simple attempt at understanding the different results is to replace the Extremal Dependence Analysis of Hernández Campos et al. (2003) with a log-log correlation analysis. When this is done (in Section 1.1), the same contradiction remains. Thus the difference in results is not caused by the method of analysis. A deeper look at the respective analyses, reveals another difference. The Zhang, et al. (2002) analysis restricted attention to "large responses" by thresholding above 5 seconds in duration, while the Hernández Campos et al (2003) analysis thresholded to responses that were larger than 100 kilobytes. This seemingly minor difference turns out to be surprisingly important.

This difference between thresholding is carefully studied in Section 2, by combining a wide range of possible thresholds with some useful visualizations. This analysis explains the widely different conclusions that were drawn by the two papers, and casts considerable doubt on the practice of thresholding before computing log-log correlations.

In Section 2.2 it is seen that this conclusion is not restricted to the particular data at hand, by applying the same analysis to a simulated Gaussian distribution (on the log-log scale), fit to the data. Even in this simple case, thresholding the larger time values gives answers similar to those of Zhang et al (2002), while thresholding the larger size values gives answers similar to those of Hernández Campos et al (2003).

??? Add section on NZ data. ???

We conclude that thresholded log-log correlation is an unreliable approach

to understanding relationships between large values of random variables. Instead we recommend the Extremal Dependence Analysis proposed by Hernández Campos et al (2003).

1.1 Earlier Results

The goal of this section is to directly contrast the results of Zhang et al (2002), with those of Hernández Campos et al (2003). Table 1 allows this by summarizing the respective results. The variables studied are:

- D Duration (sec), the total time needed for transfer of data.
- S Size (bytes), the size of the file being transferred.
- R Rate (bytes/sec), the overall transfer rate (throughput), defined as size divided by duration.
- IR Inverse Rate (sec/byte), the inverse of the overall rate, defined as duration divided by size.

Because larger files need more time to transfer, strong correlation is expected between D and S. However, here we study the larger values of these variables, where the presence or absence of this correlation is not obvious. The inverse rate, IR, is considered because large sizes might correlate with slow rates, i.e. large inverse rates. Note that on the log scale: $\log R = -\log IR$.

The results reported by the two studies are summarized in Table 1, for simple direct comparison. Zhang et al (2002) considered a variety of 8 different traces, and the ranges of thresholded log-log correlations, for the three pairs of variables, are summarized in the middle column of Table 1. Hernández Campos et al (2003) considered 21 traces representing 3 time blocks (8:00AM - 12:00Noon, 1:00PM - 5:00PM and 7:30PM - 11:30PM), on each of the seven weekdays, gathered in April of 2001, on the main internet link of the University of North Carolina, Chapel Hill. This time correlation is expressed in terms of the Extremal Dependence Measure, defined in Hernández Campos et al (2003), with the threshold parameters used there set at 2000. The last column of Table 1, summarizes results for the 15 weekday time blocks. Weekends are excluded in Table 1, because they exhibited increased variation which increased the lengths of the intervals, see Section 3.3 of Hernández Campos et al (2003) for further discussion. Table 1 also gives a single word summary of the conclusions reached, in each paper, about the relative (in)dependence of the pairs of variables.

	Zhang, et al (log-log Corr)	Hernández Campos, et al (E D A)
S vs. D	Independent 0.10 – 0.30	Inconclusive 0.50 – 0.65
S vs. R	Dependent 0.84 – 0.89	Independent 0.22 – 0.38
D vs. IR	Inconclusive 0.18 – 0.45	Dependent 0.55 – 0.76

TABLE 1: *Comparison of thresholded log-log correlation results by Zhang et al (2002), with extremal dependence results by Hernández Campos et al (2003). Shows very strong difference in conclusions.*

The contrast between the results is very stark. Perhaps most striking is the case of S vs. R, ??? Felix, should we argue that this is the most important case? ???, where diametrically opposite conclusions are reported. The other differences are also striking.

As noted in Section 1, there are several possible explanations of these differences, including the different data sets studied, and the different analysis methods used. A simple first step in exploring these explanations is to apply the log-log correlation analysis, used by Zhang et al (2002), to the data from Hernández Campos, et al (2003). The results are shown in Table 2, whose middle column gives the range of log-log correlations for the same 15 weekday data sets as shown in the last column of Table 1. These results carry the same qualitative lessons as for the extremal dependence measure from Table 1. Thus, type of analysis method can be ruled out as the cause of the striking differences. This leaves at least two other possible explanations: differences in the data, or differences in the thresholding.

