Statistical Characterization of Wide-Area Self-Similar Network Traffic

Matthew T. Lucas † *Dallas E. Wrege* † *Bert J. Dempsey* ⋆ *Alfred C. Weaver* †

† Department of Computer Science
University of Virginia
Charlottesville, VA 22903
{matt, dallas, weaver}@Virginia.edu

⋆ Manning Hall, Campus Box 3360
School of Informaion and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
bert@ils.unc.edu

October 9, 1996

Abstract

Background traffic models are fundamental to packet-level network simulation since the background traffic impacts packet drop rates, queuing delays, end-to-end delay variation, and also determines available network bandwidth. In this paper, we present a statistical characterization of wide-area traffic based on a week-long trace of packets exchanged between a large campus network, a state-wide educational network, and a large Internet service provider. The results of this analysis can be used to provide a basis for modeling background load in simulations of wide-area packet-switched networks such as the Internet, contribute to understanding the fractal behavior of wide-area network utilization, and provide a benchmark to evaluate the accuracy of existing traffic models. The key findings of our study include the following: (1) both the aggregate and its component substreams exhibit significant long-range dependencies in agreement with other recent traffic studies, (2) the empirical probability distributions of packet arrivals are log-normally distributed, (3) packet sizes exhibit only short-term correlations, and (4) the packet size distribution and correlation structure are independent from both network utilization and time of day.

Key Words: Traffic Modelling, Traffic Characterization, Network Simulation, Statistical Characterization, Self-Similar Traffic.

University of Virginia Technical Report CS-96-21

1 Introduction

Simulation modeling of computer networks is a powerful technique for evaluating the design and performance of network, transport and application-level protocols. Background traffic models are a fundamental component of packet-level network simulators since the network load drives the packet drop rate, queuing delay, end-to-end delay variation, and available network throughput [6]. Developing background traffic models suitable for use in a large-scale, packet switched network simulation (e.g., an Internet backbone network simulator) is a difficult problem because wide-area traffic dynamics are not well understood. Characterizing wide-area traffic is difficult for several reasons: (1) backbone networks are often inaccessible for measurement and study, (2) the nature of Internet applications, user populations, and user demand is constantly changing, and (3) network traffic is shaped by network switches as well as end-system congestion control protocols. This paper focuses on characterizing the statistical properties of network traffic exchanged between the campus network at the University of Virginia, a state-wide educational network, and a large Internet service provider. The data and analysis presented in this study can be used to develop background traffic models, contribute to understanding the fractal behavior of wide-area network utilization, and provide a benchmark to evaluate the accuracy of existing models.

The analysis presented in this paper focuses on 90 minute traces taken from a week-long trace of packets leaving the UVA campus network during periods of low, medium, and high network utilizations. Section 2 of this paper describes how these packet traces were collected and gives an overview of the data. Section 3 presents the statistical properties of the aggregate packet stream generated by the UVA campus network. We present the density function and correlation structure of the packet sizes, and find that the packet sizes have only short-range dependencies with a density that is independent of the network load. We next study the properties of the packet arrivals. We demonstrate that the arrival density follows a log-normal distribution and exhibits significant long-range dependencies over the entire range of network utilization. In Section 4, we present density and arrival correlation analysis of substreams that result from partitioning the aggregate traffic stream into substreams based on destination IP addresses. We show that a small range of the class B and C address space comprises the majority of the aggregate traffic, and find that the larger substreams exhibit statistical properties similar to those of the aggregate stream. However, the substreams which contribute less that 3% of the aggregate packet stream exhibit little long-range dependencies and do not follow a log-normal arrival distribution. Conclusions and a discussion of parsimonious models that efficiently generate traffic loads consistent with the empirical findings are discussed in Section 5.

2 Empirical Traces of Wide-Area Traffic

The analysis presented in this paper is based on a week-long trace of nearly one billion IP packets exchanged between the University of Virginia's campus network (UVAnet), the Virginia Educational and Research Network (VERnet), and BBNplanet (at the time, UVA's Internet service provider). The network monitor used to collect the trace consists of a powerful workstation[1], a kernel customized to have large network buffers, and a kernel-level packet filter [1]. The network monitor provides a timestamp resolution within 100μsec and an observed drop rate of less than 0.001% over the entire trace.[2]

Figure 1 shows the experimental setup. As shown in the figure, three routers and the network monitor are connected to a single Ethernet hub. The VERnet and BBNplanet routers are each connected to three T1 links, while the UVAnet router is connected to UVA's backbone FDDI concentrator. The filter is configured to listen promiscuously on the Ethernet and capture all IP packets sent between the UVAnet, VERNet and BBNPlanet routers. The filter captures the IP header and saves the IP source, IP destination, timestamp, and size of each packet to disk. After compression, approximately six bytes are saved per packet. The week-long packet trace (consisting of 6GB of data) is publicly available at [2].

