In Search of Invariants for E-Business Workloads

Daniel Menascé‡
Flávia Ribeiro†

‡ Dept. of Computer Science
George Mason University
Fairfax, VA 22030 USA

menasce@cs.gmu.edu

Virgílio Almeida†
Rodrigo Fonseca†

† Dept. of Computer Science
Univ. Fed. Minas Gerais
Belo Horizonte, MG 31270
Brazil

virgilio@dcc.ufmg.br

Rudolf Riedi§
Wagner Meira Jr.†

§ Dept. of Electrical and
Comp. Engineering
Rice University
Houston TX 77251 USA

riedi@rice.edu

ABSTRACT

Understanding the nature and characteristics of e-business workloads is a crucial step to improve the quality of service offered to customers in electronic business environments. However, the variety and complexity of the interactions between customers and sites make the characterization of e-business workloads a challenging problem. Using a multi-layer hierarchical model, this paper presents a detailed characterization of the workload of two actual e-business sites: an online bookstore and an electronic auction site. Through the characterization process, we found the presence of autonomous agents, or robots, in the workload and used the hierarchical structure to determine their characteristics. We also found that search terms follow a Zipf distribution.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—Web-based Services; C.2.4 [**Computer-communication networks**]: Distributed systems — client /server systems

General Terms

Design, Measurement, Performance

Keywords

e-commerce, WWW, workload characterization, performance modeling, heavy-tailed distribution

1. INTRODUCTION

Understanding the nature and characteristics of e-business workloads is a crucial step to improve the quality of service offered to customers in electronic business environments. E-business workload characterization can lead to a better understanding of the interaction between customers and Web sites and can also help design systems with better performance and availability.

There are very few published studies of e-commerce workloads because of the difficulty in obtaining actual logs from electronic companies. Most companies consider Web logs to be very sensitive data. Past studies of WWW workloads concentrated on information provider sites and found several characteristics common to many sites. Some of the characteristics considered important are: file size distributions, file popularity distribution, self-similarity in Web traffic, reference locality, and user request patterns. A number of studies of different Web sites found file sizes to have heavy-tailed distributions and object popularity to be Zipf-like. Other studies of different Web site environments demonstrated long-range dependencies in the user request process, primarily resulting from strong correlations in the user requests. In particular, [4] identified ten workload properties, called *invariants*, across six different data sets, which included different types of information provider Web sites. Some of the most relevant invariants are: i) images and HTML files account for 90-100% of the files transferred; ii) 10% of the documents account for 90% of all requests and bytes transferred; iii) file sizes follow the Pareto distribution, and iv) the file inter-reference times are independent and exponentially distributed. Shortly after, [3] found that the popularity of documents served by information provider Web sites follows a Zipf's Law. In [7], the authors pointed to the self-similar nature of Web server traffic. All these studies were performed almost five years ago. Since then, several major changes have been observed in the WWW. The most important are: clients now have much larger bandwidth, the number of users has grown exponentially, and e-commerce became one of the major applications on the Web. A question that naturally arises is: are the characteristics and invariants found in information provider Web sites still valid for e-business workloads?

To answer this question, we define a hierarchical structure to characterize the workload and apply this structure to two different types of actual e-business sites: an online bookstore and an electronic auction site. We examine statistical and distributional properties of the e-business workloads and compare these properties across the two datasets. As much as possible, we compare the features of these workloads with the invariants that were discovered for information dissemination Web sites.

EC'00, October 17-20, 2000, Minneapolis, Minnesota.

Agents and robots are expected to constitute an important part of the load of Web sites in the future. We present an analysis of robots and agents, that we found in the bookstore workload. Autonomous agents, also known as bots or robots, will play an increasingly important role in the Web and in particular in e-business applications. By 2005 [5], "at least 25% of the then-current PC/workstation user base, will allow their personal agents to anticipate their needs." Therefore the importance of understanding the e-business workload generated by agents.

The rest of the paper is organized as follows. Section two briefly describes related work. Section three shows the approach used to characterize e-commerce workloads. The next section describes the data collection process. Session five analyzes two logs from actual e-business sites and characterizes the workload at the request level. Characterizations of higher levels are provided in sections six and seven. Section eight discusses characteristics of workloads generated by software robots and agents that were identified in the logs. Finally, section nine presents concluding remarks.

2. RELATED WORK

To evaluate the performance of an e-business site, one needs a solid understanding of its workloads. Most of the existing references on workload characterization in the WWW focus only on information provider Web sites [13]. A few references have addressed the problem of workload characterization for e-commerce sites. Existing work concentrates on characterizing Web workloads composed of sequences of file requests [4, 7]. The characteristics and statistical properties of workloads on the Web have been studied by many papers [3, 4, 7, 13]. A number of studies of different sites identified WWW workload properties and invariants. For instance, file sizes have heavy-tailed distributions (e.g., Pareto distribution), object popularity follows Zipf's Law and across WWW traffic is bursty across several time scales [7].

In [9], the authors introduce the notion of session, consisting of many individual HTTP requests. However, they do not characterize the workload of e-commerce sites, which is composed of typical requests such as browse, search, select, add, and pay. The analysis focuses only on the throughput gains obtained by an admission control mechanism that aims at guaranteeing the completion of any accepted session. The work in [14] proposes a workload characterization for e-commerce servers, where customers follow typical sequences of URLs as they move towards the completion of transactions. The authors though do not present any characterization or properties of actual e-commerce workloads. In [11], the authors propose a graph-based methodology for characterizing e-commerce workloads and apply it to an actual workload to obtain metrics related to the interaction of customers with a site. For example, the paper shows how to obtain information such as the number of sessions, average session length, and buy-to-visit ratio. Reference [12] presents several models (e.g., customer behavior model graph and customer visit model) for workload characterization. It also shows how models can be obtained from HTTP logs.

3. APPROACH

E-business workloads are composed of sessions. A *session* is a sequence of requests of different types made by a single customer during a single visit to a site. During a session, a customer requests the execution of various e-business functions such as browse, search, select, add to the shopping cart, register, and pay. A request to execute an e-business function may generate many HTTP requests to the site. For example, several images may have to be retrieved to display the page that contains the results of the execution of an e-business function.

