EECS598, Section 5
Paper Review of the Fifth Topic: Internet Traffic Engineering and Self-Similarity

Chih-fan Hsin
08507134

I. Overview

Recent measurements of local-area and wide-area traffic have shown that network traffic exhibits variability at a wide range of scales. Such scale-invariant variability is in strong contrast to traditional models of network traffic. These four papers study the causes and effects of traffic self-similarity.

II. Paper 1: On the Self-Similar Nature of Ethernet Traffic

A. Introduction

The main object of this paper is to establish in a statistically rigorous manner the self-similarity characteristic of the very high quality, high time-resolution Ethernet LAN traffic measurements in the LAN traffic data collected by an earlier paper. Using the degree of self-similarity as a measure of burstiness, [1] shows that the burstiness of LAN traffic typically intensifies as the number of active traffic sources increases.

First, [1] introduces different manifestations, such as slowly decaying variances, long-range dependence, and a spectral density obeying a power-law, of the property of the covariance stationary process X, namely that X is asymptotically or exactly second-order self-similar. [1] also introduces three methods, which are the R/S plot, the variance-time plot, and the periodogram-based analysis, to estimate the degree of self-similarity. Then, [1] uses trace data collected to explore the properties of self-similarity.

Four data sets are considered in this paper. The first two sets were from an Ethernet network serving a laboratory of researchers. However, the second one was taken after the Sun-3 class workstations were replaced with RISC-based workstations. Because of the increased computing power of the machines connected to this segment, the network load increased appreciably. The third data set came from an Ethernet cable that linked the two wings of the facility that were occupied by a laboratory. It was also the segment serving links to the outside Internet world. The last data set was taken from the building-wide Ethernet backbone. The overall traffic levels were much lower than for the other three sets. The traffic is primarily router to router rather than host to host.

B. Results from Collected Trace Data

Ethernet traffic over approximately a 24-hour period is self-similar, with the degree of self-similarity increasing as the utilization of the Ethernet increases.

For traffic over a four-year period, [1] uses the maximum likelihood type estimates to estimate H and considers a 95%-confidence interval. [1] shows that Ethernet traffic during normal traffic hours can be considered to be exactly self-similar rather than asymptotically self-similar. For long time series considered, graphical estimation methods based on R/S or variance-time plots can be expected to be very accurate. The Hurst parameter corresponding to the typical normal and busy hours, respectively, are comparable with slightly higher H-values for the busy hours than for the normal traffic hours. The aggregate traffic tends to become less smooth as the number of active sources increases. The results also show some typical features of nonstationarity, which are characteristic of stationary long-range dependent process.

For external ethernet traffic, in terms of the self-similar nature, external traffic does not differ from the internal traffic. A more significant change in the Hurst parameter occurs during the low traffic hours. While the internal traffic data yield an H about 0.7, the external traffic data yield and H about 0.5. This means conventionally used short-range dependent models seem to describe the data accurately for external ethernet traffic during the low traffic hours.

C. Engineering for Self-Similar Network Traffic

To create the self-similarity, [1] suggests that a simple renewal reward process is an adequate traffic source model for an individual Ethernet user.

For measuring "burstiness" for self-similar network traffic, [1] claims that the commonly used measures such as the index of dispersion, the peak-to-mean-ratio, or the coefficient of variation are no longer meaningful for self-similar traffic but can be replaced by the Hurst parameter.

[1] also discusses two methods for generating asymptotically self-similar observations. The first one simulates the buffer occupancy in an $M/G/\infty$ queue, where the service time distribution G satisfies the heavy-tail condition. An infinite variance service time distribution results in an asymptotically self-similar buffer occupancy process. The second method exploits a convergence result that when aggregating many simple AR(1)-processes, where the AR(1) parameters are chosen from a beta-distribution on [0,1] with shape parameters p and q, then the superposition process is asymptotically self-similar.

D. Some Thoughts

In order to generate synthetic traces of self-similar traffic, exactly self-similar models such as fractional Gaussian noise, or some nonlinear transformation of fractional Gaussian noise or asymptotically self-similar models such fractional ARIMA processes can be used. However, parameter estimation techniques for these models often turn out to be computationally too intensive in order to work for large data sets. Therefore, [1] discusses two computationally unintensive methods, $M/G/\infty$ and AR(1), to generate synthetic traces.

III. Paper 2: On the Relationship Between File Sizes, Transport Protocols, and Self-Similar Network Traffic

A. Introduction

[2] examines a mechanism that gives rise to self-similar network traffic and presents some of its performance implications. The mechanism is the transfer of files or messages whose size is drawn from a heavy-tailed distribution. The mechanism [2] proposes is motivated by the ON/OFF model. The ON/OFF model shows that self-similarity can arise in an idealized context as a result of aggregating a large number of 0/1 renewal processes whose ON or OFF periods are heavy-tailed.

The network model is given by a directed graph consisting of n nodes and m links. A node is a server node if it has a probability density function $p_i(X)$ where $X \geq 0$ is a random variable denoting file size. This probability density function is a Pareto distribution. $p(x) = \alpha k^{\alpha} x^{-\alpha-1}$. A node is a client node if it has two probability density functions $h_i(X),\ d_i(Y),\ X \in \{1, ..., n\},\ Y \in \mathbf{R}_+$, where h_i is used to select a server, and d_i is the inter-arrival time (or the idle time) distribution. Based on this model, [2] uses *ns* simulator to obtain its results.

B. Results

[2] uses the variance-time plot and the R/S plot to show the following result. The degree to which file sizes are heavy-tailed can directly determine the degree of traffic self-similarity at the link level. This means that there is a nearly linear relationship, which is close to $H = (3-\alpha)/2$, between H and α over a wide range of network conditions. This causal relationship is robust with respect to changes in network resources (bottleneck bandwidth and buffer capacity), network topology, the influence of cross-traffic, and the distribution of inter-arrival times.