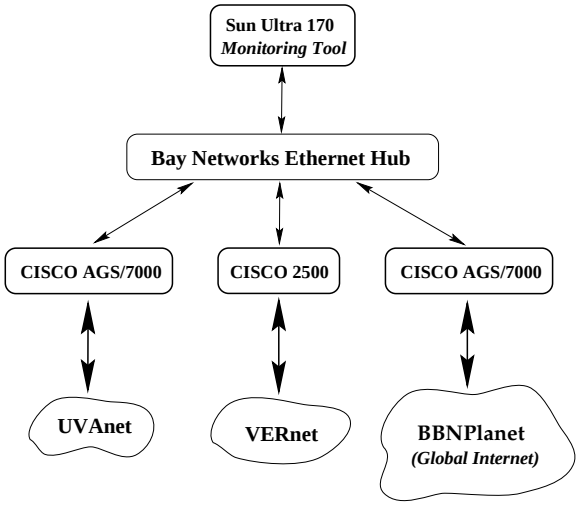


Figure 1: Experiment Setup

Figure 2 depicts the nine-day packet trace captured by the packet filter. The figure plots the number of packets exchanged between the three networks per 100-second interval as a function of

[1]Sun UltraSparc Model 170, 100MB RAM, 8GB HD running Solaris 2.5

[2]The drops reported by the filter occurred in isolated bursts.

time. There are two periods where the monitor workstation went off-line. The first period occurred between 8PM Wednesday and 8AM Thursday due to a disk problem, and the second failure occurred at 11PM on the second Tuesday due to a campus-wide power outage. Two interesting observations about the data are: (1) the ratio of the peak to the minimum data rate is approximately 8:1, which is bursty at this timescale, and (2) the packet rate is cyclical with periods of low utilization occurring around 5AM and peak utilization occurring around 4PM.

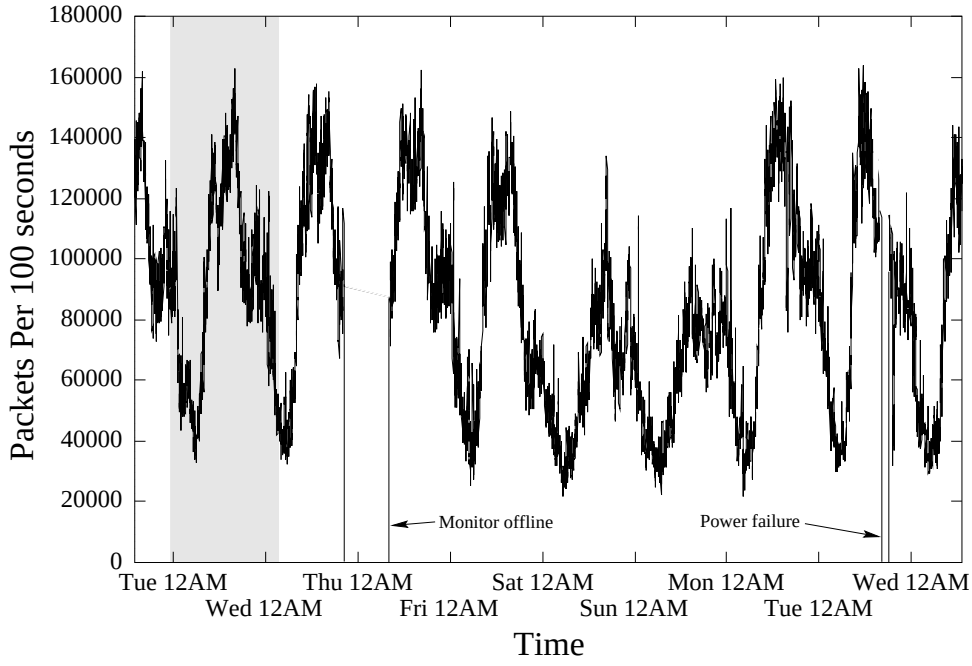


Figure 2: Packets per 100 seconds for 9 day packet trace.

3 Empirical Data Statistical Analysis

In our statistical analysis, we consider the packets leaving UVAnet destined for either BBNplanet or VERnet during the one-day period highlighted in Figure 2. We characterize outgoing packets since the end-goal of this research is to develop a model of packet arrivals generated at a wide-area backbone network access point (e.g., packet arrivals for a large campus network or small internet service provider).

Figure 3 depicts the number of packets generated by UVAnet per 10-second interval over the 27 hour trace. The network monitor experienced a single burst of drops during the 27 hour period when, just before 12PM, the monitor timed out for exactly ten seconds and dropped 9,784 packets. Our analysis focuses on the three 90 minute intervals highlighted in Figure 3. These intervals are