Workload characterization can be accomplished at many levels: user level, application level, protocol level, and network level. An e-business workload can be viewed in a multilayer hierarchical way, as shown in Fig. 1. This paper focuses on the characterization of three levels, represented by the request layer (protocol level), function layer (application level), and session layer (user level). This hierarchy can be used to capture changes in user behavior and map the effects of these changes to the lower layers of the model.

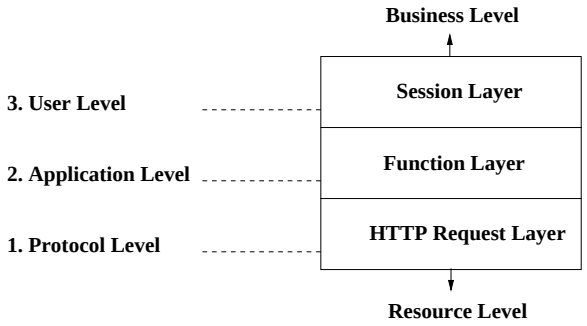

Figure 1: A hierarchical workload model.

Our approach is to analyze each layer individually in order to obtain a characterization of the arrival process and usage statistics. The latter includes multiscale statistical analysis, long range dependency (LRD), and burstiness. The former covers information such as: session interarrival times, session interrarival times for specific e-business functions, session length distribution, e-business function distribution per session, and number of active sessions and initiated sessions.

4. DATA COLLECTION

The data used for the workload characterization came from two actual e-businesses, an e-tailer and an auction server. The e-tailer is a bookstore that sells exclusively on the Internet. The auction site sells Internet domains. In both cases, the data consist of access logs recorded by the WWW server of each e-business.

The data comprises two weeks of accesses to each of these sites. The bookstore logs were collected from August 1st to August 15th, 1999, while the auction server logs are from March, 28th to April 11th, 2000.

During these two weeks, the bookstore handled 3,630,964 requests (242,064 daily requests on average), transferring a total of 13,711 megabytes of data (914 MB/day on average). The auction server has a smaller load, and answered 466,058 requests (31,071 requests/day) which amounts to 1,863 megabytes of data (124 MB/day). Most of these requests are for embedded images in the response pages. In the case of the bookstore, images account for 71% of the requests, while in the auction workload they represent 85.27% of the requests.

For workload characterization purposes, we only considered requests that invoke the execution of e-business functions (e.g., search, register, pay). Thus, requests to images were eliminated from the analysis. E-business function related requests amounted to 26.32% and 14.73% of the bookstore and auction requests, respectively. The difference in percentage between the two sites is explained by the larger number of images used by the auction site. Thus, the bookstore executed 63,711 e-business functions per day, and each service response had 12,618 bytes, on average. We should note that service-related requests are responsible for most of the network traffic, comprising 84.62% of the data sent by the bookstore server and 92.18% of the data sent by the auction server. This is explained because most of the gif files embedded in pages are usually already cached and are not transmitted back to the client. Although the auction pages contain more images than the bookstore pages, the auction server employs fewer images, normally used for banner advertisement and page layout, while the bookstore images may be book covers and comprise a larger number of files.

5. REQUEST CHARACTERIZATION

In this section we study the nature of the arrival process of requests in statistical terms, with mainly two goals in mind: i) the extraction of statistically significant features towards classification, understanding and modeling of request workload, and ii) eventual prediction of workload, allowing for an adaptive provisioning of resources towards optimal performance.

It is now a well accepted fact that strong correlations are present in various aspects of the World Wide Web, from request arrivals on servers to packet arrivals on the network. These correlations lead to "burstiness" or high variability which may degrade performance and throughput if not accounted for. We carry out a statistical analysis across all time scales to detect correlations and assess their strength.

The fact that statistical analysis and modeling has to incorporate different methods according to time scale is most apparent as we attempt to accommodate various trends. On the largest time scale of days (in our study), the weekend produces somewhat less volume, while on the scale of hours the presence of a periodic sleep-wake pattern per day is visually obvious. It is not our intent to explore these patterns. On finer time scales, structure is much less obvious and it is our goal to present a simple analytical tool which allows one to distinguish scales of "noisy oscillations" from scales with strong correlations and "trends."

5.1 Multiscale analysis of requests

Let us first study the overall arrivals of requests at the e-commerce site. A visual inspection of the number of requests arriving at the bookstore on different time scales, i.e., in time intervals of varying length (see Fig. 2) reveals, even to the unexperienced eye, an apparent strong dependence which shows long sequences of increase or decrease of volume (trends), particularly pointedly at intermediate time scales of the order of minutes.

A simple quantification of such a dependence is achieved by computing the sample variance by time scale: if arrivals are independent, the sample variance doubles if the length of the interval doubles. The variance exceeds twice the variance of the original interval if the arrivals are positively correlated and will not reach twice the variance if the

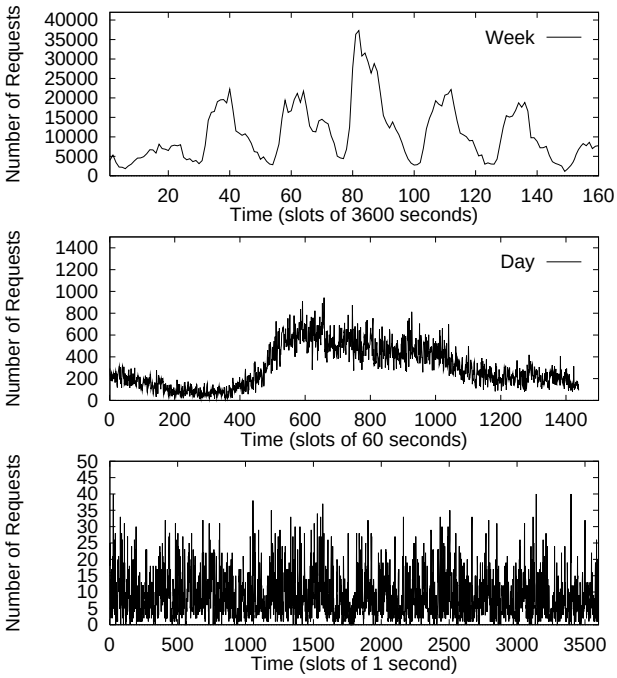

Figure 2: Number of requests arriving at the bookstore in time intervals of varying length, from top to bottom: 3,600 sec, 60 sec, 1 sec