	Hernández Campos, et al (log-log Corr, S-thresh)	Hernández Campos, et al (log-log Corr, D-thresh)
S vs. D	0.56 – 0.67	0.05 – 0.31
S vs. R	–0.09 – 0.11	0.83 – 0.90
D vs. IR	0.73 – 0.82	0.24 – 0.39

TABLE 2: *Comparison of log-log correlation results, for the data of Hernández Campos, et al (2003), for Size and Duration thresholding. Shows this explains the very divergent early results.*

In Section 2, a careful analysis, involving some novel visualizations, of a wide range of potential thresholds reveals that the latter is the cause. In particular, in Zhang et al (2002), consideration was restricted to "large response", by thresholding to response durations that were more than 5 seconds. However, in Hernández Campos, et al (2003), "large durations" were defined as those with size more than 100 kilobytes. This seemingly mild difference turns our in Section 2 to be critical. The point is illustrated in the last column of Table 2, where log-log correlations for the same 15 weekday data sets from Hernández Campos, et al (2003), are summarized, but this time the data are "Duration-thresholded" (i.e. only responses with duration more than 5 seconds are used), as opposed to the "Size-thresholding" (size more than 100 kilobtyes) that was used in the middle column of Table 2. This rules out different data sets as the explanation for the differences observed in Table 1. This suggests that the critical difference between reported results is due to the type of thresholding employed to arrive at "large data values". Visual insight into this phenomenon, together with a demonstration that this is quite generally a serious issue, is given in Section 2.

2 Global Thresholded Analyses

In Section 2.1, a visualization is developed which clarifies all of the seemingly contradictory results shown in Tables 1 and 2. As suggested by Table 2, the pivotal issue is the type of thresholding. In Section 2.2, it is seen that this unacceptably strong dependence, of the thresholded log-log correlation on the type of thresholding, is not merely an artifact of these data sets, but may be expected to hold quite generally. This is done by showing that the same phenomenon exists even for purely Gaussian distributions. We conclude that a better notion of "(in)dependence of large values" is Extremal Dependence, proposed by Hernández Campos et al (2003).

Because individual data set visualization is the key to this analysis, only the single data set of Wednesday afternoon is considered here. This time block was chosen as "quite representative" by Hernández Campos et al (2003), who also used it in their detailed single data set analyses.

Single packet responses are recorded with a duration D of 0, because the available time information is based on packet time stamps. To handle the difficult case of $\log 0$, we eliminate these responses from the calculations done in this paper.

2.1 HTTP Response Data

Figure 1 illustrates how the difference between the thresholding methods of Zhang et al (2002) and Hernández Campos et al (2003) led to the surprisingly large difference in conclusions. The main idea is to embed the thresholds used in those two papers among a bigger set of thresholds. The full collection of thresholds considered are shown using vertical and horizontal lines. Additional visual insight comes from overlaying the log-log scatterplot of these two variables.

The 3 vertical and 5 horizontal lines divide the plane into $(3+1)\times(5+1) = 24$ cells. Because of the interest in large values of both variables, each cell determines a sub-population which includes all of the cells above and to the right.

Insight into the relative sizes of these sub-populations comes from the numbers shown in parentheses in each cell, which are the respective percentages (of the number of data in the "upper right") relative to the full population. Note that the lower left cell shows (%100), because the corresponding sub-population is the full data set. The numbers decline when moving either to the right, or upwards, because these movements eliminate either rows or columns of data.

The other numbers in the cells show the correlation of the respective upper right sub-populations. The correlation in the lower left cell, 0.54, is the full population correlation. This is not of central interest here, because it feels aspects of the data, that are not central to the issue of ??? Felix, help here please ???. The approach to addressing this issue of Zhang et al (2002) is to only consider responses with a duration of more than 5 seconds, $D > 5$. This results in computing the correlation for only the responses that appear above the top horizontal line in Figure 1, giving 0.19, which is in the range shown in

the last column of Table 2. The approach of Hernández Campos et al (2003) is to only consider the responses with size larger than 100 kilobytes, $S > 10^5$. This is the sub-population to the right of the right-most vertical line, whose correlation is 0.63, which is in the range shown in the middle column of Table 2.