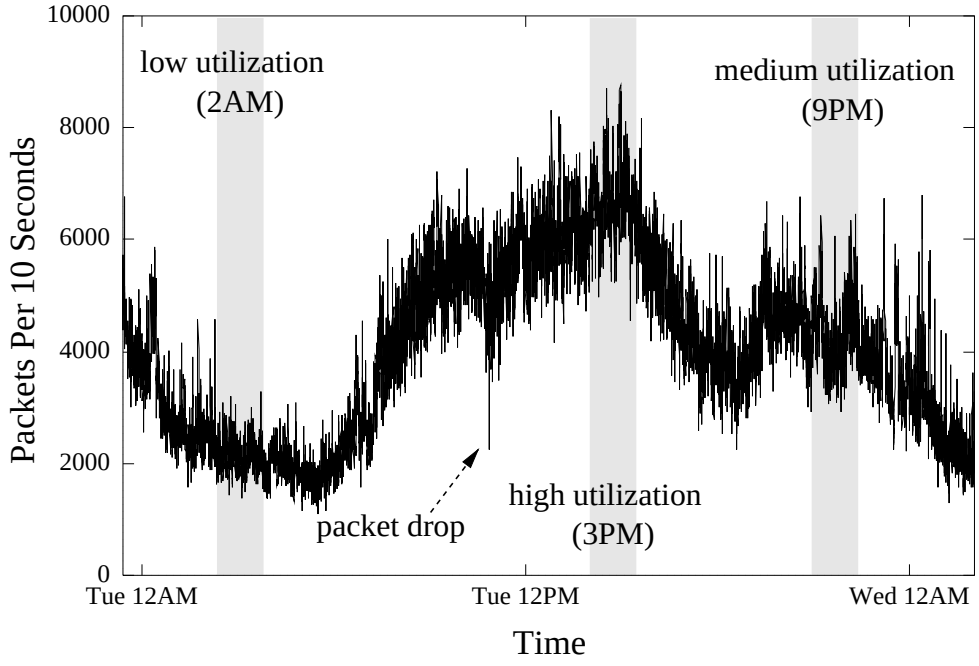


Figure 3: Packets per 10 second interval for Tuesday, June 4th, 1996 packet trace.

from 2:15AM – 3:45AM ("2AM trace"), 2:00PM – 3:30PM ("3PM trace"), and 9:00PM – 10:30PM ("9PM trace"). We selected these intervals because they correspond with low, high and medium network utilizations, respectively, and the arrival processes are stationary over these durations.

In the remainder of this section, we study the statistical characteristics of each traffic stream. We first examine the probability density and autocorrelation structure of the packet sizes in Section 3.1, and we then investigate the number of packet arrivals per unit time in Sections 3.2 and 3.3.

3.1 Packet size analysis

Figure 4 shows the empirical probability distribution of packet sizes for the 2AM, 3PM and 9PM traces. We plot the density on a logarithmic scale as a function of the packet size. We see in the figure that a small number of packet sizes dominate the trace. In particular, approximately 75% of the packets are either 40-44 byte or 552 byte packets. Inspection of the distribution also reveals pronounced "spikes" at $55, 60, 75, 144, 576$ and 1500 byte packets, accounting for 12% of the packets. A key observation in Figure 4 is that the densities are nearly identical for all three traces, suggesting that the distribution of packet sizes is independent of network utilization.

We next consider the correlation structure of the packet sizes. For a random process $\{X_i\}_{i=0,1,\ldots,N}$ with sample mean $\overline{X}$ and sample variance of S^2, the autocorrelation function r can be estimated

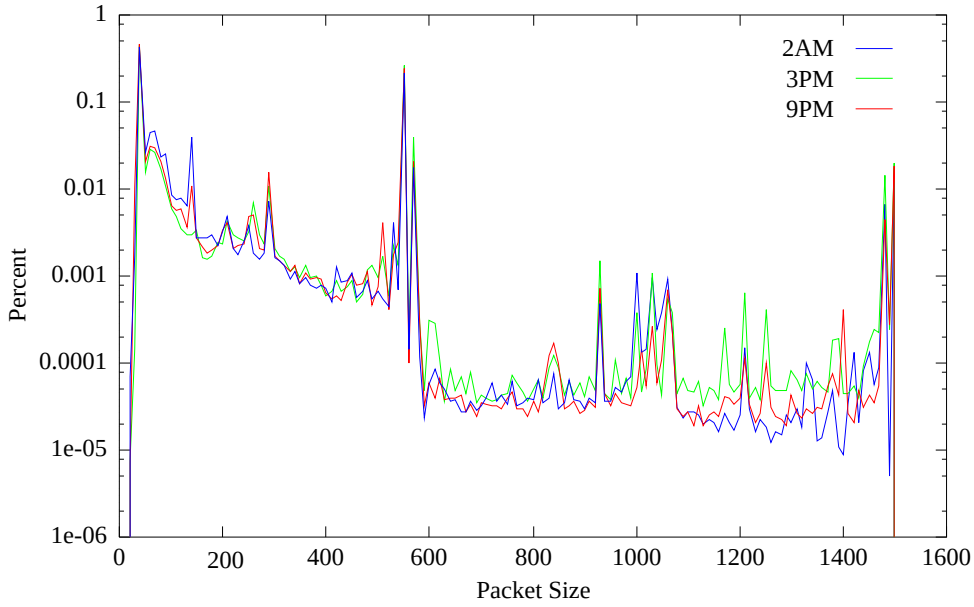


Figure 4: Probability density function of packet sizes.

for all lag k as follows:

$$r(k) = \frac{\sum_{i=1}^{N-k}(X_i - \overline{X})(X_{i+k} - \overline{X})}{(N-k)S^2} \quad (1)$$

Figure 5 shows the autocorrelation $r(k)$ of the packet sizes plotted as a function of the lag k for each trace. The figure shows that the correlation of packet sizes is significant only until a lag of 10 packets. Also observe that the correlation structure is similar for all three traces, indicating that packet size correlation is independent of network utilization and time of day. The correlation and density analysis suggest that packet sizes can be accurately modeled using a simple short-range dependent process in which packet sizes are selected directly from the empirical density function shown in Figure 4.

