Modeling and Evaluation of Internet Applications

Ajit K. Jena[a], Adrian Popescu[b] and Arne A. Nilsson[b]

[a]Computer Center, Indian Institute of Technology, Bombay-400076, India

[b]Dept. of Telecommunications and Signal Processing, Blekinge Institute of Technology,
371 79 Karlskrona, Sweden

The paper presents a modeling and evaluation study of the characteristics of SMTP and HTTP applications in terms of user behavior, nature of contents transferred and application layer protocol exchanges. Results are reported on measuring, modeling and analysis of application level traces collected, at client and server ends, from different environments such as university networks and commercial Frame Relay networks. The methodologies used for capturing traffic flows and for modeling are reported. Statistical models have been developed for diverse parameters of applications, which can be useful for building synthetic workloads for simulation and benchmarking purposes. Both applications possess a session oriented structure. Within each session, a number of transactions may occur, and this number has been modeled as well.

1. INTRODUCTION

The purpose of the paper is to present a characterization study of traffic collected at clients and servers for SMTP and HTTP, to be further used in a client-server simulation framework [5]. Detailed results are reported on measuring, modeling and analysis of traffic collected from different sites together with methodologies used for capturing traffic flows and for modeling. The highlight of our work is the way we model the structure of Web pages, to consider the number of embedded items as well. For complete rendering of a Web page, the base schema and the embedded items are needed. In particular, we confirm results of former studies showing that the distribution of HTTP document sizes can be modeled by a mixture of Lognormal and Pareto distributions as well as that diverse classes of Web servers seem to show structural similarities in their distributional properties [1,2,8,11]. Furthermore, it is also shown that the number of embedded documents in a Web page can be well modeled by a Negative Binomial distribution and a model is put forth that can be used for the generation of the number of embedded objects in a Web page. In addition, other several properties observed earlier are revealed, which seem to be specific for SMTP like for instance the bimodal structure in the body size distribution for SMTP messages [11]. Finally, an important point in our study is regarding the common modeling structure used to model the considered applications. This takes the form of transaction oriented structure of application sessions, which is useful in developing the simulation model.

The paper is organized as follows. Section 2 describes the measurement infrastructure. The collection of data traffic is presented in section 3. Section 4 describes the modeling methodology. Sections 5 and 6 present the SMTP, respectively HTTP, characteristics considered in the study and the summary statistics for the collected data sets. A model is put forth for the number of embedded documents in a Web page. The paper is concluded in section 7.

2. MEASUREMENT INFRASTRUCTURE

Several measurement utilities have been used for traffic monitoring and analysis at both client and server sides. The objectives are to capture the type and sizes of application layer objects, structural similarities and/or differences among characteristics of sessions as well as to search for possible invariant characteristics in object flows across different user organizations.

- **NIKSUN NetVCR**^{TM} monitoring and analysis system [9] is a non-intrusive system for network monitoring and analysis able to collect and storage data from LAN or WAN link equipments.
- The **tcptrace** utility, for mapping IP packets to TCP flows [10], to increase capabilities of NetVCR^{TM}.
- The **augmented tcptrace** utility is used to capture the message timings [4].
- **x_flowstat** utilities, where **x** refers to **smtp** or **http**, is used to help summarize the huge information generated from flows and to extract application specific features [4].
- **GetWeb** mirroring software, to duplicate the content tree of a specific Web server on local disks [12].
- Analysis of **server logs**, in particular SMTP logs.

3. CAPTURING APPLICATION TRAFFIC FLOWS

The results reported in the paper are based on datasets with up to about 25000 SMTP flows collected from the teachers' access subnetwork at the Blekinge Institute of Technology (BTH), Karlskrona, Sweden and from the Indian Institute of Technology (IIT), Bombay, India. Typically the time periods selected span over a week or 10 days, or sometimes even a month. It was observed that about 100 SMTP sessions were usually generated every day, giving a total of about 6000 SMTP sessions for analysis. The message body sizes have also been collected from the SMTP logs of the main mail servers at BTH and IIT. Both hosts handle the bulk of the inter-site email message loads for the respective institutions.