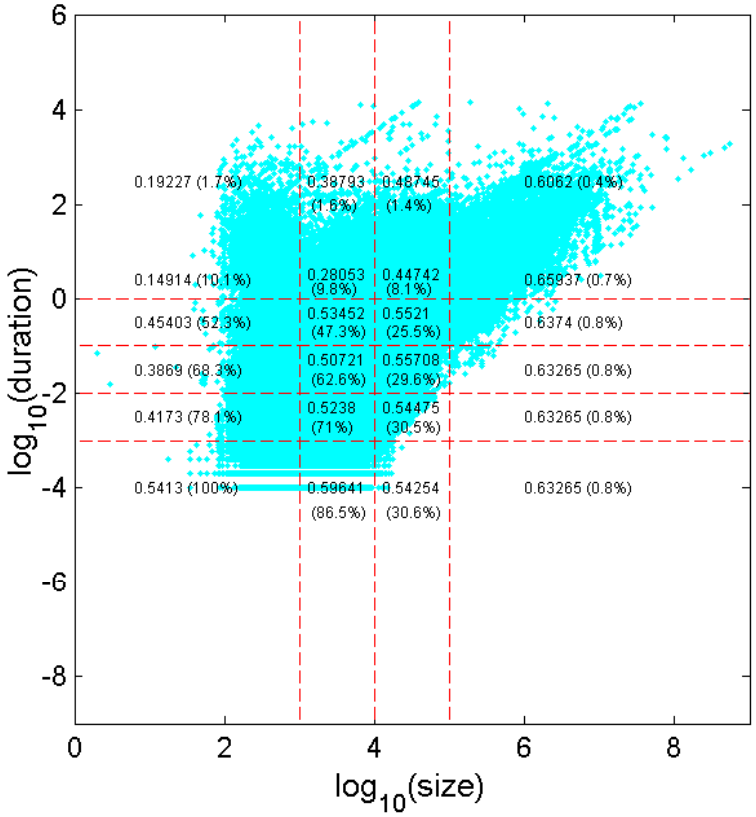

FIGURE 1: *Wednesday afternoon log(D) vs. log(S) scatterplot. Overlaid numbers are correlations (sample percentage), for data that is "upper right" with respect to each cell.*

The other cell correlations provide a bridge between these two, and give a strong impression that thresholded log-log correlation is driven very strongly by the particular thresholding scheme that is used.

Figure 2 is a similar display, for studying the relationship between the file size, S, and the overall transfer rate R. The horizontal axis, $\log S$ is the same as in Figure 1, but now the vertical axis is

$$\log R = \log(S/D) = \log S - \log D. \tag{1}$$

The same cells, and sub-populations from Figure 1 are used here (because the same thresholdings were used by Zhang et al (2002) and Hernández Campos

et al (2003) for these variables as well. The vertical lines are the same as in Figure 1. The horizontal lines from Figure 1, now appear as slanted lines in Figure 2. Because of the minus sign in the transformation (1) the ordering of the lines is reversed, and the sub-populations are now below and to the right of each cell. Thus, the full population cell is now at the upper left. The cell corresponding to the Zhang et al (2002) analysis is in the lower left, with the correlation of 0.86 shown, which lies in the range shown in the last column of Table 2. The cell representing the Hernández Campos et al (2003) analysis is now in the upper right, showing the correlation of –0.05 which fits in the range given in the middle column of Table 2.

