Email Message Interarrival Time Analysis

Lourens O. Walters
Supervised by Prof. Pieter S. Kritzinger

Data Network Architectures Laboratory
Department of Computer Science
University of Cape Town
Rondebosch, 7701
South Africa

E-Mail: {lwalters, psk}@cs.uct.ac.za

Abstract

The arrival pattern of email messages at an email server shows daily and weekly arrival patterns. The interarrival distribution is skewed to the right and has a large variance which suggests heavy-tailed behaviour. Further analysis shows that although the data is severely skewed the tail of the distribution does not follow a power rate and is therefore not heavy-tailed.

1 Introduction

This paper presents the analysis of email message interarrival times at the Department of Computer Science at the University of Cape Town's email server. In particular it investigates message arrival patterns and the possibility of a heavy tailed interarrival time distribution.

The analysis presented is the first part of a study to find the best fit probability distribution for email interarrival times which in turn is part of a study to derive traffic models for Email, File Transfer and Web traffic.

2 The Simple Mail Transfer Protocol

The Simple Mail Transfer Protocol (SMTP) is defined in RFC 821[Pos82]. It is a host-to-host mail transport protocol and traditionally operates over TCP using port 25. Figure 1 illustrates the components involved in email transactions.

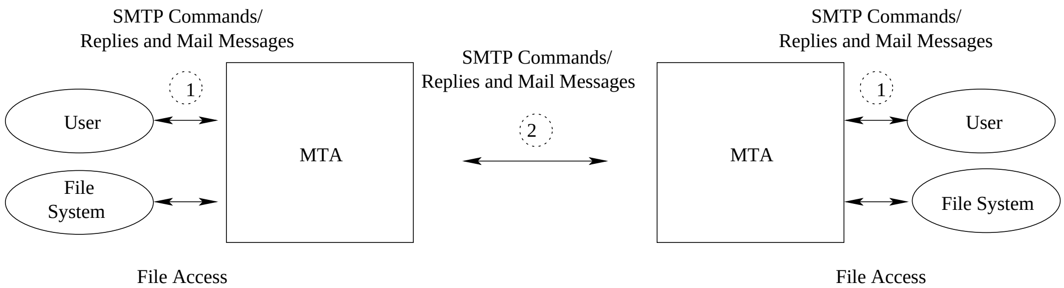

Figure 1: Model of Use for SMTP

Both the MTA's in Figure 1 are email servers. They receive email messages from local users (Action 1) or from other email servers (Action 2). We captured all messages arriving at a server, both from local users and other email servers.

3 Traffic Trace

A 3 month long traffic trace was recorded using the tcpdump and tcpshow Unix utilities. The trace was recorded over the period of 1 November 2000 to 31 January 2001 which fell within the University's holiday period. Even though traffic was less than usual we recorded sufficient data for our purposes.

The information necessary for analysis was extracted from the trace by means of PERL scripts. The email server had approximately 400 user accounts at the time of the measurements.

4 Arrival Pattern

We start by analyzing the arrival pattern of email messages in order to ascertain whether the interarrival process is stationary.

Figure 2 plots the number of arrivals per hour over a five day subset of the data (20 - 24 November). As expected a clear daily pattern is evident.

