Journal of Web Engineering, Vol.9, No.1 (2010) 25-47
© Rinton Press

INVESTIGATING THE DISTRIBUTIONAL PROPERTY OF THE SESSION WORKLOAD

JAMES MILLER TOAN HUYNH
University of Alberta, Canada
jm@ece.ualberta.ca huynh@ece.ualberta.ca

Received September 2, 2008
Revised September 8, 2009

Companies now rely on the World Wide Web for communication with their customers. As reliance on web servers grows, the need for companies to better understand the workload placed upon these servers also increases. The session workload unit is a popular unit of measurement used to analyze recorded information from server logs. In fact, many web applications, from shopping carts to online banking systems, require session information to function correctly. Web data mining is also dependent on session workload information. However, the distributional properties of this session workload are not understood. Whether the session workload can be described as a short-tailed or heavy-tailed distribution is a fundamental question for the investigation of the session workload unit. This paper empirically explores claims that the session workload can be described using a heavy-tailed distribution. The paper concludes that, for the samples used in this paper, a method to accurately determine whether the session workload is drawn from a heavy-tailed distribution does not exist. Hence, the conclusion that they are drawn from such a distribution cannot be made.

Key words: Web session length; Session workload property; Web log analysis
Communicated by: D. Lowe & E. Mendes

1 Introduction

The World Wide Web is now the most popular component of the Internet [2]. The Web can be utilized for many purposes ranging from information retrieval to fully interactive e-commerce stores. Companies increasingly use the Web to reach their customers. As the Web's popularity increases, so does the need for companies to better understand the workloads placed upon their servers. Web mining allows companies to further understand their users' behavior and demographic information, which in turn allows the organization to maximize sales. It can also provide critical workload information, such as hits per user or session, enabling system administrators to improve usability, availability and reliability of their websites. One of the most popular units used to analyze traffic, workload and user behavior is the session workload unit. Many researchers have investigated the session workload. However, previous studies have used very short (duration) data sets and many have not considered data from commercial websites, especially mission-critical websites.

Furthermore, the investigations into the distributional properties of the session workload lack rigorous analysis. In fact, Goševa-Popstojanova et al. [24] is the only known study to provide a detailed analysis of the measure's characteristics. However, this study only considers "are sessions lengths sampled from a heavy-tailed distribution" without convincing evidence that this characterization is definitive. The implications of whether the session length is heavy-tailed can have a significant impact on the formulation of many website models. For example, Tian et al. [47] proposed a reliability model for websites based on a short-tailed distribution which would be invalid if the session length is heavy-tailed. Furthermore, let's consider constructing a simple reliability model of a website. If we assume that the probability of any software failure per input or hit is constant, *p*, we have a simple binomial process. The number of failures f_n after *n* inputs is given by the binomial distribution:

$$P(f_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (1)$$

Therefore, the probability of the system failing after n hits occurs whenever $f_n > 0$. *Hence,*

$$\begin{aligned} P(f_n > 0) &= 1 - P(f_n = 0) \\ &= 1 - (1-p)^n \end{aligned} \qquad (2)$$

The system administrator might want to think about the defect rate of the system as a function of time rather than as a function of the number of inputs or hits.

$$1 - (1-p)^n = 1 - (1-p)^{At} \qquad (3)$$

where *A* is the average inputs per time unit (*t*). Further, considering the data presented in the previous sections, *p* is obviously small and *n* is obviously large allowing our binomial process to be approximated by an exponential distribution.

$$= 1 - e^{-Atp} \qquad (4)$$

If the distributional property is heavy-tailed, this model would be invalid because the average inputs per time unit *A* is infinite. These types of models are neither new nor unique to reliability. Many dynamic characteristics of websites may be approximated by such models. However, if the workloads are heavy-tailed, many of these models will be invalid because either the mean or the variance is infinite. That is, they require an estimation of one of the moments of the workload variable; yet, the moments are infinite in heavy-tailed distributions.

The session workload unit has also been used to mine web usage for web personalization [16][37]. This personalization process allows websites to customize themselves to match the users' usage patterns. For example, Amazon.com uses web mining data from user sessions to recommend books to their customers. Jasen and Spink [32] examined user sessions to determine how web search engines are utilized and which search results are being viewed by the users. Cherkasova and Phaal [10] proposed a session-based load management for commercial websites to improve quality of service; they utilized a simulation to model the session workloads in their study. All approaches mentioned are dependent on the session workload model. Hence, the acceptance of the conjecture that workloads are sampled from heavy-tailed distribution has serious ramifications for future research and analysis of the "behavior" of websites. Therefore, this paper re-evaluates the results presented by Goševa-Popstojanova et al. [24] which concludes that session length data is sampled from a heavy-tailed distribution. The conclusion was based on results from the analysis of the log-log complementary distribution plots (LLCD) and the Hill estimator. However, we believe a more rigorous empirical investigation into session length and its potential distributional properties can be performed.

This paper extends Goševa-Popstojanova et al. [24] by applying the evaluation to two new websites. One of which is a mission-critical commercial website. The logs investigated for this commercial website cover a 27 month period, an extensive time period. Other investigations are "focused" on high throughput web sites for a short period. However, the authors believe that examining a website over a long calendar period is essential as many "external actions" which impact the characteristics of the site happen infrequently as hence a true sense of the historical norm of a website's characteristics is only available over an extended period. Furthermore, additional tests, such as the Heavy-tailed Autocorrelation Function (ACF) method, "wobble analysis" and Q-Q plots (Q stands for quantile), are performed to determine if session length can really be modeled by a heavy-tailed distribution. The results from this paper show that, for the samples used in this paper, a method to determine whether the session workload can be modeled by a heavy-tailed distribution does not exist.

The remaining sections of this paper are organized as follows: Section 2 describes the data analysis step and provides a brief description of the websites under investigation. Section 3 re-evaluates the heavy-tailed property of session lengths. It investigates the validity of using log-log complementary distribution (LLCD) plots and the Pareto distribution to model the session length as presented by Goševa-Popstojanova et al. [24]. Section 4 discusses the results from this study versus the previous study. Finally, Section 5 presents our conclusions.

2 Data Analysis

2.1. Description of the Websites under Investigation

Server logs from two websites are investigated in this paper. The first website (Site A) is a website for a company that specializes in online databases. This is a commercial website that is critical to Company A’s operation. The website charges customers for the time used to access its online database; hence an outage means that revenue will be lost. This website represents one of the core revenues streams for this organization. The PHP (http://www.php.net) scripting language, MySQL (http://www.mysql.com) database and Apache HTTP Daemon are technologies used by the website. 27 months of operation from December 2004 to February 2007 are examined in order to observe potential trends and patterns for this mission critical website. The website is dynamic – the pages are generated dynamically depending on the customers’ inputs; its users are customers who are either looking to purchase a product or to register for a training course. For the 27 months covered, Site A received approximately 3.6 million hits and 117,246 “unique” visitors. The site transferred 67 Gbytes of data. It is believed that this log represents the longest period of capture and the only truly “mission critical” log reported within the research literature.

The second website is the website for the Department of Electrical and Computer Engineering at the University of Alberta. Although the site is important to the organization, it is non-commercial and not mission critical. This website is also dynamic and utilizes the ColdFusion[a] scripting language and the Apache HTTP Daemon (http://httpd.apache.org). To investigate the data, the log files were chosen to cover 11 months of data. For this period, the ECE website received approximately 2.42 million hits, 203,896 “unique” visitors and transferred a total amount of 22.6 Gbytes of data. The data from this website is served as a cross reference to ensure that the trends observed are not unique to one particular website.

The log files are stored in the Combined Log Format[b] for Site A and the Common Log Format (CLF)[c] for ECE. The approach used to extract the data can be seen as a deep log analysis technique [35][36][37]. Since the session length estimation requires at least two requests: one to mark start time of the session and one to mark the end time of the session, all users with only one request are removed from the log files.

2.2 Comparison of the log data versus previous studies

Table 2.1 provides a summary of the data used in previous studies and this study. Websites with an asterisk(*) are commercial websites.

Although the websites examined by previous studies have higher traffic intensity, the periods covered are shorter. In fact, this table shows that the longest period this study examined is 27 months, whereas the longest period previous studies have performed is 7 months. Furthermore, the periods covered for commercial websites are extremely short, 1 to 2 weeks. This study

[a] http://www.macromedia.com/software/coldfusion

[b] http://httpd.apache.org/docs/1.3/logs.html#combined

[c] http://httpd.apache.org/docs/1.3/logs.html#common

investigates the log file from a commercial website for a much longer period (27 months). This long data period provides several benefits over short data periods.