For HTTP data traffic, we used the NetVCR^{TM} system and the Web mirroring software to collect data from a number of sites. No considerations are given to HTTP traffic dynamically generated via mechanisms like Common Gateway Interface (CGI). The content pages on WWW servers are created by the hosting organizations towards specific goals of furthering own interests. We hypothesized that organizations with similar business have similar content structures and should therefore exhibit similarities in structure. Accordingly, three classes of servers were identified: educational; media companies; and companies engaged in information technology business. For each category, three to five different servers were probed. Starting at their homepages, and using a recursive technique, we descended 10 levels of Web hierarchy and collected their characteristics. This process resulted finally in a collection of 2500 to 10000 observations for each of the servers involved in the experiment.

The other aspects of characterizing the WWW application objects were on the client side. The experiment involved collecting packet traces using NetVCR^{TM} from a user environment, i.e., the student and staff access subnetworks at BTH. Over measurement periods, a total of about 18000 observations were collected from the staff access subnet and 20000 observations were collected from the student subnet.

4. MODELING METHODOLOGY

Detection and estimation of heavy-tailed properties in the distributions of application layer objects is an important aspect in performance modeling. Often the random variable possessing a heavy tail appears hidden behind another distribution. For instance, it has been observed that the (real) distribution of the sizes of HTTP objects can be described as Lognormal and this can contribute with about 80 to 95 % of the probability mass with the remaining mass contributed by a Pareto distributed random variable [11]. The crux of the problem lies in determining the cutoff point between the two distributions. The procedure for fitting probability distributions for mixture models is as follows [6,11]:

- The first step is to investigate the histogram, the empirical distribution plot and the summary of the observations (e.g., minimum, maximum, mean, median, variance). This helps in eliminating some candidate distributions, and may also help in determining if the observation data set can be modeled using a single distribution or mixture models are required.

- Visual techniques like Hill estimator plot and Complementary Cumulative Distribution Function (CCDF) plot on a log-log scale can be used to see whether the set of observations has a heavy tail.

- The next step is to try some candidate distributions and perform the null hypothesis test to find out the significance level of the observations having been drawn from the candidate distribution. The Anderson-Darling (AD) test is used [3]. In the case of mixture model, a method of successive right-censoring is used to partition the observations. The AD test needs also to be modified to be applicable to censored samples [3,11]. If the sample does not meet the significance criteria, then the process is repeated with smaller subsamples. In the case of mixture models, a suitable cutoff point was selected where the null hypothesis is best satisfied for both distributions.

- The parameters for the distributions that passed the null hypothesis are obtained by using the Maximum Likelihood Estimation (MLE) method. Special methods to estimate parameters from censored samples are used [6].

- Having obtained the distributions and the parameters thereof, a measure of the discrepancy between the observations and the model can be obtained by using the λ^2 criteria [11].

5. SMTP

SMTP (RFC821, RFC822) deals with the electronic mail service on Internet. Users interact with it by means of *User Agents (UAs)*, which allow them to compose and send mail messages. On the receiver side, each user has a separate mailbox that is accessed by UA to render message contents on user screen. The exchange of email is done by *Message Transfer Agents (MTAs)*. The MTA processes the messages arriving from local users and from remote MTAs. It acts as a postal sorting clerk by posting the messages to local mailboxes or passing the outbound messages to the *outbound messages queue*. On the transmit side, the MTA takes the messages from the queue and sends them directly to the destination host.

5.1. SMTP characteristics

The traffic generated by SMTP is a combination of protocol messages and contents. The sender and recipient MTAs behave as ON-OFF sources. ON periods are determined by protocol message elements and user message contents. OFF periods depend on the response from the peer process, thus giving rise to a lock-step behavior. The following statistics have been collected:

- Session arrival timings.
- Number of e-mail transactions within a session.
- Message body lengths. To characterize the distributional properties of the message body sizes, it is necessary to look into application layer payload information. This is however not possible as it is usually not only controversial but even criminal. Another solution is to do analysis of SMTP responder logs. Without having access to message contents, these logs can be used instead to collect message body sizes and number of recipients each message is addressed to.
- Message timings. These are the timestamps at which different messages are observed by the monitor.