3.2 Packet arrival distribution analysis

Figure 6 plots the distribution of packet arrivals for UVAnet per 100ms interval. In the figure, solid lines give the empirical probability density of the traces, while the dashed lines show analytical log-normal distributions fit to the empirical densities using a maximum likelihood estimator (MLE). As the figure shows, the fit appears to be good except for a pronounced deviation at the distribution peaks. To better evaluate the goodness of the fit, we show Q-Q plots in Figure 7 that plot the quantiles of the empirical data against the quantiles of the fitted distribution. Figure 7 shows that the log-normal distribution is very good across the entire range of the distribution except for the

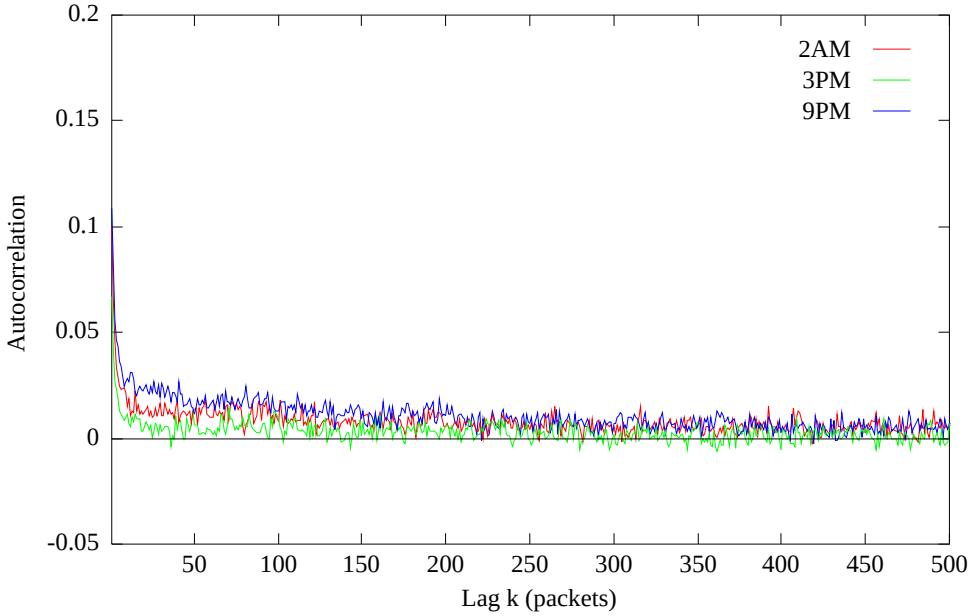


Figure 5: Autocorrelation function for packet sizes.

tail at the right where the log-normal approximation overestimates the empirical data.

3.3 Packet arrival correlation analysis

In this section we examine how packet arrivals are correlated over time. As we will see, the packet arrivals have long-range dependencies and are fractal in nature; that is, exhibit burstiness over multiple time scales (for a review of this subject, see [5]). Background traffic models must be able to model such long-range dependencies, otherwise simulated packet drop rates, variation in network transmission delay, and available network throughput may be severely underestimated.

An important class of processes that can model fractal traffic are so-called self-similar models such as fractional Gaussian noise[12] and fractional ARIMA processes[3]. The advantage to models which approximate fractional Gaussian noise[7, 9, 13] is that the the long-range dependent component can be characterized using a single parameter called the *Hurst parameter* H.

The important property of a self-similar arrival processes is that they exhibit "similar" burstiness properties independent of the time scale at which they are viewed. In particular, if we are given a stationary process $\{X_i\}_{i=0,1,\ldots,N}$, we consider its associated *aggregated arrival processes* $\{X_i^{(m)} \mid m = 1, 2, \ldots N\}$ which is given by:

$$X_i^{(m)} = 1/m \sum_{k=im}^{i(m+1)-1} X_k \tag{2}$$

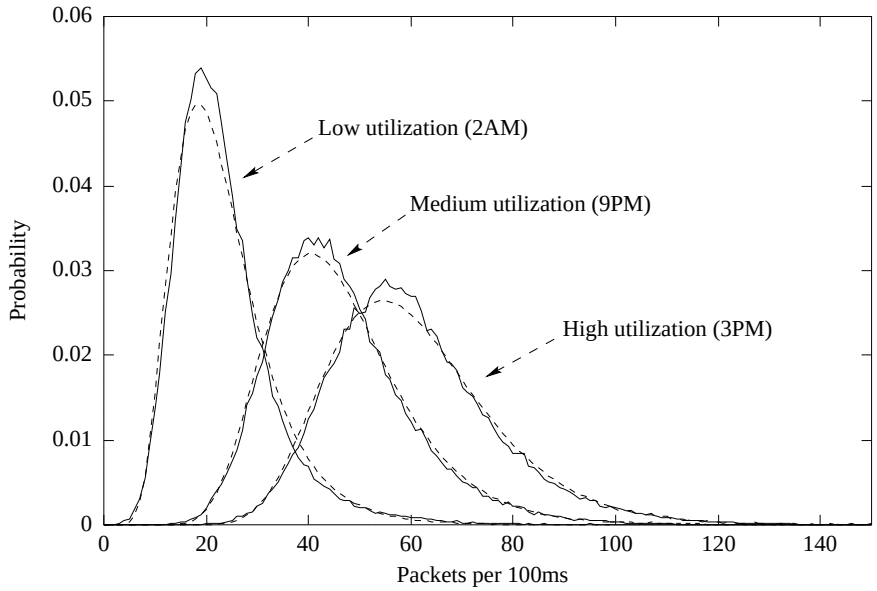
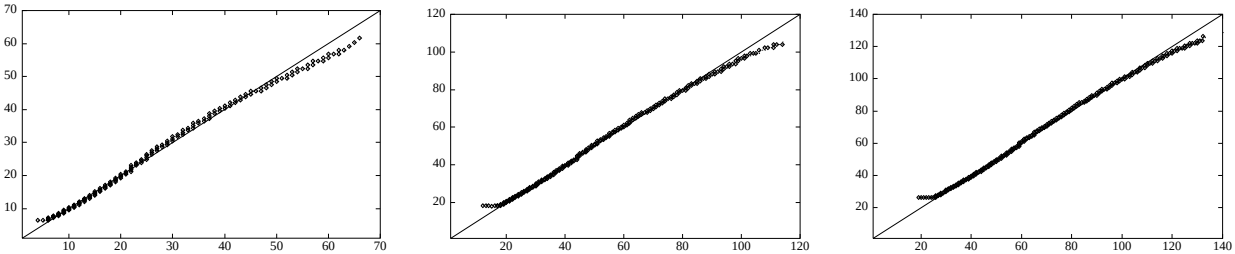