arrivals are negatively correlated. More specifically, if X_k denotes the number of arrivals in time interval $[k\ \delta, (k+1)\ \delta]$, where δ is the finest time resolution one is interested in, then the following traditional procedure $X_k^{(n)} = 2^{-n} \times (X_{k2^n} + X_{k2^n+1} + \ldots + X_{k2^n+2^n-1})$ *averages* the arrivals in $[k2^n\delta,\ (k+1)2^n\delta]$ and can be computed efficiently through the recursion $X_k^{(n)} = (X_{2k}^{(n-1)} + X_{2k+1}^{(n-1)})/2$. (Note that Fig. 2 shows the actual *number* of arrivals on three different time scales, not the averages.) The log-log plot of the variance against scale, i.e., $\log_2$ var $X^{(n)}$ versus n, is called variance time plot (VTP). This plot has the slope -1 for independent data (recall the normalization factor 1/2 necessary to provide averages instead of total counts) and a different behavior for dependent data. The extreme case of positive correlation is a constant series X_k with a flat VTP.

A more interesting case of dependent behavior constitutes the so-called statistical self-similarity which is defined by the requirement that var $X^{(n)} = \sigma^2\ 2^{2H-2}$ where the Hurst parameter H lies between 0 and 1. This case is of interest due to the existence of appealing, simple, Gaussian processes with such properties, such as the fractional Gaussian noise and the auto-regressive FARIMA processes [16]. For $H = 1/2$ we find ourselves back in the case of independent data while the decay of the VTP is slower for $H > 1/2$, i.e., with slope $2H - 2$, and we have positive correlations. In the extreme case of $H = 1$ the VTP is flat and fractional Gaussian noise almost surely constant.

It is important to be able to detect strong dependencies because they degrade estimation by increasing the variance of the estimation error. On the other hand, by detecting strong dependencies, one can foresee not only mean behavior but also temporary phases of increase or decrease in volume.

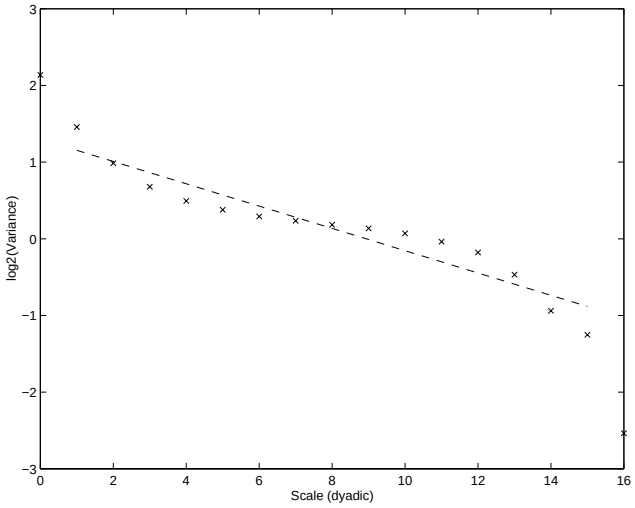

Figure 3: Variance time plot for the bookstore site.

The VTP is a crude measure of the correlation structure with known bias and poor performance as an estimator of the LRD parameter H. However, as a tool for a first look, it is completely valid (see [1, 2, 16] and references therein).

The VTP plot of the bookstore (see Fig. 3) shows an overall averaging decay corresponding to $H = 0.9273$ indicating a strong dependence. More precisely, the strongest dependence seems to be present at intermediate time scales from 16 to 4096 sec, corresponding to aggregation 4-12 in Fig. 3 (there δ corresponds to 1 sec).

This is further confirmed by the "correlation" plots of Fig. 4 for the bookstore for the two-week period. This plot gives an idea of the next-neighbor dependence on the time scales of Fig. 3 through plots of $X_{k-1}^{(n)}$ versus $X_k^{(n)}$. Note that the more the data is clustered along the diagonal, the higher is the predictability. Intermediate time scales exhibit the most convincing concentration along the diagonal. For illustration purposes, we also show in Fig. 4 the "correlation" plot of a series of independent random variables. In this case, no structure and no clear clustering is visible.

A possible explanation for particularly strong correlations on the time scale of several dozen seconds may be human think time. The overall self-similarity, at least "asymptotically," may be argued for by invoking the well known on-off process which was crucial in explaining self-similarity in network traffic loads [8]: The number of requests per session follows a heavy-tailed distribution. Since the number of requests sent by a client per time unit is limited by the nature of the HTTP protocol, sessions are thus sending requests over on-times which are heavy tailed. Numerical support for this claim comes from our analysis of session duration in Sec. 7, which shows that the distribution of session length follows a power law. The on-off model then relates the exponent of this fat tailed distribution directly with H.

We now compare the arrival process of requests at two different sites, a bookstore and an auction site. Figure 5 presents the average number of requests per day in our two-week logs for these two sites and shows marked differences that allow us to draw conclusions on client behavior. We can clearly notice the traffic reduction during the weekends, which are the days 0, 6, 7, 13, and 14. We can see that the

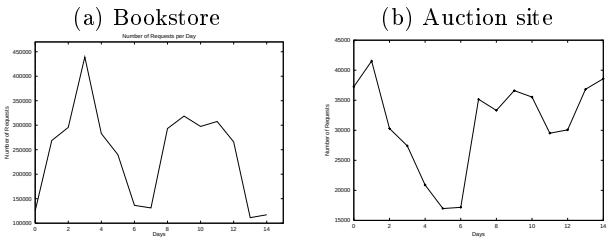

Figure 5: Number of requests arriving at the site at a time resolution of one day. (a) bookstore and (b) auction site.

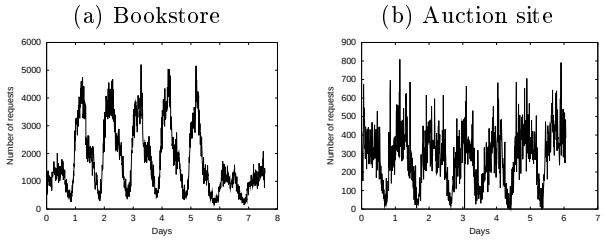

Figure 6: Number of requests arriving at a site for a time resolution of 640 secs for the second week. (a) bookstore, (b) auction site.

auction site exhibits a much higher weekly traffic variation than the bookstore. We can confirm this behavior by looking at the graphs in Figs. 6 and 7. The graph of Fig. 6 shows the number of requests for the second week at a resolution of approximately ten minutes. We can see that there are seven peaks or groups of peaks, one per day of the week for both sites. For the bookstore, there is clearly one peak per day. The variations in intensity for consecutive points in the graph of Fig. 6 are relatively small, which allows for good prediction at that time scale. When the time scale is reduced to close to one minute, as indicated in Fig. 7, the variations become much larger, reducing the predictive power at this time scale.