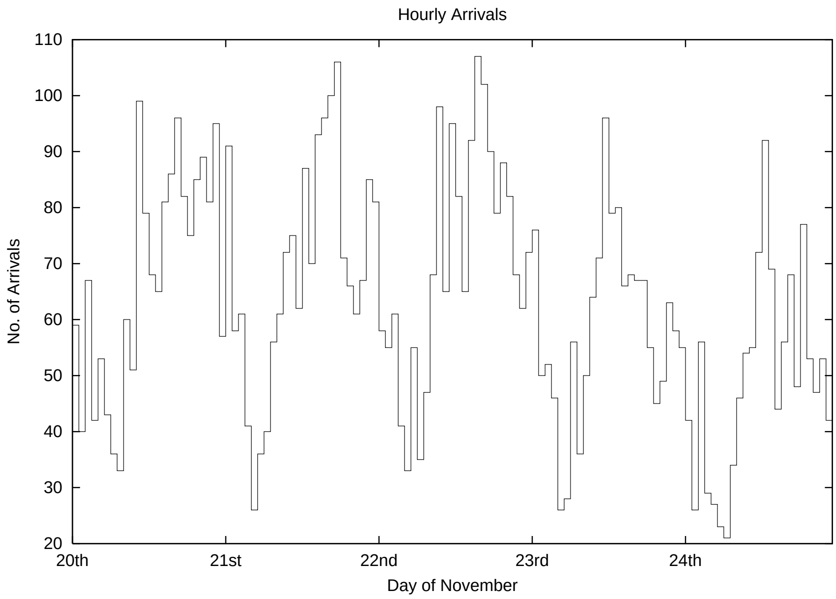

Figure 2: Daily arrival pattern for 5 days, 20-24 November2000

The hourly arrivals are averaged over the three month measurement period and plotted in Figure 3, along with averaged plots of each month's hourly arrivals. The resultant daily arrival pattern is similar for each of the months Nov-Jan, as well as for the 3 month period averaged. This confirms our expectation of a fairly unchanging day to day arrival pattern.

Figure 3 shows the arrival rate to peak between 10h00 and 13h00. The peak occurs at the time when staff and students doing vacation work arrive at university. The sharp drop in arrival rate after 13h00 is due to people going on lunch from 13h00-14h00. There is another peak between 17h00 and 18h00. The peak is caused by staff and students sending their final emails before going home at 18h00.

Surprizingly the nighttime (18h00-24h00) arrival rate is approximately the same as the daytime (10h00-18h00) rate. The arrival rate only drops below daytime levels after 24h00. This suggests high levels of nighttime activity on campus during the holidays. From 18h00 to 21h00 there is a slight decline in rate when as suggested people go home for dinner, but after 21h00 the rate picks up again. The peak midday rate is nearly reached again by midnight.

The uncommonly high arrival rates at night might be caused by the 7 hour time lag between South Africa and the USA. This time lag would mean that the USA's peak email time (10h00-18h00) would occur between 17h00-01h00 South African time, which is the time period in question. We have not investigated this theory as it is not central to our study, but the theory is supported by the high arrival rate of emails after 24h00 in the morning, which is most likely not caused by South African email as very few people are up at these hours of night.

As expected a weekly arrival pattern is observed in the data. Figure 4 shows the pattern of high arrival rates during weekdays and lower rates during weekends. The doubling of the daily arrival rate on the 13^{th}, 14^{th}, 15^{th}, 16^{th}

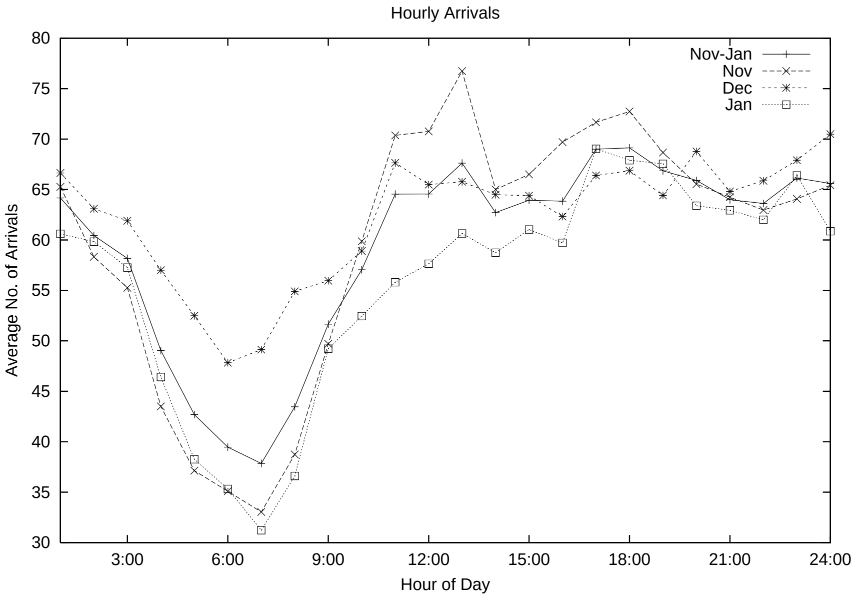

Figure 3: Average daily arrival pattern

and 17^{th} of December was in all probability caused by pre-Christmas vacation email. The drop in email activity from the 18^{th} onwards, picking up again after the 8^{th} of January, confirms this hypothesis as this period corresponds to the Christmas vacation.