Table 2.1. Overview of the log data used in previous studies and this study

		Log duration	Requests	Bytes Transferred
Goševa-Popstojanova et al. [24]	NASA-Pvt1	20 weeks	23 thousands	0.5 GB
	NASA-Pvt2	20 weeks	92 thousands	0.2 GB
	NASA-Pvt3	20 weeks	489 thousands	2.2 GB
	NASA-Pub1	20 weeks	93 thousands	9 GB
	NASA-Pub2	20 weeks	732 thousands	6.7 GB
	NASA-Pub3	20 weeks	108 thousands	4.6 GB
	CSEE	6 weeks	5.8 millions	80.9 GB
	WVU	3 weeks	37.9 millions	97 GB
	ClarkNet*	2 weeks	3.3 millions	27.6 GB
	NASA-KSC	2 months	3.5 millions	62.5 GB
	Saskatchewan	7 months	2.4 millions	12.3 GB
Goševa-Popstojanova et al. [25]	WVU	1 week	15.8 millions	34.5 GB
	ClarkNet*	1 week	1.7 millions	13.8 GB
	CSEE	1 week	397 thousands	10.1 GB
	NASA-Pub2	1 week	39 thousands	0.3 GB
This study	Site A*	27 months	3.6 millions	67 GB
	ECE	11 months	2.4 millions	22.6 GB

- An organization's behavior also affects its website traffic patterns. Advertising campaigns, various public announcements will often increase the amount of traffic. For example, GoDaddy.com's website experienced a 1500 percent increase in traffic following its Super Bowl ad campaign[d]. Other websites advertised during Super Bowl Sunday also had their traffics increased. Short term collection either overstates these actions if it is performed near a major activity or understates them if performed far from the activity.
- Well known trends and periodic patterns such as the "weekend effect" will distort short term collection resulting in skewed data. In fact, Arlitt and Jin [1] have demonstrated that websites have very different workload intensities on weekdays versus weekends. Although "weekend effects" are short, seasonal effects such as holiday seasons can last much longer. Hence, if the data period is short, the analysis will be skewed by such effects.
- Major web events will also affect the data sets gathered within a short time frame. For example, popular YouTube videos are known to result in millions of hits to YouTube's website within a short period of time before the site's traffic returns to normal. A website being mentioned on another popular website such as Slashdot will also cause the website's traffic to increase. This is commonly known as the Slashdot-effect[e].
- Short collection periods can experience distortion due to either higher than normal or lower than normal activities from robots. For example, attackers directed thousands of bots to access Facebook; and hence, they created a denial of service attack[f]. The number of requests made is large enough to deny normal users from accessing the website. Although the attack only lasted several hours, the data generated from the attack will have an effect on the short term data due to the large data size.
- Users have very low brand loyalty. If quality of service (such as response period) is poor, users leave quicker than normal (the inverse will be at some-level true) – this impacts session statistics and again short-collection periods can get skewed because of the quality of service differing from the long-term norm. For example, a user may visit a website during maintenance which may cause the website to response much slower than usual. The quality

[d] http://ir.comscore.com/releasedetail.cfm?releaseid=245204, last visited September 2, 2009

[e] http://hup.hu/old/stuff/slashdotted/SlashDotEffect.html, last visited September 2, 2009

[f] http://news.cnet.com/8301-27080_3-10305200-245.html, last visited September 2, 2009

of service during this maintenance period cannot be considered as the normal QoS for the website.

Table 2.2. Traffic by month for Site A

Month	Requests	Bytes Transferred
2004-12	99 thousands	1.77 GB
2005-1	122 thousands	2.30 GB
2005-2	110 thousands	2.05 GB
2005-3	137 thousands	2.59 GB
2005-4	120 thousands	2.34 GB
2005-5	115 thousands	2.21 GB
2005-6	123 thousands	2.42 GB
2005-7	108 thousands	2.16 GB
2005-8	116 thousands	2.17 GB
2005-9	114 thousands	2.08 GB
2005-10	120 thousands	2.27 GB
2005-11	125 thousands	2.28 GB
2005-12	110 thousands	2.14 GB
2006-1	152 thousands	2.86 GB
2006-2	137 thousands	2.78 GB
2006-3	164 thousands	3.30 GB
2006-4	138 thousands	2.67 GB
2006-5	148 thousands	2.72 GB
2006-6	134 thousands	2.41 GB
2006-7	128 thousands	2.48 GB
2006-8	144 thousands	2.69 GB
2006-9	149 thousands	2.52 GB
2006-10	146 thousands	2.76 GB
2006-11	144 thousands	2.81 GB
2006-12	115 thousands	2.28 GB
2007-1	171 thousands	3.05 GB
2007-2	181 thousands	3.20 GB

Analysis of the Site A data set shows that data traffic does change. Table 2.2 shows the monthly traffic for the 27 months examined. This table shows an increase in traffic starting in January 2006. This was when an advertising campaign was launched for Site A. In fact, the average number of requests is increased from 118,453 in 2005 to 141,668 in 2006. In January 2007, another advertising campaign was launched and the increase in traffic can be seen again. The seasonal effect is also quite evident here. Traffic in December (2004, 2005, 2006) are lower than normal traffic for the other months.

3 Investigation of the Distributional Characteristics of Session Length

Goševa-Popstojanova et al. [24][25] put forward the conjuncture that session length data is sampled from a heavy-tailed distribution. In this section we empirically examine this conjecture.

3.1. Discussion of the STT

This study uses a Session Timeout Threshold to determine the sessions. A session is defined as a sequence of actions taken by a user within a period of time. Sessions offer much finer grained information than the standard *number of users* metric. However, because the Hyper Text Transfer Protocol (HTTP) is a stateless protocol, session information cannot be easily captured. Hence, web applications often use session-based technology such as cookies [33] to simulate a stateful connection to the user. In order to determine when a session ends and the next one begins, a session timeout threshold (STT) is often used. In other words, a STT is a pre-defined period of inactivity that allows web applications to determine when a new session occurs. That is, let s be a set of sessions:

$$\forall s \in \text{SessionsFor(user)} \bullet (\text{session\_time\_start}(s_{i+1}) - \text{session\_time\_end}(s_i)) \geq \text{STT}$$

Goševa-Popstojanova et al. [24][25] assign STT to 30 minutes, because it is a common value used by other researchers [6][35][36][46]. This 30 minute figure is a value rounded up based on a mean value of 25.5 minutes determined by Catledge and Pitkow [8]. Catledge and Pitkow [8] estimate STT to be 25.5 by claiming that the most statistically significant events occurred within 1.5 standard deviations (25.5 minutes) from the mean between each user interface event which was 9.3 minutes. However, no definition of these "significant events" was given; and why 1.5 standard deviations is selected is never discussed. Hence, this paper also uses a model proposed by Huynh and Miller [30] to determine the STT. By applying the model, the STT is found to be 5 minutes for Site A and 11 minutes for ECE. As a cross-check, the results presented in this paper were replicated using STT = 30 minutes for both sites; and while the numerical values clearly changed the basic interpretation of the results remained constant.

3.2. Estimating the Tail Index α with LLCD plot

Under the assumption that the data comes from a Pareto distribution, Goševa-Popstojanova et al. [24][25] estimate the tail index of the distribution using a log-log complementary distribution (LLCD) plot. This approach has also been used in many studies which concentrate on other workload metrics for web servers [1][2][11]. LLCD plots produce an estimate of the tail index using the property

$$\frac{\partial \log(P[X > x])}{\partial \log x} = -\alpha \tag{7}$$

However, the approach does not utilize the entire distribution. The estimation of the index is only over the range $[x_i, x_{i+j}]$; and the approach simply fits an ordinary least-squares linear regression model to estimate α from the small set of values ($[x_i, x_{i+j}]$) which are assumed to represent the majority of the tail.

Downey [13][15] has shown that the LLCD plot is an ineffective mechanism at discovering long-tailed distributions. Basically, the technique cannot adequately distinguish between long-tailed distributions, such as the Pareto distribution, and "similar looking" short-tailed distributions such as lognormal distributions. Figueiredo et al. [18] further support this viewpoint and provide an extensive analysis demonstrating the inadequacy of this approach; they demonstrate that the discovery of the appearance of a linear region in a LLCD plot is by itself insufficient evidence to conclude that long-range dependence exists within a data set. Finally, Goldstein et al. [21] empirically demonstrate that the LLCD plot and associated techniques are ineffective approaches to fitting power-law distributions to experimental data and conclude that the approach should be avoided.

3.2.1. Discussions of the LLCD Plot Results

This paper uses three definitions of the tail as presented by Hernandez-Campos et al. [27]. The *extreme tail* is the part of the tail that is beyond the last data point (x_n), hence no information is available for this part. The *far tail* is the part of the tail where some data is present, but the distributional properties cannot be understood because of the minimal information available (around x_{i+j}). The *moderate tail* is the part of the tail that contains "rich" (by comparison) distributional information ($[x_i,x_{i+j}]$). Clearly, the definitions are heuristics because the boundary between the *moderate tail* and the *far tail* cannot be defined accurately. However, the definitions are required for discussions of the results in this section.

Goševa-Popstojanova et al. [24][25] have estimated α using LLCD plots. Figures 3.1, 3.4, 3.7 and 3.10 display the LLCD plots for ECE and Site A with each having two different STT values. This paper utilized Huynh and Miller [30] dynamic STT estimation model and the commonly used 30 minute constant STT value approach used by Goševa-Popstojanova et al. [24][25] to investigate if the value of STT was a covariant of the distributional characteristics of the session length. Hence, LLCD plots were created for both the dynamic model's STT values and the constant STT value. These figures show that for values below –1 on the vertical axis the distribution is generally linear until the far tail is reached. Although, linear least squares fitting can be applied to estimate α, this paper uses a numerical differential equation to estimate α at all possible data points. Figures 3.2, 3.5, 3.8 and 3.11 show results of this estimation. These figures show that α does not stabilize in the moderate tail in any of the LLCD plots. The variations are consistently too large to be explained by numerical differential estimation error. To further confirm this observation, the box plot for α for all LLCD plots are shown in Figures 3.3, 3.6, 3.9, and 3.12, and the descriptive statistics for α are shown in Table 3.1. Box plots are used in this paper for their ability to visually display different types of populations without any dependency on the statistical distribution of the data. These figures show that the range for the non-outliers varies considerably; furthermore, the outliers are numerous. Figures 3.2 and 3.8 perhaps provide the clearest evidence of α failing to stabilize within the tail of the distribution. These figures can be approximated as:

1. estimates for α are relatively "well-behaved" in the *pre-tail*;
2. estimates for α vary wildly in the *moderate tail*; and
3. estimates for α seem to be almost random values in the *far tail*.

Because the type of distribution for the data sets is unknown, Table 3.1 displays the statistics for both parametric and non-parametric distributions. This table can be seen as an exploratory tool to aid the data examination process. The table and box plots further confirm that α is not stable enough for the least-squares linear regression model.