Figure 8 shows the distribution of references per domain. If we compare these results with the ones presented in [4], we can see that the popularity of hosts is much less concentrated on few domains since e-business sites tend to have a much broader audience than content dissemination Web sites.

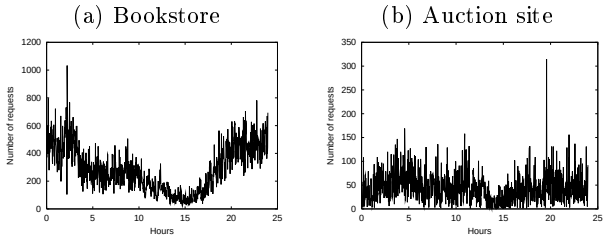

Figure 7: Number of requests arriving at a site for a time resolution of 80 secs for a period of one day. (a) bookstore, (b) auction site.

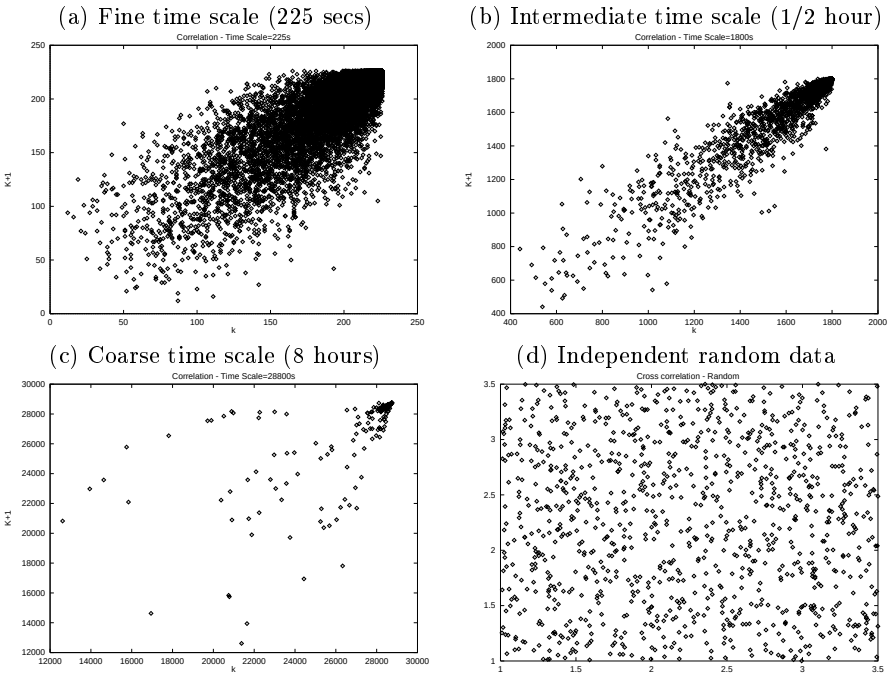

Figure 4: $X_{k-1}^{(n)}$ versus $X_k^{(n)}$ for the bookstore for the two-week period and for different time scales.

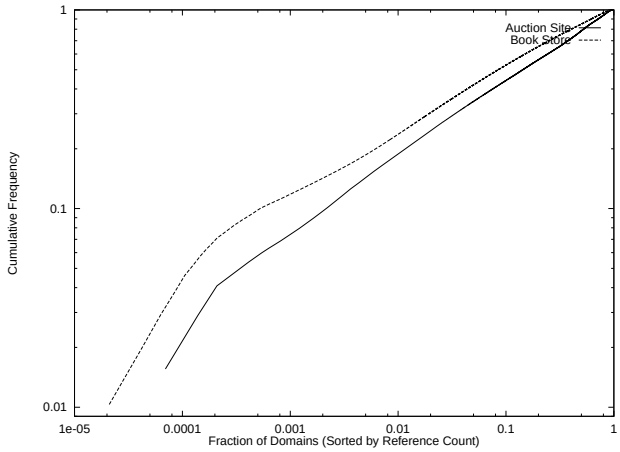

Figure 8: Distribution of References by Domain

6. FUNCTION CHARACTERIZATION

In this section, we characterize the workload at the level of e-business functions. Our first criterion is the nature of the function. When considering an online store, we may divide the functions into four groups: static, product selection, purchase, and other. Static functions comprise the home and informational pages about the store. Product selection includes all functions that allow a client to find and verify a product they are looking for: browse, search, and view. Purchase functions indicate a desire to buy, either by selecting a product for later acquisition (e.g., add to cart) or by ordering it (e.g., pay). One interesting invariant in the logs we analyzed, despite the time scale is that more than 70% of the functions performed are product selection functions. Table 1 presents a breakdown of functions for the virtual bookstore.

Bookstore		Auction	
Function	Frequency	Function	Frequency
Home	11.92%	Home	20.70%
Browse	17.72%	Browse	14.66%
Search	36.30%	Search	16.74%
View	19.99%	View	4.87%
Add	5.44%	Bid	0.08%
Pay	1.19%	Sell	7.99%
Account	2.44%	Account	5.99%
Robot	0.04%	Robot	0.06%
Info	3.66%	Info	9.44%
Other	1.31%	Other	2.31%
		Auth	9.18%
		Register	7.29%
		Admin	0.71%

Table 1: Distribution of e-business functions.

In the auction site, there are functions that relate to the process of posting items for sale. The majority of accesses also correspond to selection of products, but not as much as in the bookstore case. On both sites, the functions directly related to spending money have a very low frequency.