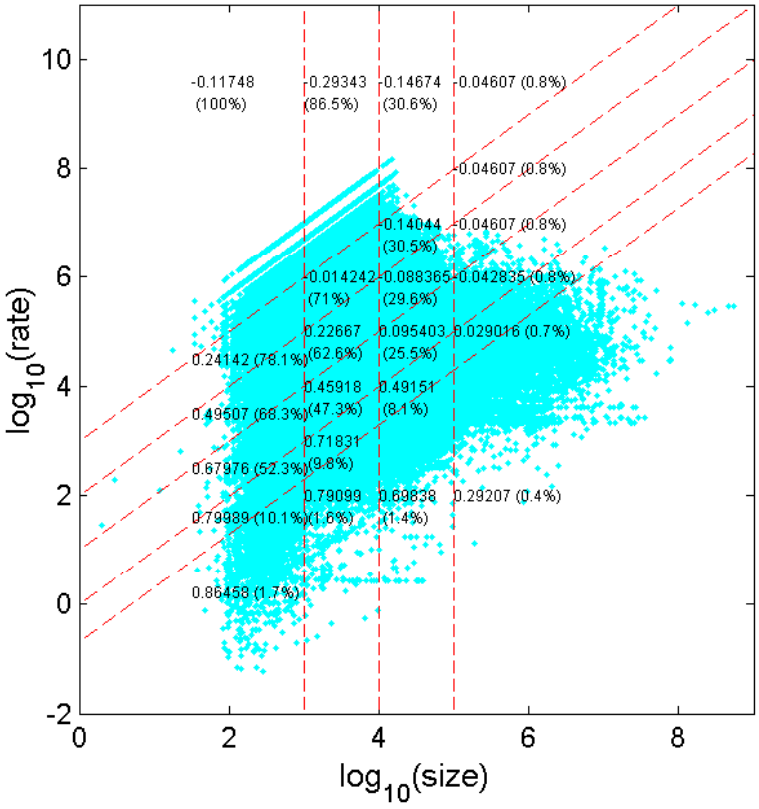

FIGURE 2: *Wednesday afternoon log(R) vs. log(S), overlay on scatterplot. Subpopulation correlations are again strongly driven by thresholding.*

Again the full range of thresholded correlations shows that many interpretations of the data are possible, depending on the choice of this threshold.

The same analysis for the variables D vs. IR resulted in very similar lessons, so the graphic is not shown here.

This casts substantial doubt on the viability of the thresholded log-log correlation as a means of understanding the relationship between the larger values

of joint distributions. However, this analysis was done completely in the context of a single data set, with some perhaps unusual structure as seen in the scatterplot. This still leaves open the important questions: Do these ideas generalize? Are they mere artifacts of this particular data set? This question is answered in the next sections.

2.2 Simulated Gaussian Data

To investigate the performance of the thresholded log-log correlation in a more general context, we first consider a simulated bivariate Gaussian data set with similar characteristics to the log data considered in Section 2.1. In particular, the same sample size is used, and the mean vector and covariance matrix are estimated from the pairs $(\log S, \log D)$.

The scatterplot of the resulting data, from using this $(\log S, \log D)$ distribution to generate the $(\log R, \log S)$ distribution as in (1), is shown in Figure 3. To illustrate the generality of the instability of the thresholded correlation, the same analysis as in Figure 2 is applied.

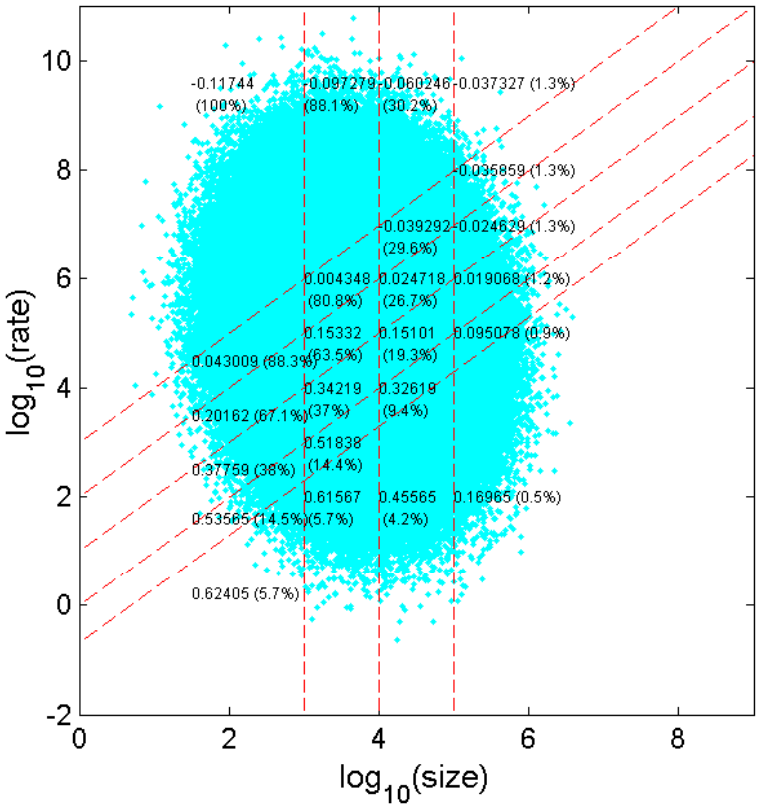

FIGURE 3: *Simulated bivariate normal log(R) vs. log(S), overlay on scatterplot. Shows both subpopulation sizes and thresholded correlations are surprisingly similar to Figure 2.*

In view of the clear non-Gaussianity of the population in Figure 2, it is perhaps surprising that both the percentages of the sub-populations, and the correlations are so similar to those of Figure 2. In particular, the main lesson that the type of thresholding critically impacts the correlation is equally clear.