5.2. Traffic analysis

SMTP exchanges have been observed to exhibit a bimodal structure [11], and these results are confirmed by our study. The control messages are short (often less than 50 bytes) whereas message bodies are bigger. Some of the main results are:
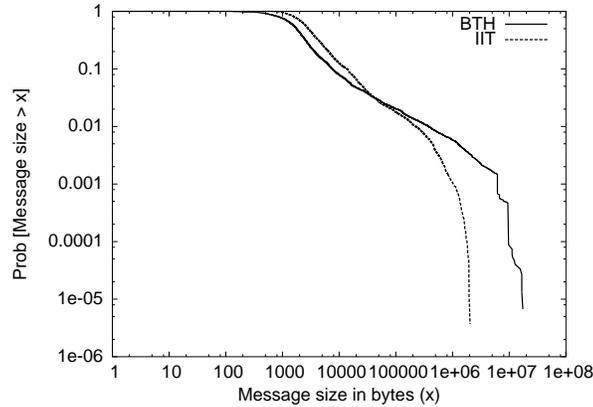
Figure 1. CCDF of SMTP message body sizes

Table 1
Model parameters for the message sizes

Org.	1st Qn, median, mean, 3rd Qn in bytes	Subs size	λ^2 value	Uniform parameters	Pareto parameters	Body mass
Bombay	1828, 2857, 12090, 5292	100	0.11	$a = 500$ $b = 3000$	$\alpha = 0.99$ $\beta = 3000$	$m = 0.52$
Karlskrona	1028, 1766, 11660, 2015	100	0.09	$a = 100$ $b = 2150$	$\alpha = 0.79$ $\beta = 2150$	$m = 0.606$

- The arrival process of user sessions for WAN traffic has been previously shown to be well modeled by Poisson processes [11]. Evaluating the monitor timings for SMTP connection inter-arrival times in the collected data sets, we have observed that this is an Exponentially distributed process, with a mean of 20 seconds.

- For simulation purposes, we have also analyzed the protocol messages between initiator and responder. The study revealed that they have a very simple structure, and message lengths can be modeled by Uniformly distributed random numbers, whereas other messages have constant sizes [4].

- The process of fitting a distribution model to the message body was initiated with the investigation of CCDF plot for message body sizes (figure 1). We observe the mixture of two distributional components, one for bulk and another for upper tail (power-law characteristics). The body and tail portions were tested for Uniform and Pareto distributions, respectively. The results of the fitting process are shown in table 1 together with the parameters of the Uniform distribution (*a* and *b*, the lower and upper limits of the Uniform random variable) and Pareto distribution (the shape parameter α and the location parameter β). We also observe the disturbing phenomenon of $\alpha < 1$, indicating unstable processes in SMTP message sizes. Such values have been disregarded in the simulation model.

- With the advent of MIME encodings, e-mail is increasingly being used to send files as attachments to the mail body. The bimodal behavior seems to be caused by this mixture. The body messages are likely to be the cause of the Uniformly distributed body portion, and the attachments are likely to generate the Pareto tail portion.

6. HTTP

A Web page consists of diverse (content) objects, which are transferred from a WWW server. On the server, each object is physically stored as a distinct file, e.g., HTML file, GIF or JPEG image, audio clip. The HTTP protocol (RFC1945, RFC2616) is used for data transfer. Characterization of the WWW services essentially involves characterization of the objects transferred with HTTP.

6.1. Session structure

Most Web pages are designed around an object called *main (primary) object*, which serves as an anchor and holds, in its body, several references to other URIs. The references can be *links* or *embedded (secondary) objects* like in-lined images. The *links* are underlined or marked with special colors or symbols when they are rendered on user screens. These are hints to the human user so the user can click on them and fetch them. The participating URIs corresponding to the *embedded objects* are implicitly fetched and rendered on the user screen. Very often, the embedded references are located on the same server as the *main page* but in a few cases they reside on a different server.