When we split requests according to the e-business functions they invoke, i.e., search, browse, add, and pay, we find two clearly distinct classes. While the behavior on large time scales of hours and days of all functions follow the already observed human behavior, their small scale behavior is quite different. For example, Fig. 9 shows the number of requests to execute searches at the bookstore and to retrieve

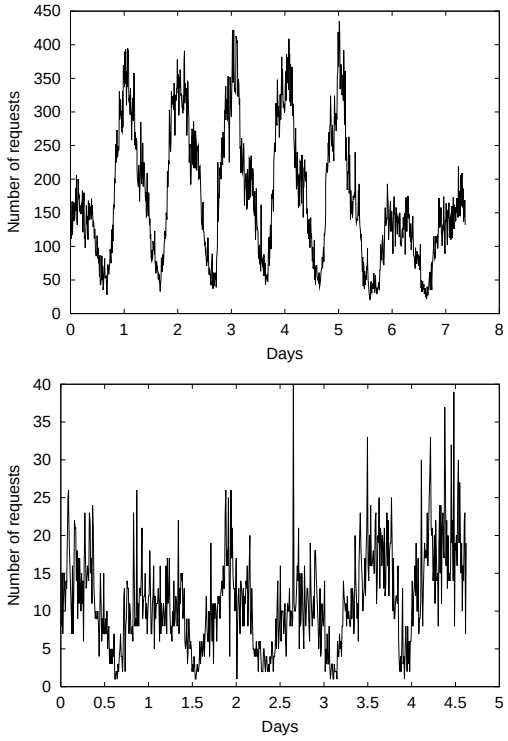

Figure 9: Number of arriving requests to execute frequent e-business functions. The time resolution is similar to that of Fig. 6. (Top) Bookstore: search, (Bottom) Auction site: home.

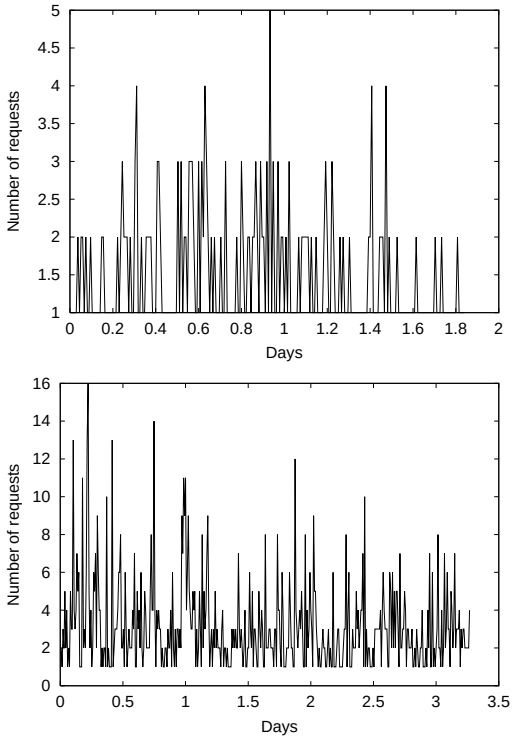

Figure 10: Number of arriving requests to execute less frequent e-business functions, The time resolution is similar to that of Fig. 6. (Top) Bookstore: pay, (Bottom) Auction site: view.

the home page of the auction site for several days for a time resolution similar to that of Fig. 6, i.e., in the order of ten minutes. If we compare Fig. 9 to Fig. 6, we see a similar pattern. This indicates that requests to execute frequent e-business functions exhibit a similar pattern of behavior as HTTP requests.

The same is not true for less frequent functions such as pay and view, as indicated in Fig. 10. This figure shows clear bursts and a very different behavior from Fig. 6. Here, a more advanced statistical analysis revealing the multifractal scaling would be in place [15] and prediction is harder. The more frequent functions such as "search" and "home" show statistics similar to the overall load of requests and are well described by Gaussian LRD processes.

This difference in small scale behavior is best understood when thinking in terms of doubly stochastic Poisson processes where Poisson arrivals are driven by a varying intensity which is itself random. As intensities are low, the spikyness of Poisson arrivals are apparent; as intensities grow, the Poisson distributions are well approximated by the Gaussian. In a unifying approach one would aim at measuring the "hidden" intensity through a Bayesian scheme, thus capturing the driving stochastics of request arrivals and allowing for a deeper understanding and more control. This is left for future investigation.

Figure 11 shows the number of search requests for the bookstore on a daily basis. We can see that this graph exhibits, in the first week, a different behavior from the request/day one (Figure 5). We attribute the difference to

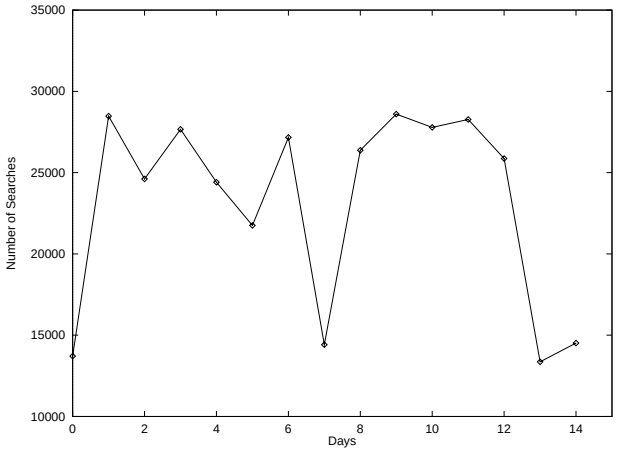

Figure 11: Number of searches arriving at the bookstore at a time resolution of one day.

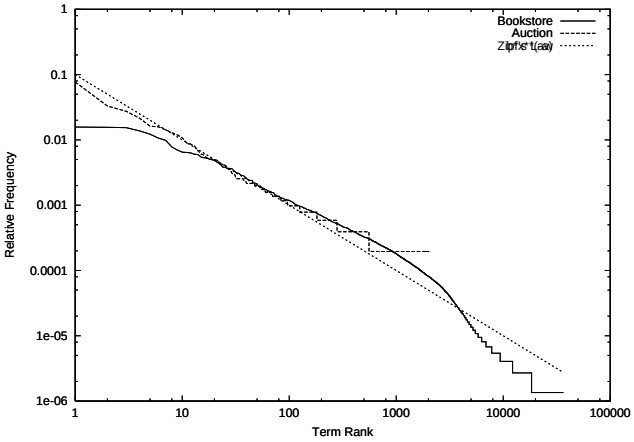

Figure 12: Popularity of search terms.