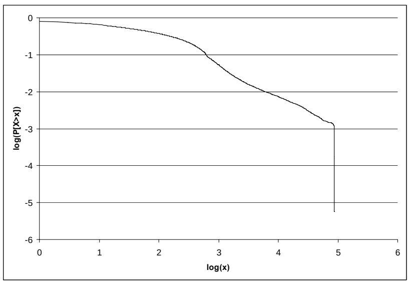

Figure 3.1. ECE 11m STT – LLCD

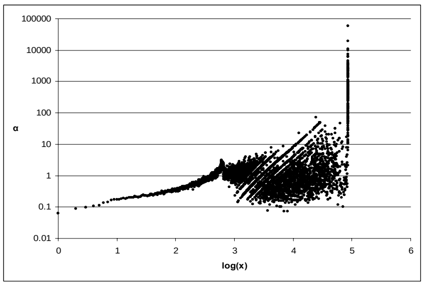

Figure 3.2. ECE 11m STT – α

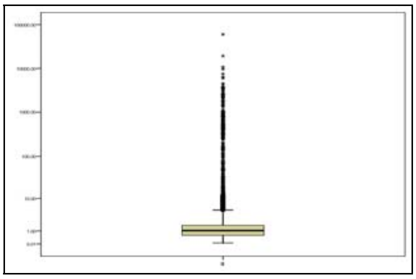

Figure 3.3. ECE 11m STT- Box plot of α

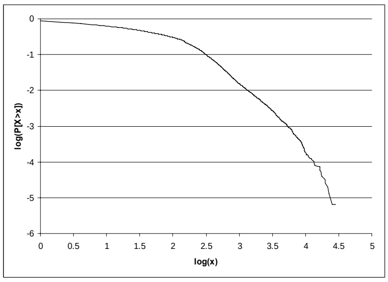

Figure 3.4. Site A 5m STT – LLCD

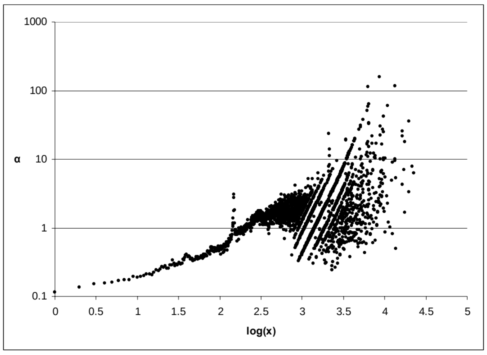

Figure 3.5. Site A 5m STT – α

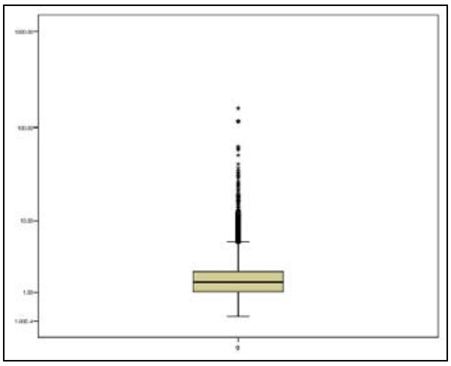

Figure 3.6. Site A 5m STT – Box plot for α

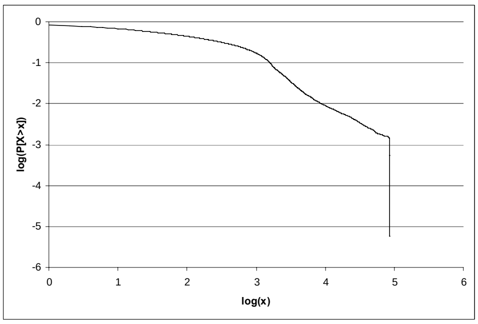

Figure 3.7. ECE 30m STT – LLCD

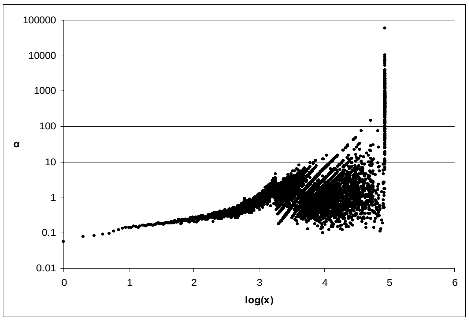

Figure 3.8. ECE 30m STT – α

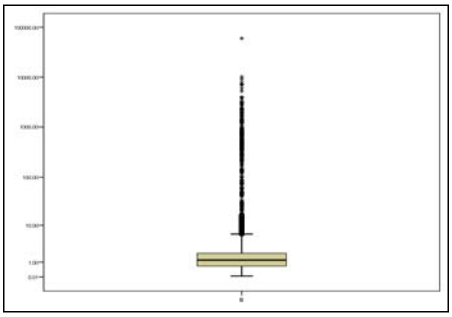

Figure 3.9. ECE 30m STT – Box plot for α

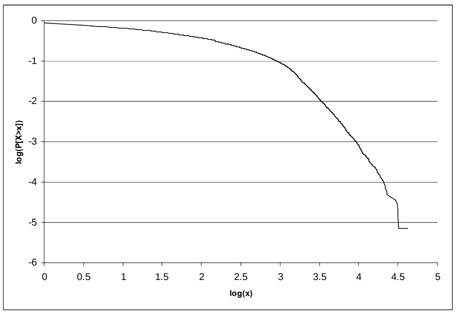

Figure 3.10. Site A 30m STT – LLCD

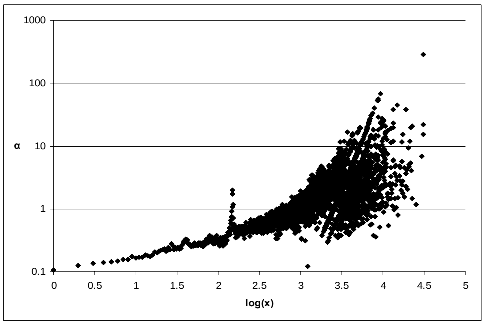

Figure 3.11. Site A 30m STT – α

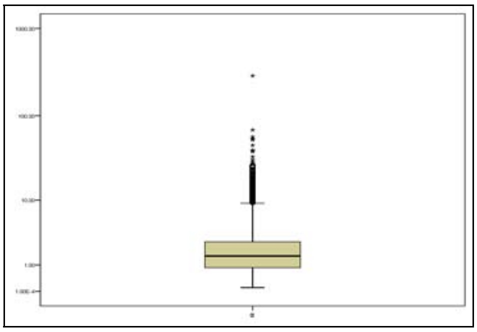

Figure 3.12. Site A 30m STT – Box plot for α

Table 3.1. Statistics for α

	ECE 11m	Site A 5m	ECE 30m	Site A 30m
Mean	49.87	2.49	35.92	2.50
Mean 95% Confidence (Lower bound)	24.60	2.27	18.43	2.33
Mean 95% Confidence (Upper bound)	75.14	2.71	53.41	2.67
Median	1.04	1.54	1.20	1.53
Variance	867872	33.61	578953	31.95
St.Dev	931.60	5.80	760.89	5.65
Minimum	0.06	0.12	0.06	0.10
Maximum	59851.53	159.08	59851.53	287.50

The figures and table show that data obtained from the proposed dynamic STT model behave similarly to the data obtained from the commonly used 30 minute STT. Hence, further data analysis methods in this paper will only explicitly examine the data set generated from the dynamic STT model as it is believed to represent the state of the art in estimating STT.

3.2.2 "Wobbles" in the Distribution

During the investigation of the session length per month plots, two interesting observations can be seen.

1. The distributions, at this level of granularity, appear to be stable. Hence, the observable phenomenon seems to repeat in a deterministic fashion.
2. The distributions are not smooth; they include several points of inflection or "wobbles". While it might initially seem reasonable to dismiss these "wobbles" as noise, the fact that they repeat across most of the monthly patterns argues that they are more likely to be signal than noise. This "wobbling" effect has been noted by several other authors investigating related phenomenon [14][34][43].