A session is defined to start at the instant the user asks for a URI by explicitly naming it or by pointing the mouse at a link reference and clicking on it. With this epoch, the client side connects to the server and starts sending the requests for the *main object* and subsequently for the *embedded objects*. The server processes the requests and sends the contents back to the client. Aborts are not considered in our model. The session is said to have ended when the last embedded object is received at the client side. At this point, the view on the browser screen is complete and the human user starts reading the page. From the traffic point of view this is a silent period, which we define as the inter-session gap (i.e., passive OFF time). This gap represents the human user behavior (user think time) and typical action may simply involve browsing, printing, or even saving on local disk for future reference.

A stochastic marked point process is used to model the HTTP session and the associated timings (fig. 2). A HTTP session is defined as being the procedure for downloading a single Web page, with associated main and secondary objects. The following variables are considered for a session:

- ON time: time duration elapsed for the fetching of main object and associated embedded objects.
- active OFF time: time duration between consecutive fetches of embedded objects (same ON period).
- passive OFF time: time duration elapsed between the completion of the transfer of one page and the beginning of the following session (user think time).
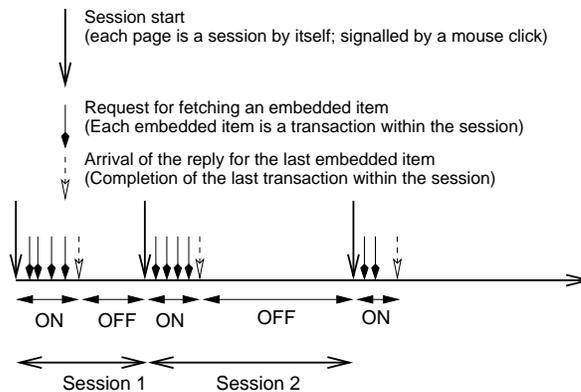

Figure 2. Model of HTTP session

6.2. HTTP characteristics

Using the software tools described in section 2, the following statistics have been collected:

- Session arrivals. These statistics indicate the intensity of WWW requests generated by users.
- Number of transactions within a HTTP session. This is essentially given by the number of embedded objects inside the Web page that constitutes the session, and it determines the session duration.
- Request message lengths. These are messages from client to server for fetching Web objects (of type *main* or *embedded*).
- Statistics about the success or failure of WWW access requests.
- Reply messages. The messages from server to client consist of header and content.
- Percentage of requested items that were not transferred due to caching on the client side.
- Type of objects transferred, i.e., HTML or non-HTML.
- Content lengths.

6.3. Traffic analysis

WWW is a server-centric application that generates most of data flow from the server towards the clients. The server has some kind of *latent* properties in terms of content pages it holds and their structural properties (e.g., number and type of *embedded objects* in a page). Some of these properties are exhibited during WWW sessions when the user first selects a specific server and then a portion is selected of the total domain on WWW pages. In terms of analogy, the situation is similar to energy systems where the objects, by virtue of their location, have potential energy. Some of the potential energy gets converted into kinetic energy when the object moves. Accordingly, one of the main goals of this experiment was to study both the *latent* properties (captured by using *GetWeb*) and the *exhibited* properties (captured by using *http_flowstat*). The main characteristics of the WWW application are as follows:

- Statistics of session arrivals (user behavior). Previous research has shown that such arrivals can be modelled by a renewal process [11]. The inter-arrival time between requests for *main objects* has been observed to be well modeled by an Exponential process with a mean of 15 seconds.
- Inter-arrival time gaps between successive WWW requests belonging to a given WWW session, which represent the active OFF times of the WWW client (fig. 2). These timings are mainly dependent on the HTTP version and the pipelining feature of the browser software. It has been clearly observed that these timings are in the range 100-500 milliseconds.
- Request message sizes. Compared to the bimodal distribution observed in [8], the analysis of our traces reveals instead that this parameter can be well modeled as a Uniformly distributed random variable in the range 180-600 [4]. Similarly, the header sizes of the response messages also show limited variability, and they can be modeled by a Uniform random variable in the range 250-700.
- The models for HTTP object sizes (collected through Web mirroring software) are presented in table 2. The CCDF plots for the server categories are shown in figure 3. The Web pages have been fitted as a mixture of Lognormal (with parameters μ and σ) and Pareto distributions (with parameters α and β). The quality of fitting is shown in the 'Remarks' column: (Lok) and ($Lpoor$) refer to the Lognormal body portion, whereas (Pok) and ($Ppoor$) refer to the Pareto tail. These remarks are based on the significance levels passed by the censored AD test. The choice of subsample sizes for applying the AD test is also shown. The relative frequencies of HTML and non-HTML types are shown. It is observed that this model does not fit well in some cases. In a majority of cases however, it appears that the tail portion comes from a Pareto process. However, in some cases (two of the news/media servers, highlighted with a '*' in the 'Remarks' column), the power-law behavior is negligible and the complete distribution can be well represented as a Lognormal random variable only. Furthermore, in the case of commercial organizations, the Lognormal distribution seems to be