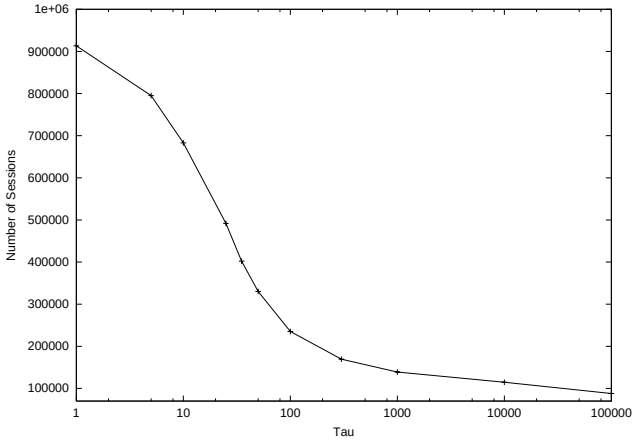

Figure 13: Influence of the threshold τ on the total number of sessions initiated.

the fact that the search function is used by robots, which behave differently from human users. For instance, the spike observed in day 3 results from an unexpected number of requests for the home page.

6.1 Popularity of Search Terms

Zipf's law[17] was originally applied to the relationship between a word's popularity in terms of rank and its frequency of use. It states that if one ranks the popularity (denoted by ρ) of words used in a given text by their frequency of use (denoted by P) then

$$P \sim 1/\rho.$$

Figure 12 shows that Zipf's law applies quite strongly to *the terms used for search functions*. This result is similar to the one found in [3], which showed that Web documents returned by Web servers also follow a Zipf's Law. The figure displays three curves: one for the bookstore, one for the auction site, and another for Zipf's Law. As it can be seen, there is a good match with Zipf's Law over an extremely wide range of popularity.

7. SESSION CHARACTERIZATION

Session boundaries are delimited by a period of inactivity by a customer. In other words, if a customer has not issued any request for a period longer than a threshold τ, his session is considered finished. Usually, sites enforce this threshold and close inactive sessions to save resources allocated to these sessions. For the auction site, we know that the HTTP server enforced a threshold of twenty minutes. Since we do not have this information for the bookstore site, we had to estimate the threshold from the log. The value of τ has an influence on the number of sessions being handled by the site.

We discuss the effect of τ in what follows. Figure 13 shows the effect of the value of τ in the total number of sessions initiated for the bookstore site. As we can see, as the threshold increases from 1 to 100 sec, the number of sessions initiated decreases very rapidly. ¿From 1000 sec on, the decrease is very small. This indicates that most sessions last less than 1,000 sec. A de facto industry-standard has been that 30 minutes (i.e., 1,800 sec) should be used to delimit sessions.

Figure 14 shows the distribution of session lengths, measured in number of requests to execute e-business functions, for both sites. The threshold for the session length is 1,800 seconds for the bookstore, while there is no threshold for delimiting the sessions of the auction site, since it implements timeouts for its sessions. The graphs of Fig. 14 show the empirical tail of the distribution of the session length X, i.e., $P\{X > x\}$ for the bookstore and auction sites, as well as the tail of the exponential and Pareto distributions. A random variable X, such as Pareto, that has a heavy-tailed distribution is characterized by $P\{X > x\} \sim x^{-a}, 0 < a < 2$. Among other implications, a heavy tailed distribution presents a great degree of variability, and a non-negligible probability of high sample values. The exponential distribution decays much faster than a heavy tailed distribution. In a log-log plot, x^{-a} is a straight line with inclination $-a$. We can distinguish two regions in the plot of Fig. 14. The first one comprises session sizes up to 100 requests, in which the curves for both sites are similar. In particular, in the region from about 5 to 100, they are fit by a straight line (not shown for clarity) with inclination ~ -2.05. For sessions longer that 100, the behavior changes. We can see that for the auction site, the probability for longer sessions falls abruptly, whereas for the bookstore it remains close to the straight-line plot of a Pareto-like distribution with $a = 1$. This "very" heavy tail is most likely due to the accesses by robots, which tend to exhibit long sessions. The auction site was not accessed by any detectable robot, and this explains why one does not see sessions much longer than 100 requests.

We can also notice the most sessions are small (about 90% of the sessions for both workloads have less than 10 requests.

7.1 Usage Analysis

The top part of Fig. 15 shows the number of sessions initiated per day at the bookstore site for various values of the threshold τ. A small value of τ corresponds to the extreme case of considering each request as a session. The picture clearly shows that there is very little difference in the number of sessions as τ is increased from 1,000 sec to 10,000 sec. This is a strong argument in favor of the 30-minute standard. A similar behavior is seen in the top part of Fig. 16.

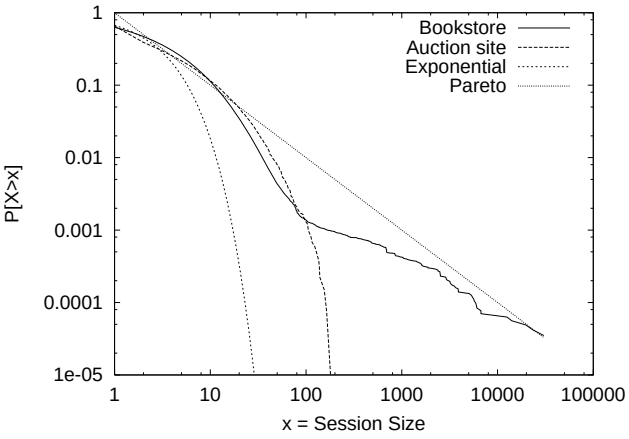

Figure 14: Session size distribution.

The bottom part of Figs. 15 and 16 indicate the number of sessions initiated per day and per hour for the auction site. If we compare the shape of the graph of the number of initiated sessions for the bookstore site for $\tau = 1000$ and for the number of initiated sessions for the auction site with the corresponding graphs of Fig. 5, for number of arriving requests, we see some degree of similarity.

Figure 17 displays the number of active sessions on an hourly basis for various values of the threshold τ. Again, very little variation is seen for $\tau > 1000$ sec. At a time scale of one hour, we observe a high variability in the number of active sessions per hour since the session timeout for the auction site or the threshold of 1,000 sec for the bookstore are of the same order of magnitude as the time scale.