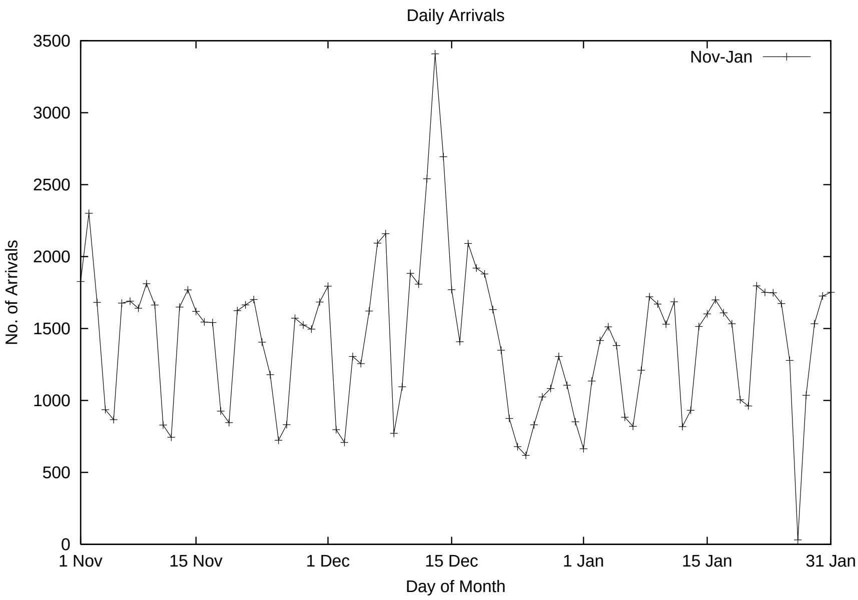

Figure 4: Weekly arrival pattern

We can conclude that the arrival rate is dependent on the day of week as well as time of day, and is therefore non-stationary.

By design a non-stationary stochastic process captures the behaviour of non-stationary traffic more accurately than a stationary stochastic process. Feldmann [Fel98] however showed that for certain modeling problems where short time-scale interactions are of importance, a stationary stochastic process can model interarrival times more accurately than a non-stationary process.

We decided to analyse the interarrival time distribution of the whole dataset as a stationary process as opposed to breaking the dataset up into non-stationary subsets and analysing these.

5 Interarrival Pattern

The summary statistics for the interarrival time data are shown in Table 1. The measurements were recorded in microseconds. There are 129 930 recorded interarrival times, measured over 92 days, the smallest element is 0.001006 seconds and the largest approximately 8 minutes. The sample mean is approximately 57 seconds.

Sample Size	129 930
Measurement Period	92 Days
Smallest Value	0.001006s
Largest Value	499.934105s
Sample Mean	57.604762s
Confidence Interval (95%)	57.230524s , 57.978999s
Sample Variance	4736.879822s
Standard Deviation	68.824994s
Coefficient of Variation	1.19
Skewness	2.37
Kurtosis	10.18
Lower Quartile	12.495278s
Median	33.227505s
Upper Quartile	292.914021s
Trimean	38.864720s

Table 1: Summary Statistics

Figure 5 plots all the interarrivals in order of occurrence from left to right.