Table 2

Modeling results for HTTP object sizes collected from different categories of Web servers

Site Name	μ	σ	α	β	Tail mass	Signif sub size	λ^2 value	cutoff point (byte)	HTML %	non HTML %	Remarks
www.cs.berkeley.edu	8.49	1.86	0.766	5717	0.13	200	0.06	81435	44	56	$(Lok)(Pok)$
www.bth.se	7.97	1.27	0.624	1160	0.24	200	0.09	11394	58	42	$(Lok)(Ppo)$
www.cs.purdue.edu	7.09	1.27	0.83	1547	0.15	200	0.06	15137	74	26	$(Lpo)(Pok)$
www.cam.ac.uk	8.12	1.28	1.03	1860	0.16	200	0.02	10964	68	32	$(Lok)(Pok)$
www.uiuc.edu	8.01	1.30	1.62	3865	0.17	200	0.04	11526	69	31	$(Lok)(Pok)$
www.hp.com	-	-	0.46	54	0.04	100	0.12	53658	40	60	$(Lpo)(Pok)$
www.ibm.com	-	-	0.63	418	0.06	100	0.13	36096	61	39	$(Lpo)(Pok)$
www.nokia.com	-	-	0.51	142	0.06	100	0.13	35428	34	66	$(Lpo)(Pok)$
www.lucent.com	-	-	1.12	16578	0.06	100	0.14	199961	60	40	$(Lpo)(Pok)$
www.rediff.com	8.50	1.56	-	-	-	100	0.05	-	67	33	*
www.nytimes.com	8.85	1.88	1.85	8423	0.29	100	0.17	16408	72	28	$(Lpo)(Pok)$
www.the-week.com	9.78	0.64	-	-	-	100	0.05	-	34	66	*
www.india-today.com	8.90	1.53	1.46	6160	0.13	300	0.05	24824	50	50	$(Lok)(Pok)$

Table 3

Modeling results for the HTTP object sizes collected through packet traces

BTH Karlskrona	Type	μ	σ	α	β	Tail mass	Signif sub size	λ^2 value	cutoff point	%
Staff Net	HTML	8.54	1.60	-	-	-	300	0.06	-	45
	non HTML	7.71	1.68	1.69	3178	0.14	300	0.04	10112	55
Student Net	HTML	6.64	1.50	1.30	1418	0.14	200	0.04	11394	30
	non HTML	7.08	1.87	0.93	1147	0.22	200	0.06	5802	70

a poor choice to fit the body portion. The HTML and non-HTML groups, when plotted separately on the CCDF plot, reveal that this behavior appears to mostly come from the HTML type documents (figure 4). From the CCDF plot it also appears that the body portion can be approximated by a Uniform distribution, and this seems to be common to all commercial servers investigated.

- The analysis of a representative set of HTTP object sizes collected on the client side (staff and students access networks at BTH) is shown in table 3 and figure 5. The qualitative aspects of model fitting are much better than shown in table 2 and hence they are not reported.