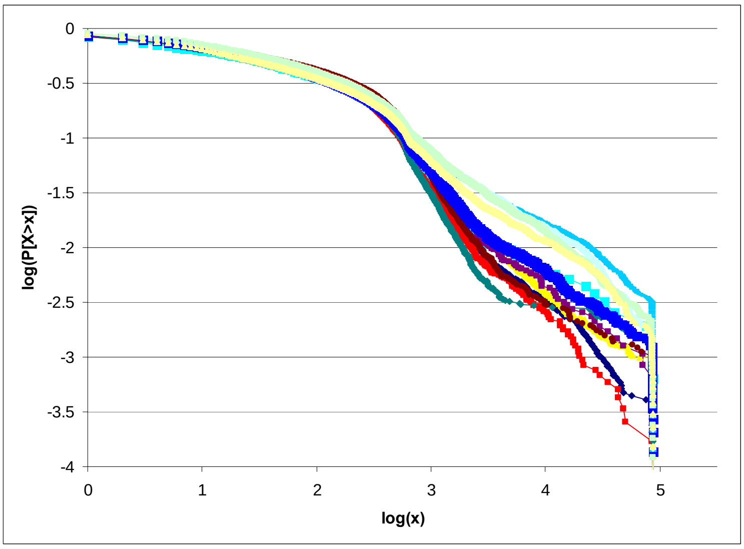

Figure 3.13. Session length by month for ECE

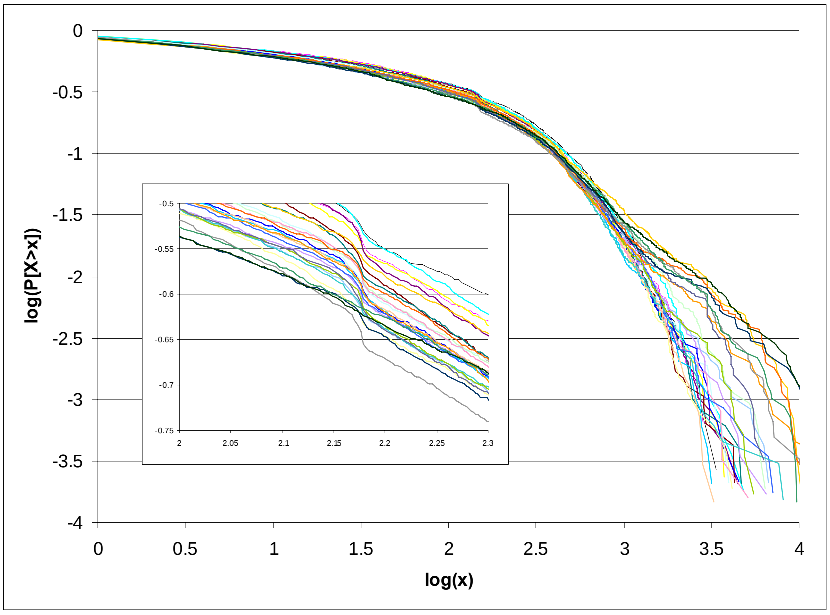

Figure 3.14. Session length by month for Site A

Figures 3.13 and 3.14 display the session length by month graphs for ECE and Site A. For the ECE site, the points of inflection can be seen at approximately 2.6, 3.4 and 3.3. For Site A, two points of inflection happen in quick succession, as can be seen by the smaller graph in Figure 3.13. This figure shows the points of inflection occur at approximately 2.16, 2.18 and 2.5.

This “wobbling” phenomenon argues simple distributions such as Pareto or lognormal distributions cannot be used to model the session workload. Hence, attempting to fit the session length into the Pareto distribution will lead to the wrong conclusion. Various researchers have investigated a range of more complex models to fit this phenomenon:

- Arlitt et al. [1][3], Barford et al. [4][5] and Downey [13][14] have all investigated hybrid models that combine a lognormal distribution with a Pareto tail;
- Mitzenmacher [34] investigate, amongst others, a double Pareto distribution;
- Reed and Jorgensen [43] investigate a double Pareto-lognormal distribution.

While all of these models can provide a superior fit to the “wobbling” phenomenon, there exists no real causal theory that they are an accurate model of the general phenomenon. The alternative argument is that the superior fit is simply the consequence of the greater number of free variables they possess compared to the simpler distributions.

3.2.3. Discussions of the Pareto Distribution

Previous studies have demonstrated LLCD plots are not effective at discovering heavy-tailed distributions because of the similarity between the Pareto distribution and the lognormal distribution. Hence, this paper will perform an investigation to determine the Pareto distribution’s effectiveness at describing the data. Downey [13][15] and Goševa-Popstojanova et al. [25] applied the curvature test to explore Pareto and lognormal distribution with conclusions stating that the data can be either Pareto or lognormal. Goševa-Popstojanova et al. [25] provides an explanation that the similarity is due to the lack of data at the *far tail*. However, as discussed, the *far tail* of a heavy-tailed distribution will never contain enough data points for any reasonable analysis. Hence, this paper will utilize the Q-Q plot [19] to visually observe the Pareto and lognormal distributions. This is the same approach used by Hernandez-Campos [27] to investigate Pareto and lognormal distributions. The Q-Q plot allows the quantiles of the data set to be graphically compared against the theoretical distribution (Pareto and lognormal for this investigation). The horizontal axis of the Q-Q plot contains the theoretical quantiles while the vertical axis contains the sorted data values. The natural log-log scale is used because of the possible large values. The curve generated should follow the 45 degree line if the data quantiles are the same (or very similar) to the theoretical quantile.

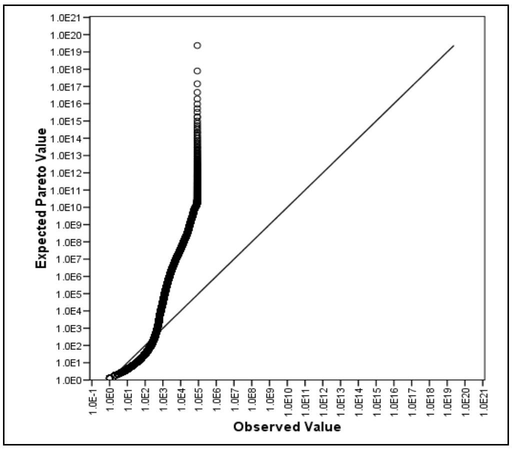

Figure 3.15. Pareto Q-Q Plot for ECE showing the observed values are not near the expected values

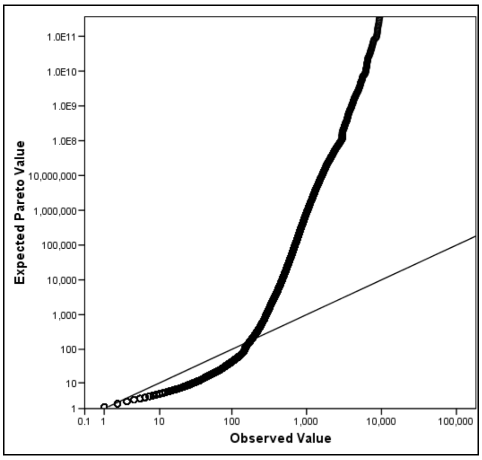

Figure 3.16. Pareto Q-Q Plot for Site A showing the observed values are not near the expected values

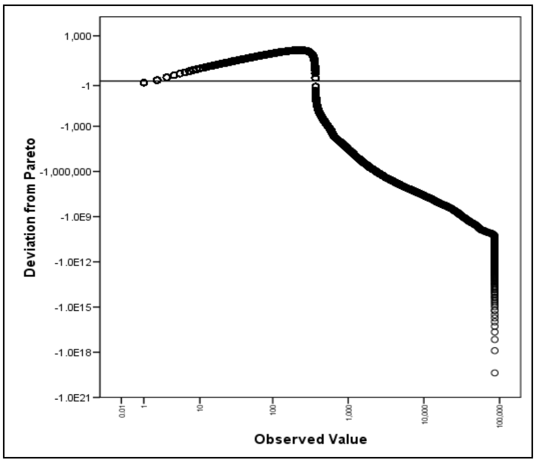

Figure 3.17. Detrended Pareto for ECE showing extreme deviations from the line in the Q-Q plot

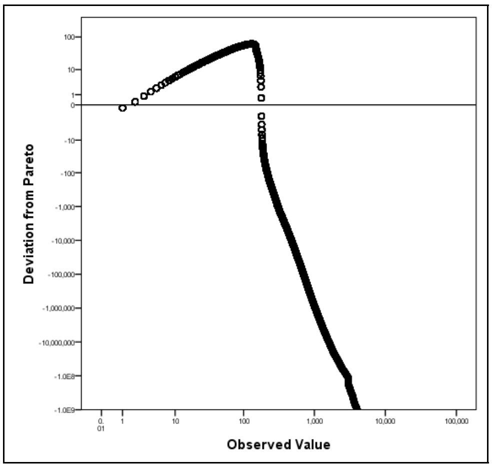

Figure 3.18. Detrended Pareto for Site A showing observed values are not near the expected values

Figures 3.15 and 3.16 show the Pareto Q-Q plots for the ECE and Site A sites respectively. Visual observation of these figures shows that the Pareto distribution does not fit extremely well to the data set as the curve does not accurately match the 45 degree line. Further confirmation of this observation can be seen with the detrended Pareto graphs as shown in Figures 3.17 – 3.18. If the plot generated by the detrended graph is not near 0 on the x-axis, then the data set is unlikely to be a good match for the distribution. Once again, these figures show that the observed values deviate from the Pareto distribution very quickly.

The lognormal Q-Q plots and detrended plots for the ECE and Site A sites can be seen in Figures 3.19 - 3.22. These figures show that the lognormal distribution also does not describe the distribution of the data accurately.

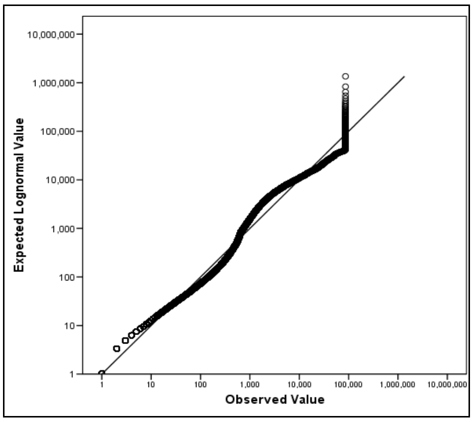

Figure 3.19. Lognormal Q-Q for ECE showing the observed values are not near the expected values

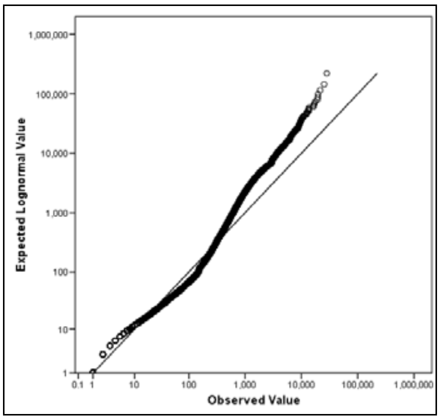

Figure 3.20. Lognormal Q-Q for Site A showing the observed values are not near the expected values