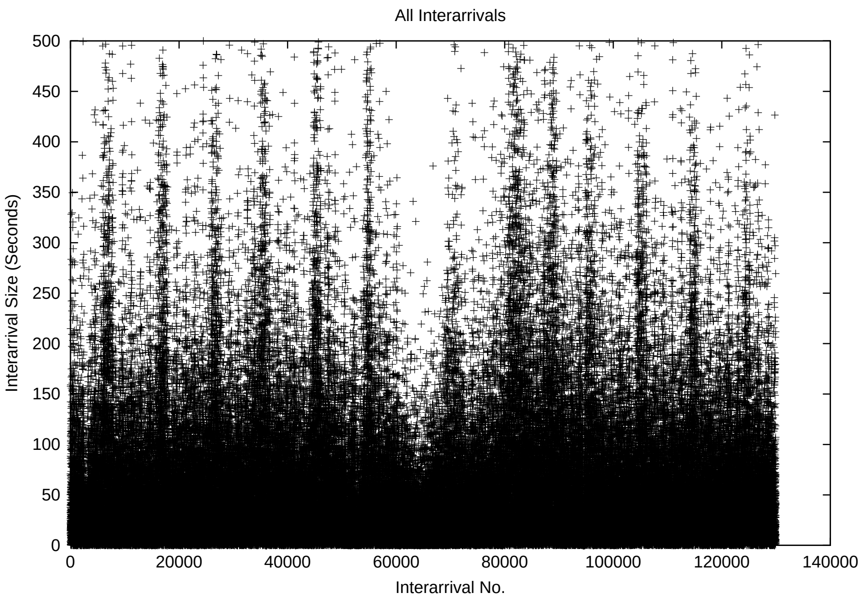

Figure 5: Interarrival Pattern of Whole Trace

The most striking observation is the weekly pattern we have observed before, where large interarrivals occur over weekends. The majority of interarrival times are smaller than 200 seconds but a non-negligible amount of interarrivals are more than twice as large. This large difference in size (approximately 200s) between the larger and smaller values suggests that the data is skewed to the right.

The *five number summary* corroborates this fact:

$$(0.010006s,\ 12.495278s,\ 33.227505s,\ 76.508591s,\ 499.934105s)$$

The *five number summary* shows the values of the smallest number, the first quartile, the median, the upper quartile and the largest number. The range between the smallest value and the median is 33.217499 seconds while the range between the median and the largest value is 466.706600 seconds.

The *skewness* value is 2.37 (positive values indicate skew to right).

The standard deviation is approximately 68 seconds which is 1.19 times larger than the mean. This indicates that the data is not exponentially distributed as the mean value is equal to the standard deviation for exponentially distributed data. The skewness as well as the large variance of 4736.879822s indicates a *heavy tailed* probability distribution.

Figure 6 shows the empirical probability density function of the data. The skewness of the data is evident when considering the very long upper tail of the distribution. The vertical lines indicate the 25^{th}, 50^{th}, 75^{th}, 95^{th} and 99^{th} percentiles of the data. The first 25% of the data is spread over the short interval between $0-12$ seconds. The last 25% of the data is spread over the very large interval between $77-500$ seconds. It can be seen that the rate of decline in the gradient of the distribution decreases between the 75^{th} and 95^{th} percentiles showing evidence of a *heavy tail.* The bins after the 95^{th} percentile contain one or two arrivals each and don't add much to the data. We will ignore this data in our analysis of the upper tail.