6.4. Structure of Web pages

The structure of Web pages is characterized in terms of statistical properties of the number of embedded objects and types (HTTP vs non-HTTP). Qualitative design aspects of the content reflect properties of a specific organization. Intuitively, it is expected that Web page design features at universities may be different from those for commercial purposes, and the differences are both quantitative and qualitative. The density of *embedded objects* is expected to be higher in the case of news/entertainment and commercial sites. On the other hand, the bulk of pages on newspaper sites will be of HTML type and the sizes should be small. The university sites are expected to be rich with research reports having Postscript or PDF format. These files tend to be large in size and therefore contribute towards the power-law behavior in the tail. Based on this, and on the collected data traces, we have developed a stochastic model for the number of *embedded objects* that may occur in a Web page. The modeling methodology is as follows:

- The basic dataset used in the modeling process is the output of the **GetWeb** software that probed a large number of WWW servers and recursively downloaded their content pages.
- The data generated in the previous step was grouped using two levels of indexing, the *serverid* and the *pageid*. The design of indexing mechanism was geared towards the goal of retaining the server
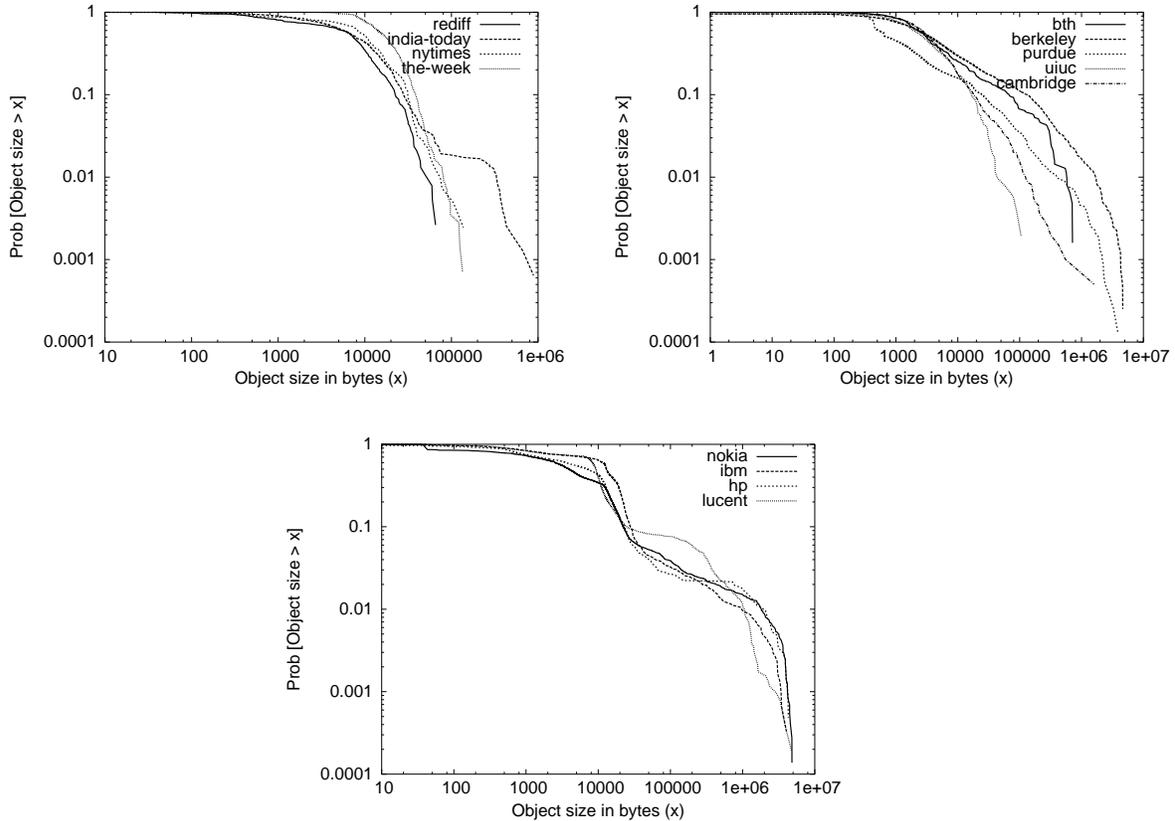
Figure 3. CCDF of different servers: entertainment (left top), educational (right top) and commercial (bottom)

information and also the particular page information within the specific server. It was observed that in many cases a single *embedded object* occurred many times (e.g., GIF files representing the bullets in a HTML page layout). The modeling procedure counts the contribution of such multiple occurrences as 1. In other words, only unique *embedded objects* are considered and duplicate occurrences of these objects are rejected. This corresponds to a realistic scenario because after the GIF object is downloaded the first time, it is available in the local cache, and hence it is not downloaded again.