Figure 6: Histogram and log-normal fit of packet rates at low, medium, and high network utilizations. Empirical traces are shown as solid lines, while their log-normal approximations are depicted as dashed lines.



(a) 2AM trace. (b) 9PM trace. (c) 3PM Trace.

Figure 7: Q-Q plots of 2AM, 9PM and 3PM empirical traces versus fitted log-normal distributions.

A process $\{X_i\}$ with infinite samples is called *exactly second-order self-similar* if the autocorrelation $r^{(m)}(k)$ of each aggregated process is given by [11]:

$$r^{(m)}(k) = r(k), k \geq 0 \tag{3}$$

and the variance is given by [11]:

$$\text{Var}(X^{(m)}) = Var(X)m^{-2(1+H)} \tag{4}$$

The Hurst parameter H varies between 0.5 and 1, where a larger value indicates a higher degree of self-similarity. For a short-range dependent process, such as the Poisson-based models in [8, 14], the Hurst parameter will be approximately 0.5. Recent studies have estimated the Hurst parameter to be up to 0.82 for Ethernet LAN traffic [11].

Figure 8 shows the autocorrelation of packet arrivals per 10ms intervals as a function of the time lag (expressed in seconds). The solid lines give the empirical autocorrelation functions for each stream, while the dashed lines illustrate the autocorrelation function for processes that are exactly second-order self-similar with values of H ranging from 0.68 to 0.8.

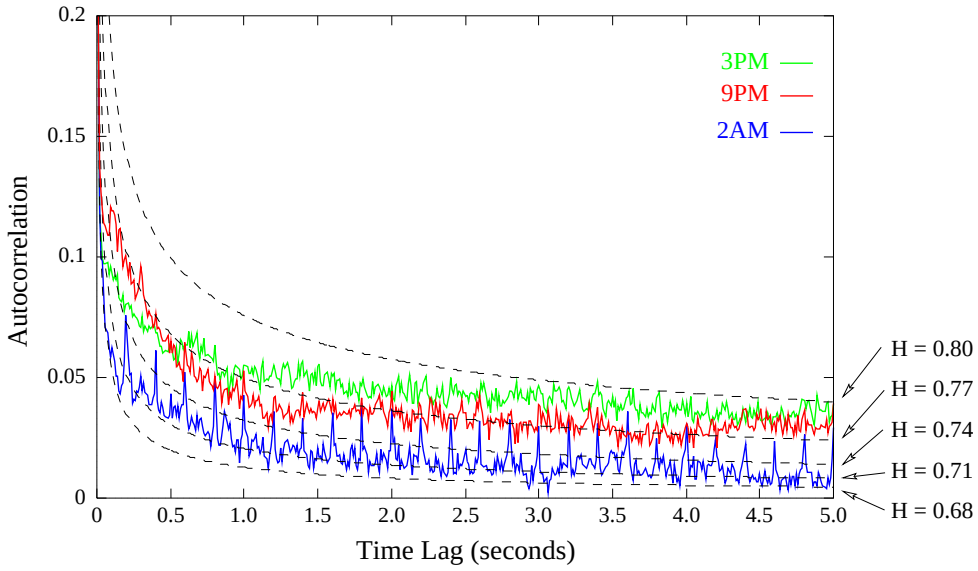


Figure 8: Autocorrelation.

The autocorrelation functions for the packet arrivals are hyperbolically decaying, suggesting that the arrival process is long-range dependent. Comparing the empirical functions with the self-similar reference curves, we see that the correlation structure of the empirical traces demonstrate long-range dependencies similar to that of Ethernet traffic.