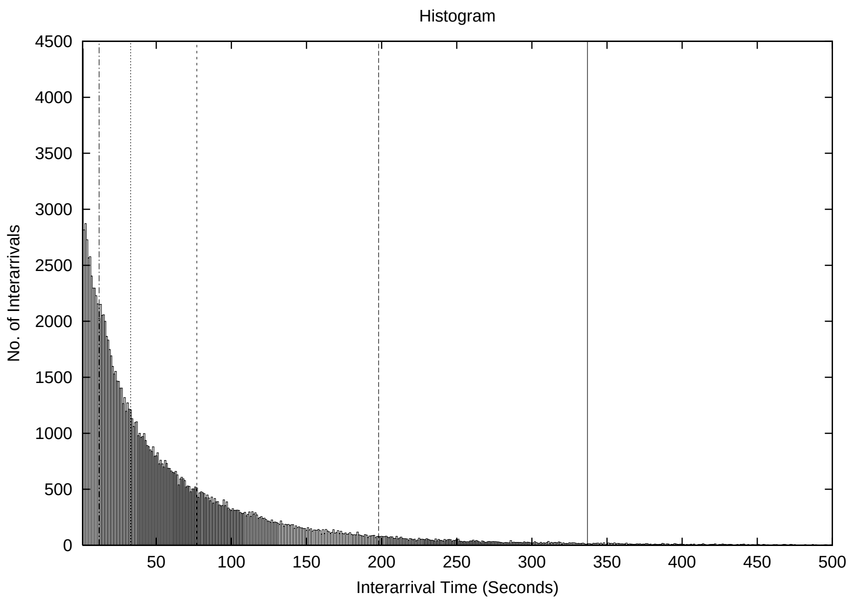

Figure 6: Empirical Probability Density Function of Whole Trace

The data is centered very closely around the mean as 50% of the interarrivals fall in the interval between $12-77$ seconds. The *kurtosis* has a value of 17.012 which corroborates this fact.

We conclude that the distribution has a high variance and is skewed to the right. Further investigation into the upper tail of the distribution is necessary. The *trimean* is a more reliable measure of location for skewed data because it takes into account outlying values but balances this by the robust qualities of the median incorporated into it. The trimean has a value of 39 and is a better reflection of the "average interarrival time" than the mean.

5.1 Upper Tail

A heavy tailed distribution is a distribution for which the *upper part* or *tail* of the distribution declines according to a power rate rather than an exponential rate. Distributions commonly used to model network characteristics such as the exponential and normal distributions have tails which decline exponentially or faster. Heavy tailed distributions therefore have tails that decline slower than these distributions which result in a greater degree of variability in the size of observations. The probability of larger observations occurring is greater for data following a heavy tailed distribution than it is for data following a non-heavy-tailed distribution. We say that a random variable X follows a heavy-tailed distribution with tail index α if

$$P[X > x] \sim cx^{-\alpha}, \ \ as\ x \to \infty, \ \ 0 < \alpha < 2, \tag{1}$$

where c is a positive constant, and where $\sim$ means the ratio of the two sides tends to 1 as $x \to \infty$ [CTB98]. This distribution has infinite variance, and if $\alpha \leq 1$ it has infinite mean. Heavy tailed behaviour can be detected in a dataset by inspecting the dataset's *complementary cumulative distribution function* (ccdf) which is defined as

$$P[X > x] = \bar{F}(x) = 1 - F(x),$$

where F(x) is the cumulative distribution function $F(x) = P[X \leq x]$ [CTB98]. Datasets which exhibit heavy tailed behaviour have the property that the tail-end of their ccdf plotted on a log-log scale is linear. This follows from Equation 1 from which we can derive

$$\lim_{x\to\infty} \frac{d \log \bar{F}(x)}{d \log x} = -\alpha$$

so that for large x the ccdf should appear to be a straight line on log-log axes with slope $-\alpha$.

Figure 7 shows the ccdf of the dataset with values above 250 seconds removed as explained in Section 5. The vertical lines indicate the 25^{th}, 50^{th}, 75^{th}, 95^{th} and 99^{th} percentiles of the data. A straight line is fitted to the distribution for abscissa values greater than 1.85 (71 - 250 seconds). This data corresponds to the upper 25% of the dataset. The gradient of the line was found to be -1.81122 and α is therefore predicted as -1.81122. This value falls within the (0,2] interval that characterizes heavy tailed behaviour. We conclude that the upper 25% of the dataset exhibits heavy tailed behaviour according to the ccdf test.