- Subsequently, a mathematical model for this specific case was built [7], which is useful for simulation purposes. The number of *embedded objects* was found to be well modeled by a Negative Binomial distribution (Fig. 6). This distribution is used, for instance, to find the number of Bernoulli trials needed to find the r-th success. Furthermore, it has been also shown that the Negative Binomial distribution can be well approximated as a mixture of Poisson distributions where the expected values of the Poisson distributions vary according to a Gamma distribution [6].

Accordingly, a simple model of the number of embedded documents in a Web page can be as follows. Assume that, on a specific server i, the number of *embedded objects* in the Web page j is $X_{i,j}$. This number is conditioned by the intensity λ_i, which is a server characteristic that represents the density of presence of *embedded objects* in the Web pages resident at server i. In other words, the quantity λ_i is implicitly fixed when the user selects a particular server during a Web session. The (conditional)
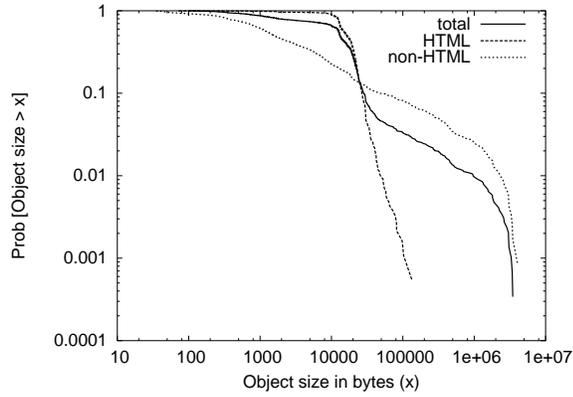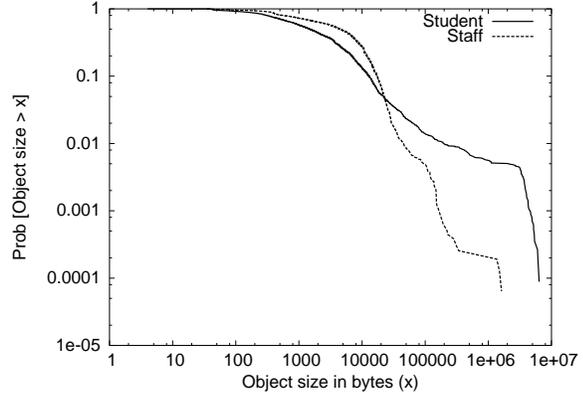
Figure 4. CCDF of ibm.com servers

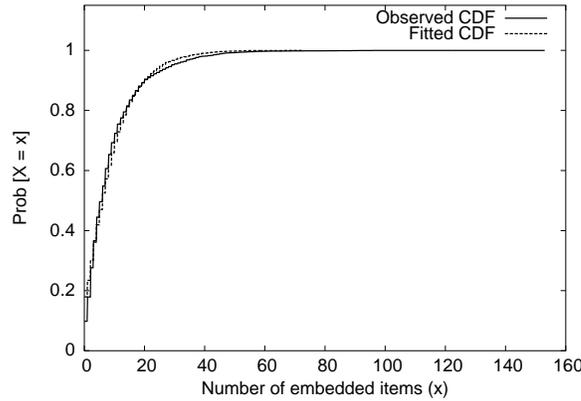Figure 5. CCDF of BTH Karlskrona (client side)