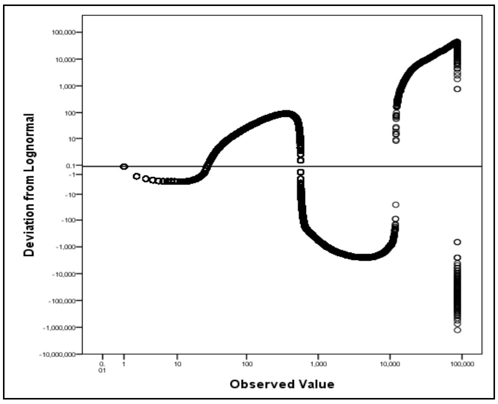

Figure 3.21. Detrended lognormal for ECE showing extreme deviations from the line in the Q-Q plot

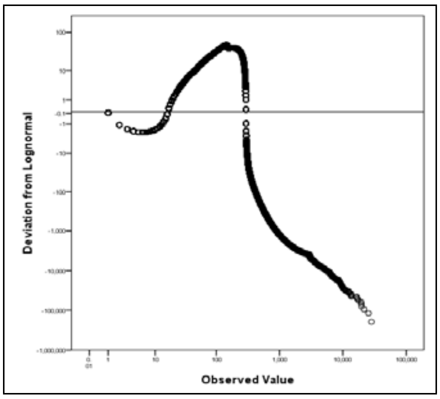

Figure 3.22. Detrended lognormal for Site A showing extreme deviations from line in Q-Q plot

To further examine the deviations in the detrended graphs between the lognormal and Pareto distributions, a statistical significance test (t-test) was conducted. The results are presented in Table 3.2. Based on the results, the null hypothesis that the Pareto distribution has a smaller mean (closer to 0) can be rejected. Hence, the alternative hypothesis, which is the mean for the lognormal distribution is “closer” to 0 than the Pareto distribution, can be accepted.

Table 3.2. t-Test to compare the lognormal distribution versus the Pareto distribution

	Site A		ECE	
	lognormal	Pareto	lognormal	Pareto
Mean	95795.9	-3.5x10^{8}	-4299959.0	-3.5x10^{20}
t Stat	9.3		9.9	
t Critical	2.0		2.0	
$P(T \leq t)$	5.0x10^{-11}		4.5x10^{-12}	

However, one can argue that the Pareto distribution is only applied to the tail of the distribution, making a formal analysis difficult to generate due to the lack of definition of the range of the tail. Hence, we fail to find any evidence using this approach that a Pareto distribution is superior to a lognormal distribution in terms of fitting the underlying data. This observation is consistent with the findings of Downey [13][15] and Goševa-Popstojanova et al. [25].

3.3. Estimating the Tail Index α with the Hill Estimator

Goševa-Popstojanova et al. [24][25] present a second alternative mechanism for estimating the tail index, the Hill estimator [28]. Again, the basic idea behind the Hill estimator is to assume that a part of the distribution is a Pareto distribution and to search for it. That is, the algorithm searches for the range $[x_i, x_{i+j}]$ once again. The algorithm uses the estimator in conjunction with a plot to search for this range. The Hill estimator uses the *k* upper-order statistics from $x_1 \geq \ldots \geq x_n$ and is defined as:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log x_i - \log x_{k+1} \qquad (8)$$

Using this estimator, $H_{k,n}$ is plotted for all *k*:

$$1 \leq k < n$$

At this point, as Resnick [44] states, one would "hope that the graph looks stable so you can pick out a value of α." Resnick [44] further states that "Sometimes this works beautifully but sometimes there are problems and it pays to be on good terms with a higher power." Resnick [44] provides a detailed discussion of the problems and issues that are encountered when using this approach; and finishes with the following brief summary of the difficulties:

"….

> 1. How do you get a point estimate from a graph? What value of *k* do you use?
> 2. The graph may exhibit considerable volatility and/or the true answer may be hidden in the graph.
> 3. The Hill estimate has optimal properties only when the underlying distribution is close to Pareto. If the distribution is far from Pareto, there may be outrageous bias even for sample sizes such as 1,000,000.

…."

Although it is not clear why this approach would be chosen, it is far from clear that this approach is inferior to the other alternatives. Tsourti and Parnaretos [48] and Rezaul and Grout [45] both undertake empirical comparisons, by simulation, of semi-parametric[g] (including Hill's) estimators for estimating tail indexes in heavy-tailed distributions. While they provide results which are not

[g] Semi-parametric estimators only estimate a single parameter at a time. Hence, here, they are used to estimate the tail index while ignoring the location parameter.

completely compatible, both papers imply that no single optimal approach exists; and that the “best estimator” tends to be dependent on the actual type of heavy-tailed distribution under investigation and the sample size. However, both papers imply that the Hill estimator performs well when the distribution is Pareto in nature and the tail index is moderate. In fact, Rezaul and Grout [45] recommend it as the “estimator of choice” when $0 \le \alpha \le 2$ and Tsourti and Parnaretos [48] when $0 \le \alpha \le 1$.

3.3.1. Discussions of the Hill Estimator Results

Using the technique discussed, which is also utilized by Goševa-Popstojanova et al. [24][25], the Hill plots for *k* was created for both sites. Goševa-Popstojanova et al. [24][25] used 10% and 14% of the upper tail in their Hill plot because *k* appears to settle to a constant value after those points. However, the Hill plots in this paper will be displayed across the entire tail to better display the stability of *k*. The Hill estimator can only be performed on the tail of the distribution. Hence, the tail was estimated using the method Goševa-Popstojanova et al. [24][25] proposed – even though Section 3.2 shows that this approach is not accurate. In order to examine the Hill plot’s behavior , a smaller range (0-5) is used for the y-axis as shown in Figures 3.25 – 3.26. These figures show that again α does not stabilize in any part of the graph. In fact, it decreases as the *k* value is increased. There does not appear to be a cut-off point as stated by Goševa-Popstojanova et al. [24][25]. The Hill plot results further confirms that the heavy-tailed property of the session length may not be an accurate model over the web sites under investigation.

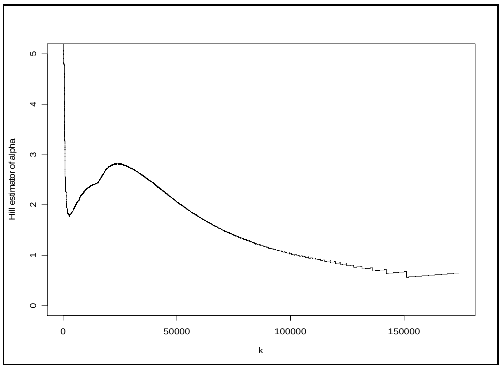

Figure 3.25. Hill estimator for ECE at a smaller range for the y-axis

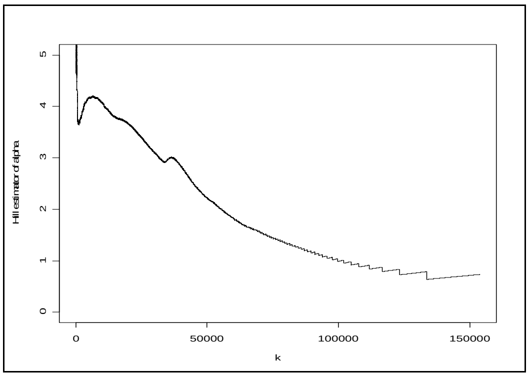

Figure 3.26. Hill estimator for Site A at a smaller range for the y-axis

4 Results Discussion

The results from this study show that the session length data may not fit a heavy-tailed distribution. The findings do not confirm the results discovered by Goševa-Popstojanova et al. [24][25]. However, it should be noted that the websites used in this study have different properties than the websites used in the previous study. Table 2.1 shows that the durations of the log files used in Goševa-Popstojanova et al. [24][25] are short. This study performs the investigation over a much longer period of time. Furthermore, although Goševa-Popstojanova et al. [24] and [25] examined a commercial website, the duration is also very short (2 weeks and 1 week), whereas this study examined the commercial website for a 27 month period.

Besides the difference in the duration of the log files, the traffic intensity between the websites in this study and Goševa-Popstojanova et al. [24][25] are also vastly different. The websites investigated by Goševa-Popstojanova et al. [24][25] have a much heavier traffic load than this study. The busiest website for Goševa-Popstojanova et al. [24] received 37.9 millions hits and transferred 97 GB of data during a 3 week period. Goševa-Popstojanova et al. [25]'s busiest website received 15.8 million hits and transferred 34.5 GB of data. This study's busiest website, which is ECE, received approximately 2.4 million hits and transferred 22.6 GB of data. The difference in traffic intensity is another possible cause for the different results obtained in this study.

5 Conclusions

This study examines claims that session length data are sampled from a heavy-tailed distribution. The dependency of the data, the LLCD plot of the data, a QQ plot comparison of the performance of the Pareto distribution against the lognormal distribution in fitting the data, and a Hill estimator approach to estimating the tail index of the distribution are all examined in detail. The investigation shows that the data may be dependent; however, the results are disputable because the formulation cannot be extended to cover all possible cases. Furthermore, this study confirms that LLCD plots may not be ideal for investigating the heavy-tailed property of session data. The α obtained from the LLCD plot does not stabilize during any part of the tail. Additionally, the Pareto distribution itself is not sufficient for modeling heavy-tailed data because of the "wobble" effect as demonstrated. The Hill estimator was examined and was shown that it also does not provide a stable α value. In fact, α does not stabilize for any k. Finally, the QQ plot suggests that the lognormal is a "better" description of the entire distribution, although we cannot rule out that a heavy-tailed distribution may be an adequate distribution of the tail of the distribution due to the imprecise definition of the tail.

Although the investigation in this study provides empirical evidence that the session data may not be heavy-tailed, the results can be disputed. The methods utilized, while popular and well known are not entirely accurate. However, no better alternatives exist; until accurate alternative approaches are presented, the heavy-tailed status of the session data is unknown. Therefore future research should consider the matter as being unresolved and should still consider producing short-tailed models to describe this phenomenon.