The Duration thresholding of 5 sec, as used by Zhang, et al (2002) results in a correlation of 0.62. This is somewhat smaller than the range of 0.83-0.9 reported in the middle column of Table 1 and the last column of Table 2. This seems to be caused by the Gaussian distribution putting 5.7% of the data in this region, vs. 1.7% for the real data.

But the Size thresholding of 100 kb, as used by Hernández Campos et al. (2003), results in the far smaller correlation of -0.04, well within the range of –0.09 – 0.11 from the middle column of Table 2. This shows that the dramatic differences in correlation, caused by the different types of thresholding, are not data set specific. In particular, even for simulated log-normal data, these same effects may be expected.

3 Extremal Dependence Analysis

The central lesson of Section 2 is that thresholded log-log correlations are an especially unreliable method for understanding (in) dependence of large values of bivariate distributions. This leads us to recommend the extremal dependence methods of Hernández Campos et al. (2003) as a more viable alternative approach.

We recommend applying extremal dependence techniques to the full data sets, but one may wonder: how stable is this method to Size and Duration thresholding. This issue is investigated in Table 3. The entries of Table 3 are the "extremal dependence measures", based on the Inverse Complementary Rank Transformation, proposed by Hernández Campos, et al (2003), using a radius threshold to the top 2000. The entries in the second column, are for the Duration type thresholding (to durations > 5 seconds), as was done by Zhang, et al (2002). The entries in the third column are for the full data set. The entries in the fourth column are based on the Size type thresholding (to Size $>$ 100 kb), as was done by Hernández Campos, et al (2003).

	D-thresh	full data	S-thresh
S vs. D	0.43	0.40	0.51
S vs. R	0.51	0.00	0.23
D vs. IR	0.28	0.25	0.69

TABLE 3: *Comparison of extremal dependence results, for the data of Hernández Campos, et al (2003), for Duration and Size thresholding as well as for the full data. Investigates robustness of extremal dependence to thresholding.*

Table 3 shows a variety of results. For the Size vs Duration pair, studied in the first row of Table 3, the extremal dependence measure values are relatively close. Thus in this case, the extremal dependence measure is quite robust to the type of thresholding. For the Size vs. Rate pair, The Duration thresholded version is much larger than the others, so in this case the extremal dependence measure is much more sensitive to thresholding. For the Duration vs. Inverse Rate pair, again there are significant differences between values, but this time it is the Size thresholded value that is much larger than the others.

We conclude that while the extremal dependence measure can have some sensitivity to thresholding, it is somewhat more robust to thresholding than simple log-log correlations. However, the (in) dependence of "large values" is most effectively studied by using extremal dependence analysis *starting with the full data set*, so we recommend this whenever possible.

4 Conclusions

This paper made contributions in two important directions. The first direction was methodological, where we show that thresholded log-log correlations provide a very unstable way of understanding (in) dependence of large values

of bivariate distributions, and recommend replacing this by the "extremal dependence measure" of Hernández Campos, et al (2003). The second direction was new ideas for networking. We showed that, contrary to previously published work, the Size and the Rate of HTTP transfers tend to be independent for large values, which is consistent with the conclusions of Hernández Campos, et al (2003). ??? Felix, want to interpret more? ??? An interesting open question is: does this conclusion continue to hold, when the focus is on all IP connections, instead of only on HTTP?

References

[1] Hernández Campos, F., Marron, J. S., Resnick, S. I, Park, C. W. and Jeffay, K (2003) Extremal Dependence: Internet Traffic Applications. Internet available at: http://www.cs.unc.edu/Research/dirt/proj/marron/ExtremalDependence/.

[2] Hernández-Campos, F., Marron, J. S., Samorodnitsky, G. and Smith, F. D. (2002) Variable Heavy Tailed Durations in Internet Traffic, Part I: Understanding Heavy Tails, *MASCOTS2002*, internet available at: http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails/.

[3] Zhang, Y., Breslau, L., Paxson, V. and Shenker, S. (2002) On the characteristics and origins of internet flow rates, *SIGCOMM'02*.