Figure 6. CDF of the number of embedded items vs fitted model

distribution $F(X|\lambda)$ of the number X of *embedded objects* in a Web page can be modeled as:

$$F(X|\lambda) \sim P(\lambda) \tag{1}$$

where P represents the Poisson process. From the collected data (on total number of *embedded objects* occurring in a page), the number of pages having zero *embedded objects* is excluded. Consider C as being the fraction of these pages. The remaining observations (having the number of *embedded objects* greater than zero) can be used to generate a distributional model by using Gamma distribution with shape parameters α and β. The random process X occurs as a mixture of Poisson distributions and the expected intensity λ of every Poisson distribution vary according to Gamma distribution with PDF:

$$f(\lambda) = \frac{\lambda^{(\alpha-1)} exp\left(-\frac{\lambda}{\beta}\right)}{\beta^{\alpha} * \Gamma(\alpha)} \tag{2}$$

$\Gamma(\alpha)$ is the Gamma function (also known as Euler function) of variable α and the parameters λ, α and β are larger than zero. Under these conditions X has a Negative Binomial distribution (with parameters

α and β) and the PDF of the number of *embedded objects* is [6]:

$$Prob[X = x] = \int_0^\infty f(\lambda) \frac{\lambda^x exp(-\lambda)}{x!} d\lambda = \left(\begin{array}{c} \alpha + x - 1 \\ \alpha - 1 \end{array} \right) \left(\frac{\beta}{\beta + 1} \right)^x \left(\frac{1}{\beta + 1} \right)^\alpha \tag{3}$$

A model for the generation of *X* is as follows:

- Generate a Uniformly distributed random variable ω in the range 0-1. The value *C* can be estimated from the observations (direct probe) and in our case it has been estimated to be 0.1202. If $\omega \leq C$, then return the number of *embedded objects* as zero. Otherwise continue with the next step.
- Generate a random number *N* from the Negative Binomial distribution. From the collected data it was observed that the maximum number of *embedded objects* in a page is 75. Also, there were about 6% of the cases where the *embedded objects* were located outside the server hosting the *main object*. Using these observations, the value of the parameters of the distribution obtained via the maximum likelihood estimation are $\alpha = 1.386$ and $\beta = 6.985$.

Fig. 6 shows the observed CDF of the number of *embedded objects* (as collected from our real Web pages) against the distribution function obtained by the modeling process presented above.

7. SUMMARY

A characterization study of traffic collected at clients and servers for SMTP and HTTP has been presented. Real packet traces have been collected from different environments such as university networks and commercial Frame Relay networks. Statistical models have been developed for different parameters of applications. The modeling activity focuses on how the users typically access the applications, kind of demands they generate and the way these demands are handled. Both applications possess a session oriented structure. Within each session, a number of transactions are performed, and this number has been modeled as well.

REFERENCES

1. Arlitt M. and Williamson C., "Internet Web Servers: Workload Characterization and Performance Implications", *IEEE/ACM Transactions on Networking*, Vol 5, No 5, October 1997.
2. Barford P. and Crovella M., "Generating Representative Web Workloads for Network and Server Performance Evaluation", *ACM SIGMETRICS*, 1998.
3. D'Agostino R.B. and Stephens M.A., *Goodness-of-Fit Techniques*, Marcel Dekker Inc., 1986.
4. Jena A.K., *Modeling and Analysis of Internet Applications*, Lic. Thesis, Univ. Lund, Sweden, 2000.
5. Jena A.K. and Popescu A., "Traffic Engineering for Internet Applications", *The Conference on Internet Performance and QoS ITCom2001*, Denver, USA, August 2001.
6. Johnson N., Kotz S., and Balakrishnan N., *Continuous Univariate Distributions*, Vol. 1, John Wiley & Sons, 1994.
7. Kurien T.V., correspondence with T. V. Kurien (Niksun Inc.), 1999.
8. Mah B.A., "An Empirical Model of HTTP Network Traffic", *IEEE INFOCOM*, Kobe, Japan, 1997.
9. NIKSUN NetVCR^{TM}, http://niksun.com/products/netvcr.html
10. TCPTRACE, http://www.tcptrace.org
11. Paxson V., "Empirically-Derived Analytic Models of Wide-Area TCP Connections", *IEEE/ACM Transactions on Networking*, Vol. 2, No. 4, August 1994.
12. Xanthakis S., GetWeb version 2.7.2, http://www.enfin.com/getweb/.