References

[1] Arlitt, M., Jin, T., A workload characterization study of the 1998 World Cup Web site, IEEE Network, 14(3), pp30-37, 2000.

[2] Arlitt, M. F. and Williamson, C. L., Internet Web servers: workload characterization and performance implications. IEEE/ACM Transactions on Networking, Vol.5(5), pp.631-645, 1997.

[3] Arlitt, M., Friedrich, R., and Jin, T., Workload characterization of a Web proxy in a cable modem environment, ACM Sigmetrics Performance Evaluation Review, Vol.27(2), pp25 – 36, 1998.

[4] Barford, P., and Crovella, M. E., Generating representative Web workloads for network and server performance evaluation, Performance SIGMETRICS '98, pp151-160, 1998.

[5] Barford, P., Bestavros, A., Bradley, A., and Crovella, M., Changes in Web client access patterns: Characteristics and caching implications. World Wide Web: Special Issue on Characterization and Performance Evaluation, Vol.2, pp15-28, 1999.

[6] Berendt, B., Mobasher, B., Spiliopoulou, M., Wiltshire, J., Measuring the accuracy of sessionizers for web usage analysis. Proceedings of the workshop on web mining at the first SIAM international conference on data mining, pp. 7-14, 2001.

[7] Brockwell, P.; Davis, R., Time Series: theory and Methods, Springer-Verlag, 1991.

[8] Catledge, L.D., Pitkow, J.E., Characterizing browsing strategies in the World-Wide Web, Proceedings of the Third International World-Wide Web conference on Technology, tools and applications, pp.1065-1073, 1995.

[9] Chen, Y-T., On the Robustness of Ljung-Box and McLeod-Li Q tests: a simulation study, Economics Bulletin, Vol. 3(17), pp. 1 – 10, 2002.

[10] Cherkasova, L., Phaal, P., Session-Based Admission Control: A Mechanism for Peak Load Management of Commercial Web Sites, Transactions on Computers, 51(6), pp. 669-685, 2002.

[11] Crovella, M.E., Bestavros, A., Self-Similarity in Word Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Transactions on Networking, Vol. 5(6), pp. 835 – 846, 1997.

[12] Davis, R.; Resnick, S., Limit theory for the sample covariance and correlation functions of moving averages, Annuals of Statistics, Vol. 13, pp. 179 – 195, 1985.

[13] Downey, A.B., Evidence for Long-tailed distributions in the Internet, Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement, pp. 229 – 241, 2001.

[14] Downey A.B., The structural cause of fie size distributions, Proceedings of the IEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, pp. 361 – 370, 2001.

[15] Downey, A.B., Lognormal and Pareto Distributions in the Internet, Computer Communications, Vol. 28(7), pp. 790-801, 2005.

[16] Eirinaki, M., Vazirgiannis, M., Web mining for web personalization, ACM Transactions on Internet Technology, 3(1), pp. 1-27, 2003.

[17] Feigen, P.D.; Resnick, S.I., Pitfalls of fitting autoregressive models for heavy-tailed time series, Extremes, Vol. 1(4), pp. 391 – 422, 1999.

[18] Figueiredo, D.R., Jiu, B., Feldmann, A., Misra, V., Towsley, D. Willinger, W., On TCP and self-similar traffic, Performance Evaluation, Vol. 61, pp. 129 – 141, 2005.

[19] Fisher, N.I., Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography, International Statistical Review, 51, 25-58, 1983.

[20] Gabaix, X., Zipf's law for cities: an explanation, Quarterly Journal of Economics, Vol. 114(3), pp. 739 – 767, 1999.

[21] Goldstein, M.L., Morris, S.A., Yen, G.G., Problems with fitting to the power-law distribution, European Physics Journal B, Vol. 41, pp. 255- 258, 2004.

[22] Gong, W. Liu, Y. Misra, V. Towsley, D., On the tails of web file size distributions, Proceedings of the 39th Allerton Conference on Communication, Control and Computing, 2001.

[23] Goševa-Popstojanova, K., Mazimdar, S., and Singh, A., "Empirical Study of Session-based Workload and Reliability for Web Servers", 15th IEEE International Symposium on Software Reliability, pp. 403-414, 2004.

[24] Goševa-Popstojanova, K., Singh, A.D., Mazimdar, S., Li, F., Empirical Characterization of Session–Based Workload and Reliability for Web Servers, Empirical Software Engineering, Springer Netherlands, Vol. 11(1), pp. 71-117, 2006(a).

[25] Goševa-Popstojanova, K., Li, F., Wang, X., Sangle, A., A Contribution Towards Solving the Web Workload Puzzle, International Conference on Dependable Systems and Networks (DSN'06), pp. 505-516, 2006(b).

[26] He, D., and Goker, A., Detecting session boundaries from Web user logs. Proceedings of the 22nd Annual Colloquium on Information Retrieval Research, pp.57-66, British Computer Society, 2000.

[27] Hernández-Campos, F., Marron, J. S., Samorodnitsky, G., and Smith, F. D., Variable heavy tails in Internet traffic. Performance Evaluation, Vol. 58(2+3), pp. 261-284, 2004.

[28] Hill, B., A simple approach to inference about the tail of a distribution, Annuals of Statistics, Vol. 3, pp. 1163 – 1774, 1975.

[29] Huntington, P., Nicholas, D., Jamali, H.R., Website usage metrics: A re-assessment of session data. Information Processing & Management. Vol. 44., pp. 358-372, 2008.

[30] Huynh, T., Miller, J., A Formal Model for the Session Timeout Threshold. Journal of Information Processing & Management. In Print.

[31] Jansen, D.W. and de Vries, C.G., On the frequency of large stock returns: putting booms and busts into perspective, Review of Economics and Statistics, Vol. 73, pp. 18 – 24, 1991.

[32] Jansen, B.J., Spink, A., An Analysis of Web Documents Retrieved and Viewed, The 4th International Conference on Internet Computing, pp.65-69, 2003.

[33] Kristol, D.M., and Montulli, L., HTTP State Management Mechanism, RFC 2965 (http://tools.ietf.org/html/rfc2965), October 2000.

[34] Ljung, G. M. and Box, G. E. P., "On a measure of lack of fit in time series models." Biometrika 65, pp. 553-564, 1978.

[35] Mahoui, M., Cunningham, S.J., A comparative transaction log analysis of two computing collections. Lecture Notes in Computer Science. Vol 1923, pp.418-423, 2000.

[36] Mat-Hassan, M., Levene, M., Associating search and navigation behavior through log analysis. Journal of the American Society for Information Science and Technology, 56(9), pp.913-934, 2005.

[37] Mobasher, B., Cooley, R., Srivastava, J., Automatic personalization based on Web usage mining, Communications of the ACM, 43(8) pp. 142-151, 2000.

[38] Mitzenmacher, M., Dynamic Models for File Sizes and Double Pareto Distributions, Internet Mathematics, Vol 1(3), pp. 305 – 333, 2003.

[39] Nicholas, D., Huntington, P., Lievesley, N., Wasti, A., Evaluating consumer Web site logs: Case study The Times/Sunday Times Web site. Journal of Information Science, 26(6), pp. 399-411, 2000.

[40] Nicholas, D., Huntington, P., Jamali, H.R., Watkinson, A., What deep log analysis tells us about the impact of big deal, case study OhioLink. Journal of Documentation, 62(4), pp. 482-508. 2006.

[41] Nicholas, D., Huntington, P., Jamali, H.R., Watkinson, A., The information seeking behaviour of the users of digital scholarly journals. Information Processing and Management, 42(5), pp. 1345-1365. 2006.

[42] Pankratz, A., Forecasting with univariate Box-Jenkins models: Concepts and cases. New York: John Wiley and Sons, 1983.

[43] Reed, J.W., Jorgensen, M., The Double Pareto-Lognormal Distribution—A New Parametric Model for Size Distributions, Communications in Statistics – Theory and Methods, pp. 1733 – 1753, 2004.

[44] Resnick, S.I., Heavy Tail modeling and teletraffic data, The Annuals of Statistics, Vol. 25(5), pp 1805 – 1849, 1997.

[45] Rezaul, K.M. & Grout, V., A Comparison of Methods for Estimating the Tail Index of Heavy-tailed Internet Traffic, Proceedings of the 2nd International Joint e-Conference on Computer, Information, and Systems Sciences, and Engineering, 2006.

[46] Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M., A framework for the evaluation of session reconstruction heuristics in Web usage analysis. INFORMS Journal of Computing, 15(2), pp. 171-190, 2003.

[47] Tian, J., Rudraraju, S., Li, Z., Evaluating Web Software Reliability Based on Workload and Failure Data Extracted from Server Logs, IEEE Transactions on Software Engineering, Vol. 30(11), pp.754-769, 2004.

[48] Tsourti, Z., and Panaretos, J., "Extreme Value Index Estimators and Smoothing Alternatives: Review and Simulation Comparison", Athens University of Economics and Business, Statistics Technical Report No. 149, 2001.

[49] Zipf, G.K., Human Behavior and the principle of least effort, Addison-Wesley, 1949.

Appendix 1: Introduction to Heavy-Tailed and Pareto Distributions

The majority of statistical work is based on short-tailed distributions such as the normal and lognormal distributions. These distributions decay “quickly” (commonly exponentially) in contrast with heavy-tailed distributions. The rank size law[h] [49] can be used to informally describe heavy-tailed distributions. This law states that: the second largest entity is half the size of the largest; the third largest entity is one third the size of the largest, etc. That is, if the entities are ranked from largest (rank 1) to smallest (rank *n*), and their values are denoted as:

$$x_1 \geq \ldots \geq x_n$$

the rank *i* for an entity of value x_i is proportional to the proportion of entities greater than *i*. Or:

$$x_i \approx \frac{k}{i} \qquad (9)$$

for some constant *k*. More formally, Resnick [44] states that a random variable X has a Pareto tail with index α, $\alpha > 0$, if for $x > 1$

$$P[X > x] \approx x^{-\alpha}\ , x > 1 \qquad (10)$$

Many authors provide a slightly more generic distribution of a Pareto distribution by incorporating an additional multiplicative term x_{min}^{α} (the location parameter, the actual term is *L(x).* For the Pareto distribution, $L(x)= x_{min}^{\alpha}$), where x_{min}^{α} is a positive minimal value of *X; i.e.* $\forall x \bullet x > x_{min}^{\alpha}$.