8. AGENT CHARACTERIZATION

We observed in the previous sections that long-lived sessions account for a significant part of the traffic served by e-business servers. One of the explanations for such sessions are agents, also known as robots. Autonomous agents are an essential component of major portals and information service sites. On behalf of these sites, agents endlessly catalogue the Web. E-business sites that offer catalogs of products are constantly visited by agents looking for product information for shopping comparison services. Also, Web users have access to shopping agents that request information from commercial sites. There are also meta-search engines that allow price comparison among several stores, and are also considered as agents for the sake of workload analysis. The increasing number of sources of information on the Web and the development of agent technology indicate that the number of agents tends to increase, justifying a more detailed study of the workload associated with these agents.

In this section, we use the proposed characterization approach for characterizing the load generated by agents in e-business sites. Detecting whether an agent is submitting requests to an e-business server is important because we can use this information for both limiting their access (using the "robots.txt" file for instance) or prioritizing the access of real clients to the server as a means of guaranteeing the quality of service provided by the site [10].

Several criteria can be used for identifying agent-generated

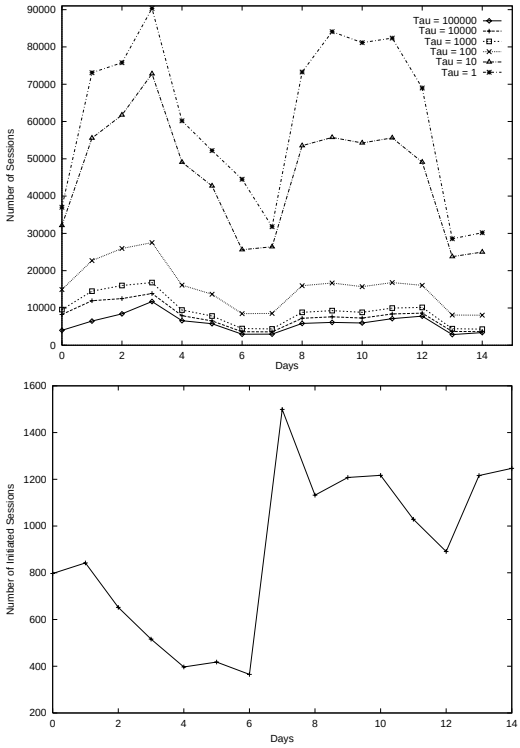

Figure 15: Number of initiated sessions per day for the bookstore (top) and the auction site.

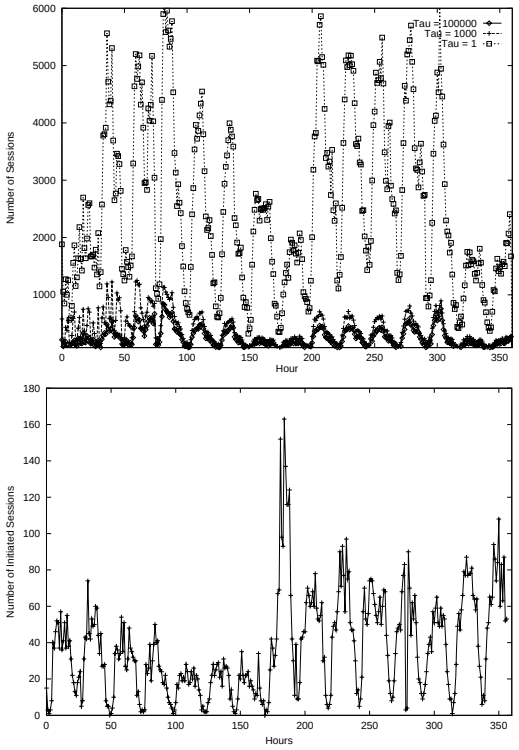

Figure 16: Number of initiated sessions per hour for the bookstore (top) and the auction site.

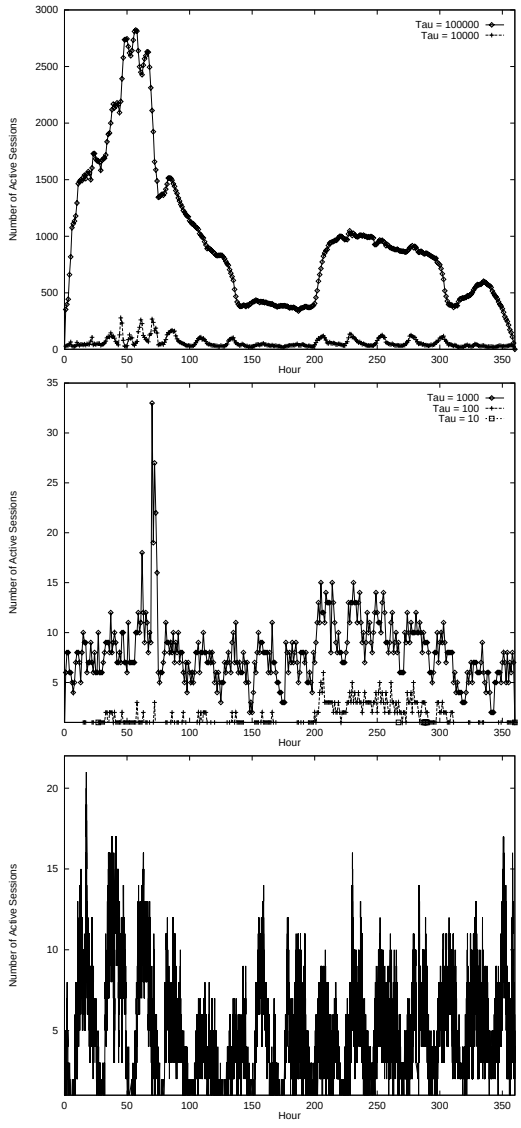

Figure 17: Number of active sessions. Top two figures are for the bookstore.

sessions. Obviously, these criteria depend on the nature of the business supported by the server and how the site is organized. As we discuss next, agents may be identified through criteria applied to the various levels of characterization. For a standard e-tailer, we distinguish seven criteria that can be used for identifying agents accessing the server. Next we discuss each of these criteria and their application to 130,314 sessions determined using a 30-minute threshold.

Session Layer Indicators

Function: One session-layer criterion that clearly identifies agents are the functions most frequently requested. For instance, a criterion is to have more than 95% of searches in a session. In our sample log, this criterion found six sections that are clearly from agents and comprise 151,588 requests.