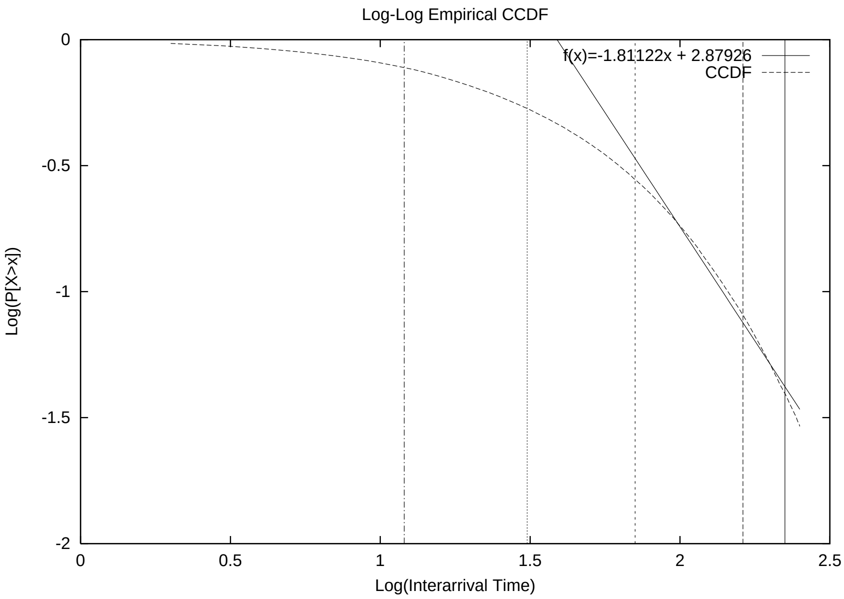

Figure 7: Complementary Cumulative Distribution Function with Straight Line Fitted

The ccdf method of estimating α can not be used on its own to decide whether a distribution has a heavy tail. We obtain two more estimates of α by means of the:

- Hill Plot [DdHR00]
- Scaling Method [CT99]

The Hill plot is a plot of the Hill estimator for a range of values of k, which is defined as the number of upper order statistics used in the calculation of the Hill estimator. The Hill estimator is defined as:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}} \tag{2}$$

where X is the order statistics $X_1, \ldots, X_n$ of the trace such that

$$X_{(1)} > X_{(2)} > \ldots > X_{(n)}$$

and k is the number of these upper order statistics used in the estimation as mentioned before [DdHR00]. The Hill Plot is therefore the plot of

$$((k, H_{k,n}^{-1}), 1 \leq k < n)$$

Figure 8 shows the Hill Plot of the dataset. The plot is interpreted by finding the value of the ordinate where the plot starts to "stabilize" by turning into a line parallel to the x-axis. The Hill Plot in Figure 8 does not stabilize, it turns away from its path parallel to the y-axis but never runs parallel to the x-axis. The fact that the plot turns away from its parallel path to the y-axis indicates that there is sub-exponential decline in the tail, but not necessarily power-rate decline i.e. heavy-tailed.

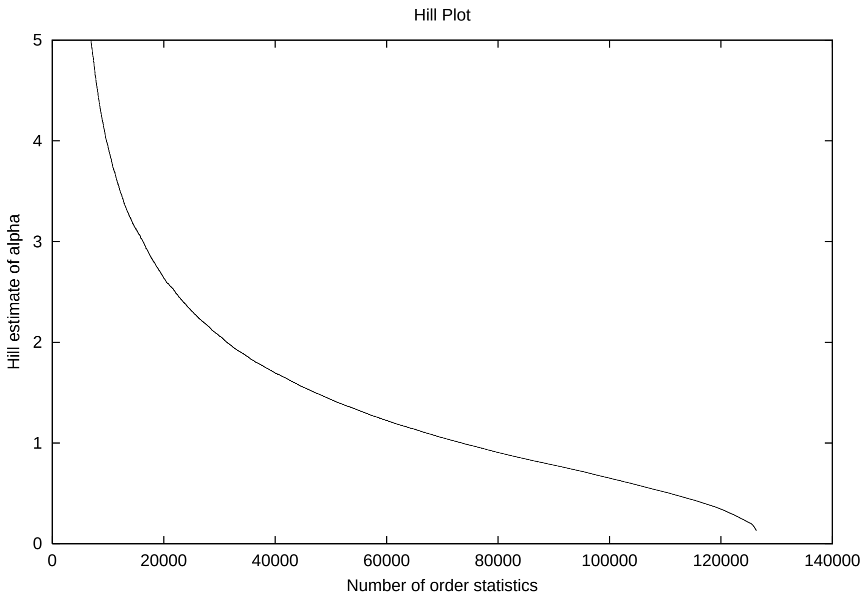

Figure 8: Hill Plot

The value for α is difficult to derive from this plot as there is no definite point where the plot "stabilizes". The most we can say is that the value is between 1 and 3 which is the range over which the plot turns away from the y-axis.