Examination of the Pareto distribution (which is a commonly examined heavy-tailed distribution) involves analysis of the tail index α. Hence, α is examined with the common approach of setting $x_{min}^{\alpha} = 1$ and the requirement for the additional inequality (Equation 10). Technically, the above distribution is defined in a continuous domain; however, within this investigation's domain, the estimation of values clearly has a defined resolution. So strictly speaking X is a discrete random variable; and the discrete probability distribution analogue to the Pareto distribution applies. Therefore, the zeta distribution, or the Zipf distribution, is the actual distribution under analysis. However, the distributions only differ in their definition of the multiplicative term *L(x)* and hence the above definition resolves the issue of having a distribution defined in a continuous domain being applicable on a discrete random variable.

The Pareto distribution is an example of a wider set of distributions, namely heavy tailed distributions. X has a heavy tailed CDF *F(x)* if

$$1 - F(x) = P[X > x] = x^{-\alpha} L(x) \qquad (11)$$

where L is slowly varying; i.e.

$$\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1 \qquad (12)$$

The Pareto distribution is the "simplest" example of a heavy-tailed distribution and is used throughout this paper; and hence the more general definition can be considered solely for information purposes.

[h] The rank size law is a good approximation for entities of high rank, but not for the largest.

The implications of deciding that X is from a heavy-tailed or Pareto distribution are severe as the definition of the standardized moments become problematic. For the Pareto distribution, the first two moments are defined as:

$$E(X) = \frac{\alpha x_{\min}}{\alpha - 1} \quad (13), \qquad Var(X) = \frac{\alpha x_{\min}^2}{(\alpha - 2)(\alpha - 1)^2} \quad (14)$$

This implies that, for $\alpha \leq 1$ the expected value is infinite; and for $\alpha \leq 2$ the variance infinite. Clearly, this demonstrates serious limitations on the types of models which can be constructed using Pareto distributed variables. In addition, these definitions are unrealistic in many situations because the distribution of X will be bounded by physical constraints. Hence, a more rigorous and realistic definition requires the above to hold over a finite range $[x_i, x_{i+j}]$ where the distribution applies.

Although this might seem an unimportant technical point, it is actually a recurring theme in this domain. Basically, all common methods of exploring potentially Pareto distributed variables follow this pattern where the investigation is only carried out within a finite range. Hence, the approaches introduce a bias because they only investigate a small component of the distribution, namely the "tail". x_i is often considered to be the start of the tail, although there is no method of evaluating *i* and no definition of the term *tail*. x_{i+j} is commonly considered to be near x_n; i.e. the highest ranked point within the data set. Clearly, the points, which in theory exist with ranks greater than *n*, cannot be inferred. It is important to note that this range only corresponds to an extremely finite part of the distribution; it is not uncommon for the "tail component" or Pareto range to be defined for less than 1% of the sampled range $x_1 \geq \ldots \geq x_n$. Hence, it is exceptionally difficult to make accurate estimations and infer reliable facts across such amounts of data. The amounts of data are very small both in absolute terms (the raw number of points) and relative terms (the percentage of the total sample). Hence, given the difficulty of accurately characterizing information as belonging to a heavy-tailed distribution and the significant consequences in terms of undefined standardized moments, one should be careful in inferring that a heavy-tailed distribution exists.

It should not be inferred from this discussion that the shape of the Pareto, or heavy-tailed, distributions are highly distinctive from short-tailed distributions. In fact, many heavy-tailed and short-tailed distributions "look" highly similar. For example, Gong et al. [22] plot the data from the Crovella and Bestavros [11] paper, a time-series which contains file sizes transferred over a period of time. They compare the data at the 95% confidence intervals for both Pareto and lognormal models; and observe that the confidence intervals of both models grow with file size; and, at the tail, the two confidence intervals have a "large overlap which makes it difficult to distinguish them". Mathematically, Pareto and lognormal distributions also have a lot in common.

Adapting from Gabaix [20] and Gong et al. [22], consider a time series of i.i.d. positive random variables $Z_1, \ldots., Z_t, \ldots.Z_\infty$. Let Z_i be defined as:

$$Z_t = Z_{t-1} A_t, \; \mathrm{t} = 1, \qquad (15)$$

with $Z_0 = 1$. Taking logarithms yields

$$\ln Z_t = \sum_{i=1}^{t} A_i \text{ , t = 1,} \qquad (16)$$

which by the central limit theorem converges in distribution to a normally distributed random variable. Consequently, Z_t converges in distribution to a lognormal distributed random variable. Now, let's add a condition that Z_t must always exceed a threshold Δ.

$$Z_t = \max\{Z_{t-1}A_t, \Delta\}, t = 1,.... \qquad (17)$$

Gabaix [20] shows that Z_t now converges to a random variable with a Pareto distribution. That is, if $\Delta = 0$, it produces a lognormal distribution, otherwise a Pareto distribution. Because of the similarity between the two distributions, this paper also examines the lognormal distribution for the session lengths recorded.

Appendix 2: Independence of Data

Extreme value analysis methods are techniques that attempt to model rare events based on limited data. Heavy-tail analysis requires a dataset of unobtainable size; and hence, the analysis performed in this paper can be classified as extreme. Many extreme value analysis methods assume that the data set is independent. In fact, the Hill estimator is the only known estimator to perform accurately with dependent data [45][48]. Hence, if the data is considered as dependent, extreme value analysis methods need to be modified. Therefore, in this appendix, we consider this question; however, in this situation, the definition and associated tests for independence is an extremely complex subject with no single clear answer. Independence or randomness is one of the four assumptions that typically underlie all measurement processes. The randomness assumption is critically important because most standard statistical tests depend on it; the validity of the test conclusions are directly linked to the validity of the randomness assumption.

To illustrate this issue, this paper investigates the autocorrelation function (*ACF*) to test for randomness or dependence of the data set. While we will concentrate on an autocorrelation approach to the question, other approaches exist (see Brockwell and Davis [7] for a discussion of alternatives). The session length data can be seen as a time-series because each session length is recorded according to the session start time. If the time-series is completely random then the entire *ACF* should be zero or the null hypothesis is *ACF(k) = 0*; where *k* is the lag. Examining *ACF* values, and determining if they are within the 95% confidence bounds around this central value is commonly utilized as a mechanism to test this hypothesis. If there are values exceeding this bound, then the data is considered dependent. Figures A2.1a and A2.1b show the *ACF* plots for ECE and Site A.

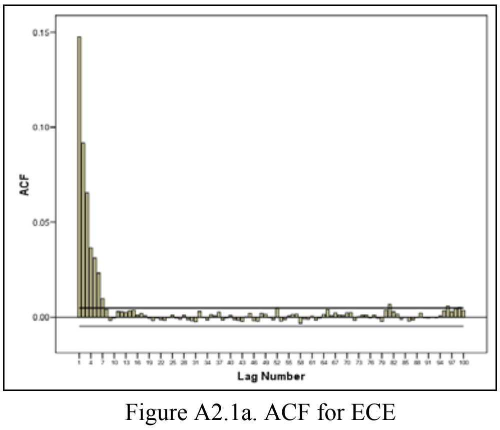

Figure A2.1a. ACF for ECE

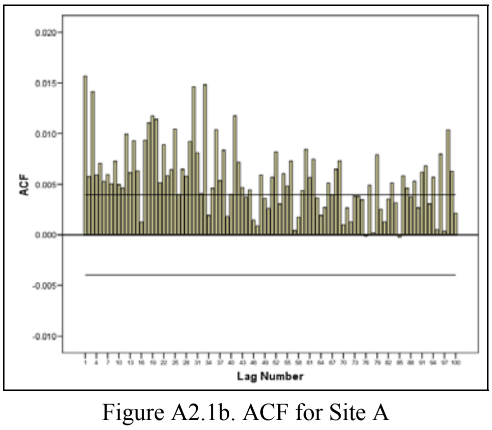

Figure A2.1b. ACF for Site A

These plots also contain the 95% confidence bounds; the plots show that 10% and 67% of the values exceed the upper bound for the ECE and Site A sites respectively, implying that the data may be dependent. However, the analysis uses Barlett's formula [42] to estimate the confidence interval. This formula assumes that the data is normally distributed, and hence the confidence bounds are meaningless if the samples are drawn from a heavy-tailed distribution. Alternatively, the Ljung-Box test [34] can be used to evaluate the null hypothesis. The Ljung-Box test utilizes the following formula:

$$Q = n(n+2)\sum_{k=1}^{m} \frac{acf_k}{n-k} \qquad (5)$$

where ACF_k is the ACF value for lag *k*, *n* is the number of samples and *m* is the maximum lag. Q is distributed as χ^2 with (*m-p-q*) degrees of freedom. We assume that $p = q = 0$; i.e. that the data sets have no trend or periodic information. Clearly, this assumption is invalid as web-sites clearly have many different types of periods with differing resolutions; e.g. day/night; weekday/weekend; non-holiday-period/holiday-period etc. However, the exact nature of the periodic information is not understood and approaches to estimating *p* and *q* can be error prone. Hence we choose to use this simplifying assumption. This assumption effectively inflates the Type II error; which is considered an acceptable risk in this situation. Using the above equation χ^2 is calculated to be 6582.68 and 586.88 for ECE and Site A respectively. These χ^2 values, with 100 degrees of freedom, correspond to a *p-value* of $p < 0.001$ for both websites. Hence, the null hypothesis can again be rejected which means that the data set is dependent, but only if it is not sampled from a heavy-tailed distribution. While this approach can be considered less distributionally restrictive than the previous approach, it is still, both theoretically [31] and empirically [9], not robust to heavy-tailed data.