In order to better evaluate the self-similar nature of the traffic, we next consider log-variance plots for the three traces in Figure 9. Log-variance plots show the degree of burstiness of an arrival process over multiple time scales by plotting the $\log_{10}$ of the normalized variance of the aggregated arrival process $X^{(m)}$ against $\log_{10}$(aggregation level m). For short-range dependent process, such as Poisson-based processes, $Var(X^{(m)})$ falls off as $1/M$. In contrast, Figure 9 shows that the variance of the arrivals for all three traces decay slowly in proportion to a self-similar process with $H = 0.65$ for small aggregation levels, and asymptotically as a self-similar process with $H = 0.8$. Using the semi-parametric algorithm developed in [10], we estimated the Hurst parameter for the 2AM trace = 0.70, 3PM trace = 0.71 and 9PM trace = 0.75.

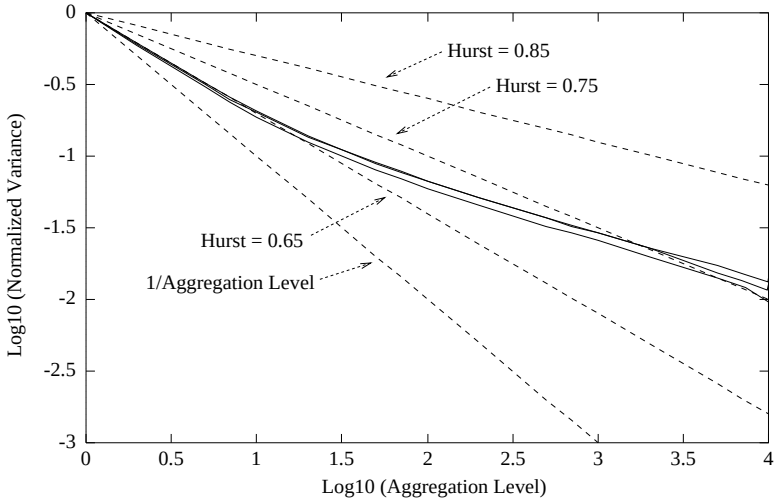


Figure 9: Log-variance of empirical aggregate traces.

4 Composite stream analysis

In the previous section we considered the traffic of an aggregate stream departing a campus network. Here we study the statistical properties of substreams obtained by partitioning the aggregate traffic along destination network addresses. Characterizing component substreams is important because background traffic models must not only construct an aggregate packet arrival process but also need to associate each packet with a destination campus address or network access point.

We divide the aggregate stream into 14 substreams based on their destination IP addresses (for a review of this area, refer to [4]). Table 1 gives the network mask used to define the component substreams and the percentage of packets each substream contributes to the aggregate stream. The network masks divide the aggregate stream such that the Class A address space (i.e., network addresses $<$ 128.0.0.0) and Class D/E address space (i.e., network addresses above 224.0.0.0) each

correspond to a substream, and the remaining 12 streams are created by partitioning the Class B and C address along bits 2-5. Note that each of the Class B/C streams is the same "size" with respect to number of network addresses.

There are several interesting observations with regard to the distribution of packets throughout the IP address space:

- Class A destinations accounts for less that 2% of the packet arrivals while consuming half of the total IP address space.
- Two of the Class C streams (i.e., those addresses in the range 192.0.0.0 – 207.255.255.255) account for 60% of the packets but consume only 1/16 of the IP address space.
- Half of the high order Class C (i.e., 208.0.0.0 – 223.255.255.255) and a quarter of the high order Class B (i.e., 176.0.0.0 – 191.255.255.255) address space had almost no arrivals. For this reason, we do not consider these substreams in the analysis that follows.

In the remainder of this section, we analyze the density, correlation structure, and long-range dependencies of substreams of the three traces described in Section 3.

Filter Mask	2AM Trace	9PM Trace	3PM Trace
0.0.0.0 – 127.255.255.255 (Class A)	1.6%	1.6%	1.7%
128.0.0.0 – 135.255.255.255 (Class B)	20%	20%	21%
136.0.0.0 – 143.255.255.255 (Class B)	6.9%	5.9%	3.9%
144.0.0.0 – 151.255.255.255 (Class B)	3.0%	3.0%	2.4%
152.0.0.0 – 159.255.255.255 (Class B)	4.2%	7.7%	6.3%
160.0.0.0 – 167.255.255.255 (Class B)	3.0%	3.0%	2.4%
168.0.0.0 – 175.255.255.255 (Class B)	0.6%	1.4%	1.0%
176.0.0.0 – 183.255.255.255 (Class B)	0.0%	0.0%	0.0%
184.0.0.0 – 191.255.255.255 (Class B)	0.0%	0.0%	0.0%
192.0.0.0 – 199.255.255.255 (Class C)	21%	26%	22%
200.0.0.0 – 207.255.255.255 (Class C)	40%	32%	39%
208.0.0.0 – 215.255.255.255 (Class C)	0.0%	0.1%	0.2%
216.0.0.0 – 223.255.255.255 (Class C)	0.0%	0.0%	0.0%
224.0.0.0 – 255.255.255.255 (Class D/E)	0.3%	0.1%	0.2%

Table 1: Network filter mask and percent of traffic for 2AM, 9PM and 3PM traces.

4.1 Density analysis of composite streams

Figure 10 shows the empirical probability density for several of the component substreams of the 3PM trace. For clarity of presentation, we include only the substreams which compose more than 3% of the aggregate stream. The solid lines depict the empirical streams, while the dashed lines illustrate an analytical log-normal distribution whose parameters were determined using the MLE. As the figure shows, the log-normal distribution provides a good fit for the streams with a larger mean. However, the log-normal fit does not accurately model substreams $152 - 160$ and $136 - 144$ because of the large number of samples with no arrivals.