The final test we employ is the "scaling estimator" derived by Crovella and Bestavros in "Estimating the Heavy Tail Index from Scaling Properties" [CT99]. The method makes use of the fact that the tail of a heavy-tailed distribution determines the scaling properties of the dataset when aggregated i.e. the tail index is unchanged when heavy-tailed random variables are summed. This effect can be observed in empirical data. The ccdf's of a heavy-tailed dataset repeatedly aggregated and plotted on log-log axes will show the heavy tails to appear as approximately parallel lines with slopes approximately equal to α. The method also results in a single value estimation of α.

We applied this test to the data by using the *aest* program written by Crovella and Bestavros [CT99], which is publicly available. The results are plotted in Figure 9.

As can be seen the lines are parallel but the slope of the lines result in a value of 2.064 for α which do not fall in the (0,2] interval which characterises heavy-tailed distributions.

6 Conclusion

We conclude that the interarrival time distribution for the dataset is non-stationary with daily and weekly arrival patterns. The average interarrival time is 39 seconds and values range between 0.001006 and 499.934105 seconds. Fifty percent of the data fall in the range 12-77 seconds.

The distribution shows evidence of heavy-tailed behaviour. Investigation shows that the tail declines faster than a power-rate. The distribution is therefore not heavy-tailed.

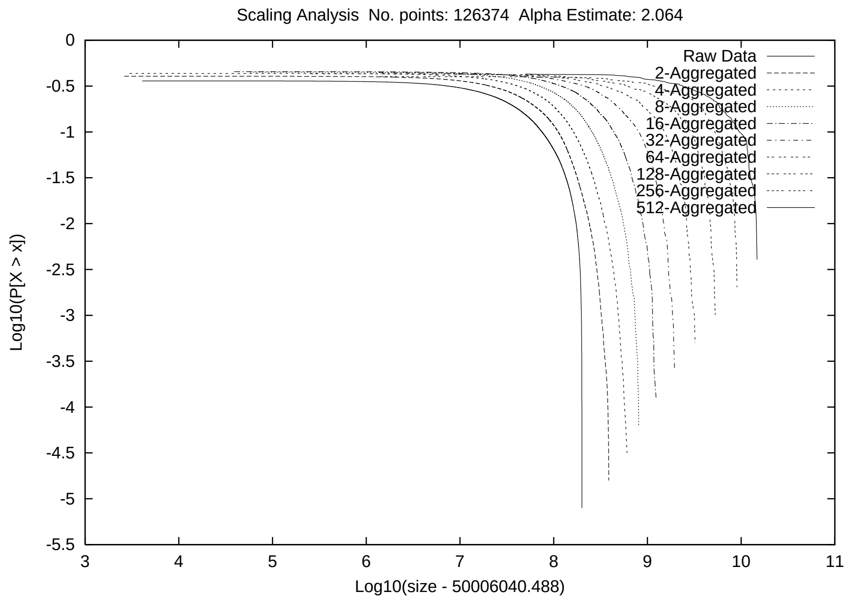

Figure 9: Scaling behaviour in dataset

References

[CT99] M.E. Crovella and M.S. Taqqu. Estimating the Heavy Tail Index from Scaling Properties. *Methodology and Computing in Applied Probability*, 1(1), 1999.

[CTB98] M. Crovella, M. Taqqu, and A. Bestavros. Heavy-Tailed Probability Distributions in the World Wide Web. In *A Practical Guide To Heavy Tails*, pages 3–26. Chapman & Hall, 1998.

[DdHR00] H. Drees, L. de Haan, and S. Resnick. How to make a Hill plot. *Annals of Statistics*, 28(25):254–274, 2000.

[Fel98] A. Feldmann. Characteristics of TCP Connection Arrivals. Technical report, AT&T Labs Research, 180 Park Av., A175, Florham Park, NJ 07932, December 1998.

[Pos82] J. Postel. RFC 821: Simple Mail Transfer Protocol (SMTP), August 1982. Internet Standard.