In addition, the standard *ACF* formula is invalid if the sample is from a heavy-tailed distribution as the formula basically measures deviations from the sample mean, while the sample mean is mathematically undefined for many heavy-tailed distributions. Fortunately, the construction of a non-centered autocorrelation function is straightforward [12]:

$$ACF_{HT}(k) = \frac{\sum_{i=1}^{n-k} X_i X_{i+k}}{\sum_{i=1}^{n} X_i^2} \qquad (6)$$

Figures A2.2a and A2.2b shows the *heavy-tailed ACF* plots for ECE and Site A. These plots show that the *ACF* values do not exceed 0.17 and 0.13 for the ECE and Site A sites respectively. However, confidence bounds estimations (or Q statistics) no longer exist; and unless specific information about the underlying distribution, including accurate values for its parameters, are known, a confidence interval cannot be defined [17].

However, several alternative approaches still exist for evaluating the null hypothesis. Feigin and Resnick [17] show that if the series can be modeled as a moving average process of lag *l* then the coefficients of the *heavy-tailed ACF* should decay to approximately zero beyond *l*; and in the limiting case where $l \to \infty$, the coefficients should again all be approximately zero. This question can be investigated by asking if the co-efficients are summable. Considering Figures A2.3 and A2.4, the answers appear to be negative in both cases. In addition, a more formal test can be constructed by forming a permutation distribution. *The heavy-tailed ACF's* behavior, with respect to the null hypothesis, can be characterized by a summary statistic; e.g. the *maximum absolute ACF coefficient*[i]; this option is recommended by [17]. The *p-value* of the observed summarizing statistic is estimated by generating 999 permutations of the time-series; computing the statistic for each permutation and counting the number (*C*) of values greater than or equal to the actual observed statistics. The *p-value* is given approximately by *((1+C)/1000)*. Clearly, this approach avoids relying in the asymptotic theory or distribution for this particular summarizing statistic; and the test is distributionally robust for heavy-tailed situations. Figures 3.3 and 3.4 display the results of the permutation test.

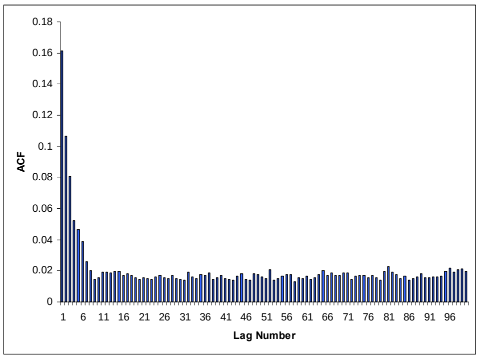

Figure A2.2a. Heavy-Tailed ACF for ECE

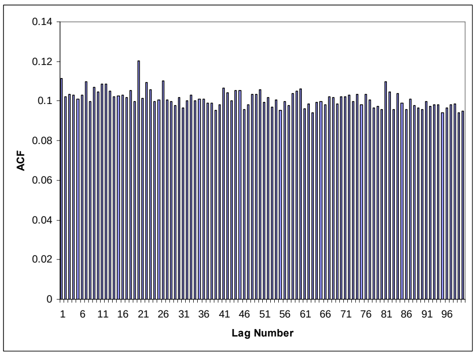

Figure A2.2b. Heavy-Tailed ACF for Site A

Visual inspection show that the majority of the *max(ACF)* of the permutations are below the actual *max(ACF)* which is represented by the horizontal line. In fact, for the ECE site, none of the permutations are greater than or equal to the actual *max(ACF)* which means the *p-value* < 0.001. For Site A, two of the permutations are greater than or equal to the actual *max(ACF)*; hence, the *p-value* < 0.003. Because the *p-value* for both websites are below the standard type I error cut-off values, the null hypothesis can be rejected which means that the data for both websites are dependent.

While our approach is now a relatively robust examination of the null hypotheses several situations still exist where the validity of our approach and hence the associated results are at best questionable and at worst non-applicable. Feigin and Resnick [17] empirically demonstrate that the *heavy-tailed ACF* tends to exhibit erratic results in the following situations:

- the presence of any non-linearities, such as the process being a bilinear process,
- when the process is a moving average*(l)* process; if $l > m$,
- the series is contaminated by (additive) outliers.

These situations clearly represent risks to the internal validity of the results presented in this appendix.

[i] Other options include the partial or biserial autocorrelations.

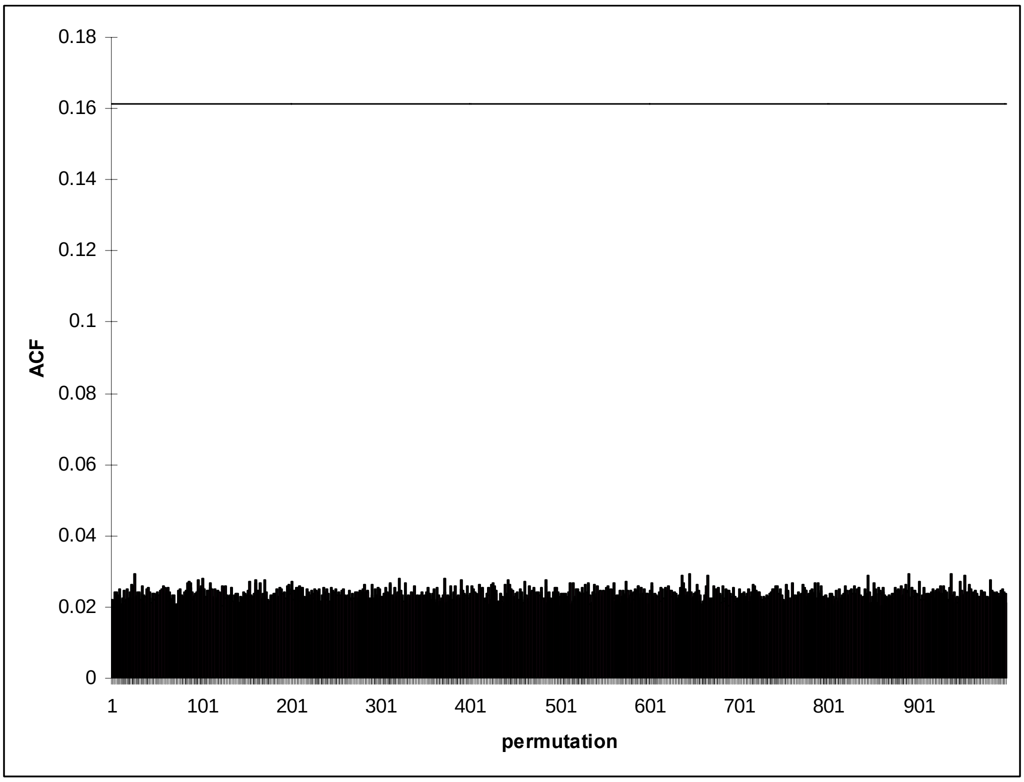

Figure A2.3. Permutation test for ECE

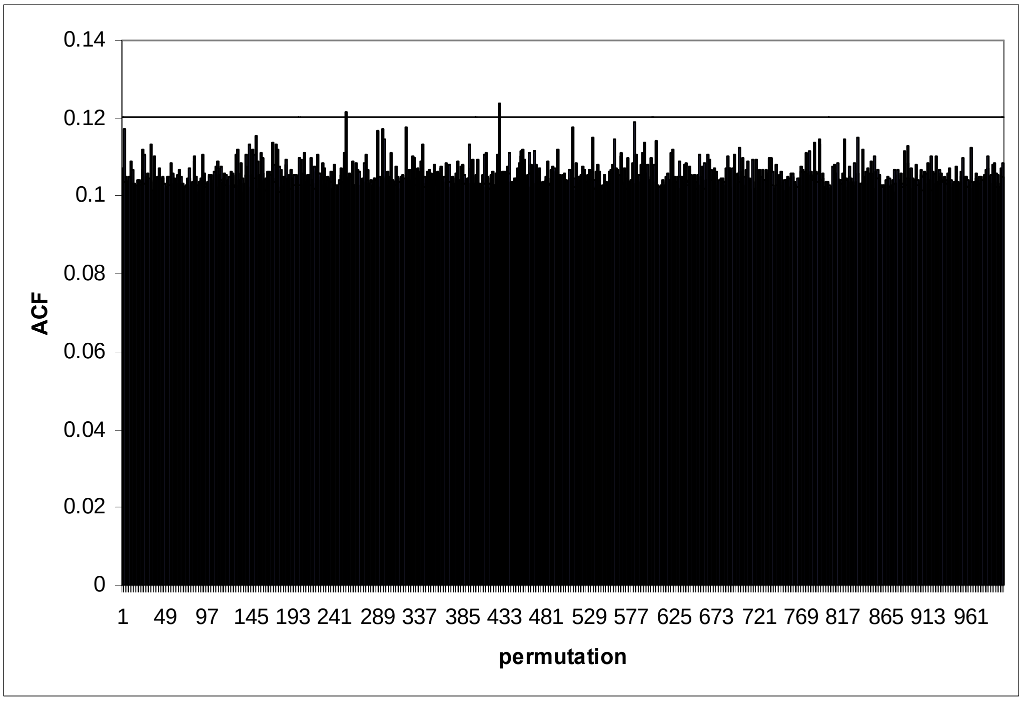

Figure A2.4. Permutation test for Site A