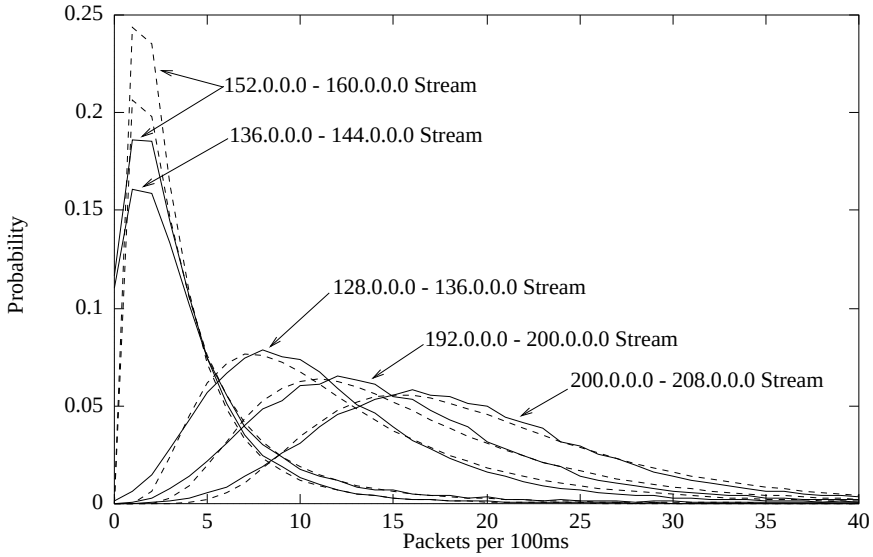
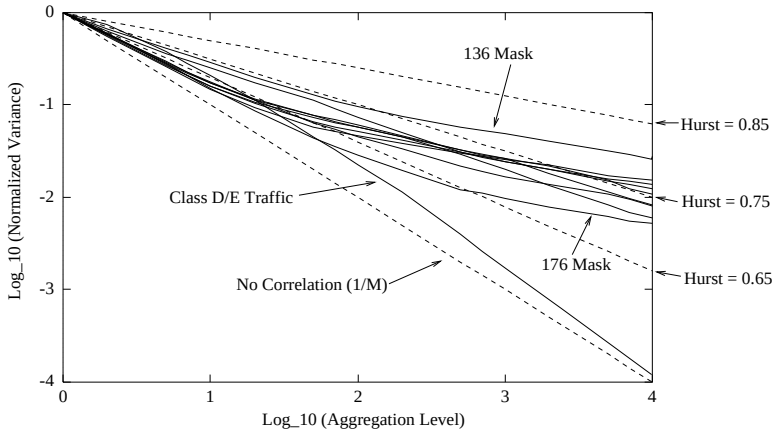


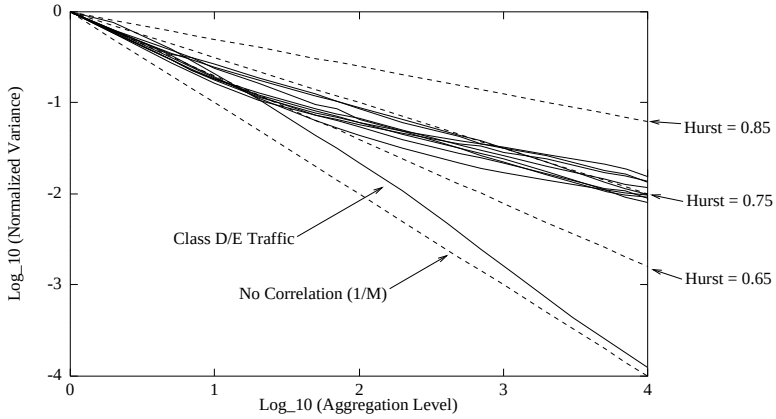
Figure 10: Packet arrival density for 3PM composite substreams.

4.2 Packet correlation analysis of composite streams

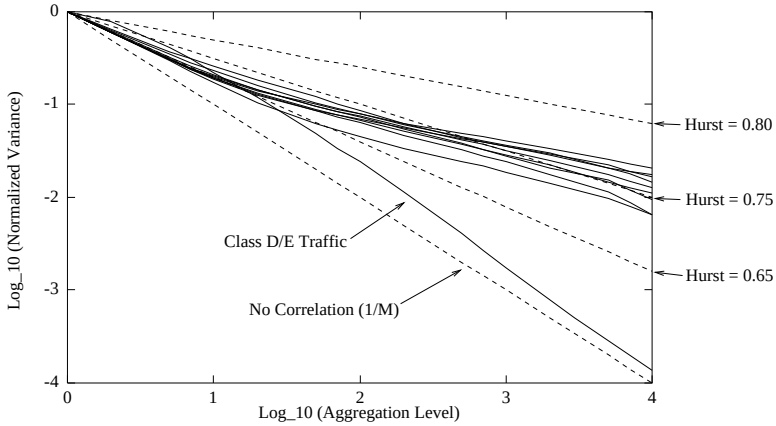
Figure 11 shows log-variance plots of the component substreams for each trace. In the 2AM trace, the degree of long-range dependence for each substream varies widely from no long-range dependence (e.g., the Class D/E substream) to an asymptotically self-similar process with $H = 0.8$. However, the 9PM and 3PM traces are very similar in that both Class D/E streams exhibit short-range dependencies, and the remaining substreams exhibit long-range dependencies similar to that of the aggregate streams shown in Figure 9. This data suggests that substreams with very low packet arrival rates (i.e., more than 3% of the 2AM trace, or more than 1% of the 3PM and 9PM traces), exhibit the same fractal properties as the aggregate streams.