Session length: Agents tend to request data to an e-business server for a longer period than customers. Thus, a session-layer criterion is to identify agents by having sessions significantly longer than the average user session. In the case of the e-tailer logs, we selected sessions that comprise at least 500 requests. We found 79 sessions that satisfy this criterion, which add to 282,478 requests.

Function Layer Indicators

Entry point: Agent requests often do not follow a logical sequence of pages. One function-level criterion is the first service in the session being different from "home." We found that 26% of the sessions do not start at the home page and contain more than 20% of the requests.

Unfeasible functions: Robots usually do not request some functions that we call unfeasible. Examples of unfeasible functions are "add to cart" and "pay." This criterion belongs to the function layer of the proposed characterization and does not seem to be very effective, since it misclassified real clients as agents because these clients visited the e-tailer without choosing or buying anything. In fact, this criterion classified 128,642 sessions as agents.

Embedded files: Robots do not request embedded files, that is, files that compose the response pages. The applicability of this criterion, however is limited because images are usually cached in the client host or proxy caches. This is a function-layer criterion, where the embedded files are associated with function pages. We found that 44% of the sessions satisfy this criterion, comprising 30% of the requests.

Request Layer Indicators

Interarrival time: An example of a request-layer criterion is the request arrival process. For the sake of identifying agents, one may expect that their requests arrive at fixed time intervals. Considering that the agents are automated, we can use the variance of these time intervals as a criterion, that is, small variances identifies an agent. It is interesting, then, to observe that the long lived sessions of agents in our trace have the same arrival statistics of the function "search" as the overall load. This indicates that agents tend to not space their requests equally, but to perform searches at best speed possible. Being dependent on answers from the server in order to formulate an intelligent new search, robots are limited by server response time, but not by think time, meaning that this criterion is not really applicable to characterize agents behavior.

Self identification: Before issuing any request, agents often request the file "robots.txt". This file states the access rules for agents to request objects from the server and allows one to identify agents using a request-layer criterion. We verified that 379 sessions requested the file. Again, we found that these sessions are probably longer than we considered, since the number of unique session identifiers is just 189.

9. CONCLUDING REMARKS

Several studies have been published regarding the workload of information provider sites. However, very few studies are available for e-business sites. This paper used a hierarchical approach for workload characterization of e-business sites. The characterization was done at the session, e-business function, and request levels. The paper shows a large number of graphs containing a detailed characterization of the two sites analyzed: a bookstore and an auction site.

Some of the findings are: i) most sessions last less than 1,000 sec, ii) more than 70% of the functions performed are product selection functions as opposed to product ordering functions, iii) the popularity of search terms follows a Zipf distribution, iv) there is a very strong correlation in the arrival process at the request level. This correlation is given by a Hurst parameter value of 0.9273, v) at least 16% of the requests are generated by robots, vi) 88% of the sessions have less than 10 requests, and vii) the session length, measured in number of requests to execute e-business functions, is heavy tailed, especially for sites subject to requests generated by robots.

The results presented in this paper should be considered preliminary since logs from only two sites were available. It is recognized by many that one of the major challenges in carrying out experimental work in e-commerce is the lack of data. Most companies regard their logs as sensitive information that should not be made public. We are in the process of obtaining additional logs to extend the scope of our analyses.

10. ADDITIONAL AUTHORS

Flávia Peligrinelli Ribeiro (flavia@dcc.ufmg.br), Rodrigo Fonseca (rfonseca@dcc.ufmg.br), and Wagner Meira Jr. (meira@dcc.ufmg.br) are all from the Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.

11. REFERENCES

[1] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. In *Self-similar Network Traffic and Performance Evaluation*. Wiley, Spring 2000.

[2] P. Abry, P. Gonçalvès, and P. Flandrin. Wavelets, spectrum analysis and $1/f$ processes. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics: Wavelets and Statistics*, volume 103, pages 15–29, 1995.

[3] V. Almeida, M. Crovella, A. Bestavros, and A. Oliveira "Characterizing Reference Locality in the WWW," *Proc. IEEE/ACM International Conference on Parallel and Distributed System (PDIS)*, Dec. 1996.

[4] M. Arlitt and C. Williamson, "Web Server Workload Characterization," *Proc. 1996 SIGMETRICS Conference on Measurement of Computer Systems*, ACM, May 1996.

[5] J. Brown and P. Duguid, *The Social Life of Information*, Harvard Business School Press, 2000.

[6] M. Calzarossa and G. Serazzi, "Workload Characterization: A Survey," *Proc. of the IEEE*, Vol. 81, No. 8, August 1993.

[7] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, 5(6), pp. 835–846, December 1997.

[8] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, pp. 1–15, 1994.

[9] L. Cherkasova and P. Phaal, "Session Based Admission Control: A Mechanism for Improving the Performance of an Overloaded Web Server," HPL-98-119, HP Labs Technical Reports, 1998.

[10] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes, "Business-oriented Resource Management Policies for E-Commerce Servers," *Performance Evaluation*, September 2000.

[11] D. A. Menascé, V. Almeida, R. Fonseca, and M. Mendes, "A Methodology for Workload Characterization for E-Commerce Servers," *Proc. 1999 ACM Conference in Electronic Commerce*, Denver, CO, Nov. 3-5, pp 119-128.

[12] Menascé, D. and Almeida, V., *Scaling for E-Business: technologies, models, performance and capacity planning*, Prentice Hall, Upper Saddle River, NJ, May 2000.

[13] Pitkow, J., Summary of WWW characterizations, *World Wide Web*, No. 2, 1999.

[14] D. Krishnamurthy and J. Rolia, "Predicting the Performance of an E-Commerce Server: Those Mean Percentiles," in *Proc. First Workshop on Internet Server Performance*, ACM SIGMETRICS 98, June 1998.

[15] R. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk. "A multifractal wavelet model with application to TCP network traffic," *IEEE Trans. Info. Theory, Special issue on multiscale statistical signal analysis and its applications*, Vol. 45, pp. 992–1018, April 1999.

[16] M. Taqqu, V. Teverovsky, and W. Willinger. "Estimators for long-range dependence: An empirical study," *Fractals*, Vol. 3, pp. 785–798, 1995.

[17] G. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.