(a) 2AM Trace - Low Utilization.



(b) 9PM Trace - Medium Utilization.



(c) 3PM Trace - High Utilization.

Figure 11: Log-variance plots of (a) 2AM, (b) 9PM and (c) 3PM component substreams.

5 Conclusions and Future Work

In this paper we presented the statistical characteristics of a week-long trace of packets exchanged between UVAnet, VERnet and BBNplanet. We focused on three representative 90 minute traces of packets leaving the UVA network. We first considered the distribution and correlation of packet sizes and found that the densities are nearly identical for all three traces, and are short-term correlated. Next, we considered the density and correlation structure of the arrivals for each trace. We showed that the arrival density can be modeled with log-normal distribution and the arrivals are self-similar exhibiting significant long-range dependencies similar to that of Ethernet traffic. Finally, we analyzed the component substreams of the aggregate traces. We showed that the component substreams which compose more than 3% of the aggregate stream are also log-normally distributed; however, the component substreams with very low packet arrivals do not exhibit significant long-range dependencies.

In future work we will look at creating an analytic traffic model which can faithfully generate traffic with the same statistical properties found in the traces presented here. Our approach is to first model the aggregate packet arrival process for the UVA campus. We use a self-similar process with a Hurst parameter that matches those found in the study to generate the long-range dependent traffic. Typically, the self-similar models generate traffic with a normal distribution, and so we use an appropriate exponential transformation that creates an aggregate stream with matching arrival density. The advantages of modeling the aggregate stream rather than individual campus streams are twofold: (1) since generating self-similar streams is computationally demanding, it is advantageous to generate only N aggregate streams rather than N^2 campus streams, and (2) background load characteristics can be changed by varying a single mean parameter and hurst parameter (as opposed to changing N mean parameters - one for each stream).

We are investigating several models which associate a destination address with each packet arrival. The simplest approach is to use the density of the destination network address to choose a network address. This approach, however, may distort the burstiness of the packet arrivals for each composite substream. The second approach is to use N Markov chains that have states corresponding with the arrival distributions of each individual substream. The destination of arrivals can then be chosen based on the state of each individual Markov process. This approach has the advantage of modeling both the aggregate and substreams such that the density of the arrivals match the empirical traces. The third model under consideration is to use the state of the Markov chain to drive the mean of a Poisson-based packet train model for each substream.

Finally, we will incorporate the most accurate of the models described above into a simulation of a large backbone network. We will evaluate the performance of the model by comparing the sim-

ulated packet drop rates, delays, and throughput to empirical statistics of ping messages collected during the week long trace.

References

[1] Snoop Packet Filter. Sun Solaris 2.5 man page. Sun Microsystems, 1996.

[2] UVA Packet Traces. Available via ftp at ftp.cs.virginia.edu in /pub/mtl8c.

[3] A. Adas and A. Mukherjee. On Resource Management and QoS Guarantees for Long Range Dependent Traffic. In *Proc. IEEE Infocom*, pages 779–787, April 1995.

[4] Douglas E. Comer. *Internetworking with TCP/IP, Second Edition.* Prentice Hall, 1991.

[5] D. R. Cox. Long-Range Dependence: A Review. In *Statistics: An Appraisal, Proc. 50th Anniversary Conference*, pages 55–74, Ames, IA: Iowa State University Press. Iowa State Univ. Press, 1984.

[6] V. Frost and B. Melamed. Traffic Modelling for Telecommunications Networks. *IEEE Communications Magazine*, 32(3):70–81, March 1994.

[7] M.W. Garrett and W. Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *ACM Sigcomm 1994*, London, UK, August 1994.

[8] H. Heffes and D. M. Lucantoni. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):856–868, September 1986.

[9] Wing-Cheong Lau, Ashok Erramilli, Jonathan L. Wang, and Walter Willinger. Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and Its Properties. In *Proc. IEEE ICC '95*, pages 466–472, Seattle, Washington, 1995. IEEE.

[10] Wing-Cheong Lau, Ashok Erramilli, Jonathan L. Wang, and Walter Willinger. Self-Similar Traffic Parameter Estimation: A Semi-Parametric Periodogram-Based Algorithm. In *Proc. IEEE Globecom '95*, Singapore, 1995.

[11] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[12] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian Motions, Fractional Noise and Applications. *SIAM Review*, 10:422–437, 1968.

[13] Vern Paxson. Fast Approximation of Self-Similar Network Traffic. Technical Report LBL-36750, Lawrence Berkeley Laboratory and EECS Division, University of California, Berkeley, April 1995.

[14] R.Jain and S. A. Routhier. Packet Trains: Measurements and a New Model for Computer Network Traffic. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):986–995, September 1986.