Then, [2] uses Pareto distribution to model the idle time distribution instead of using exponential distribution used before. For file size distribution with α close to 2, the correlation structure introduced by heavy-tailed idle time is significant relative to the contribution of the file size distribution. As α approaches 1, however, the tail weight of the file size distribution becomes the dominating term, and the contribution of idle time with respect to increasing dependency is insignificant in comparison.

Properties of the transport layer also play an important role. The reliable transmission and flow control mechanisms of TCP serve to maintain the log-range dependency structure induced by heavy-tailed file size distributions. In contrast, if a non-flow-controlled and unreliable (UDP-based) transport protocol is used, the resulting traffic shows little self-similarity: although still bursty at short time scales, it has little long-range dependence.

Increased self-similarity, results in degradation of performance. Queuing delay exhibits a drastic increase with increasing self-similarity. Throughput-related measures such as packet loss and retransmission rate increase only gradually with increasing traffic self-similarity as long as reliable, flow-controlled transport protocol is used.

C. Some Thoughts

The simulation is based on the heavy-tailed distributions. Therefore, the simulation shows the properties of self-similarity. However, it does not mean that a real network has these properties. [2] can use the trace of the real network traffic to show that its simulation results are correct.

IV. Paper 3: On the Relevance of Long Range Dependence in Network Traffic

There is much experimental evidence that network traffic processes exhibit ubiquitous properties of self-similarity and long range dependence (LRD). [3] introduces and evaluates such a network traffic model in which important parameters, namely the finite range of time scales of interest and the marginal distribution of the process, can be controlled. The model in [3] is a modulated fluid traffic model in which the correlation function of the fluid rate matches that of an asymptotically second-order self-similar process with given Hurst parameter up to an arbitrary cutoff time lag, then drops to zero.

A. Model description

The source traffic model is described by a random process $\{X_t\}$ which represents the fluid rate at time t. $X_t = \lambda(n)$ for $\tau_n \leq t < \tau_{n+1}$. The inter-arrival times $T_n = \tau_{n+1} - \tau_n$ are i.i.d. with ccdf $Pr\{T_n > t\} = F_T(t)$. The constant fluid rate $\lambda(n)$ is i.i.d. with distribution $Pr\{\lambda(n) = \lambda_i\} = \pi_i$.

[3] considers a truncated Pareto distribution as $F_T(t)$. $F_T(t) = (\frac{t+\theta}{\theta})^{-\alpha}$ if $t < T_c$ and 0 otherwise. T_c is the cutoff lag. There is no correlation in the fluid rate process beyond lag T_c. Based on this distribution, [3] derives the auto-covariance function of X_t, which is defined by $\phi(t) = E[(X_0 - \mu)(X_t - \mu)]$ where $\mu = \Pi\Lambda 1^T$. $\phi(t)$ behaves asymptotically as $t^{-\alpha+1}$ when $T_c \to \infty$. Thus, when $T_c = \infty$, $\{X_t\}$ is asymptotically second-order self-similar with Hurst parameter H such that $-\alpha + 1 = -(2 - 2H)$. When T_c is finite, the correlation drops to zero at lag T_c and $\{X_t\}$ has no LRD. This source model allows people to control the marginal fluid distribution Π, the time-scale T_c, and H.

Next [3] examines the performance of a queue with constant service rate c and a finite buffer B fed with the fluid source model. Let $Q(n)$ be the continuous random variable describing the queue occupancy at arrival instant τ_n, and

let $W(n) = T_n \times (\lambda(n) - c)$. The continuous random random variable $W(n)$ is the difference between arriving and departing work in inter-arrival interval n. The stationary density $f_Q(x)$ of the queue occupancy is derived and can be numerically solved.

Then let the random variable $W_l = (W - (B - Q))^+$ represents the amount of lost work in an inter-arrival interval. The long-term stationary loss rate defined as the ration of work lost to work arriving is $l = \frac{E[W_l]}{\overline{\lambda} E[T]}$. [3] derives a lower bound and an upper bound for l. These bounds are shown to converge rapidly to the limiting occupancy.

B. Numerical results

The loss rates considered are the average of the upper and lower bounds. The parameters of interest are the buffer size B, the cutoff lag T_c, and marginal fluid rate distribution Π.

For each buffer size, the loss rate is not significantly affected if the cutoff lag increases beyond some value. [3] refers to this value as the correlation horizon (CH). The rate at which the loss rate decreases as the buffer size increases depends on the value of the cutoff lag. For small cutoff lags, the decrease is approximately exponential. However, as the cutoff lag increases, the rate of decrease actually decreases. The simulation results are used to check these numerical results.

Next [3] considers two transformations of the marginal distribution. The first one scales the density of the marginal distribution by a constant factor α. The second one convolves the original distribution n times and renormalizes it to the original mean. The results show that the impact of the transformations on the loss rate for the range of parameters considered is much greater than that of the Hurst parameter. Small changes in the marginal scaling factor α yields dramatic changes in the loss rate.

For the numerical results above, [3] claims that the marginal distribution is a crucial parameter and it must be taken into account for accurate loss prediction. Furthermore, superposing even a moderate number of streams sharply decreases the loss rate, indicates that statistical multiplexing is an efficient mechanism to achieve high utilization while keeping loss low. Large buffers are helpful to significantly reduce the loss rate only for short-range dependent traffic.

[3] also describes a simple way to estimate the correlation horizon in terms of the various system parameters.

C. Some Thoughts

The cutoff lag and the marginal fluid rate distribution are not really the control parameters. We can not adjust theses values in a real application.

This model is not well-suited for sources with separate structures for the short term and long term correlation.

The literature on Markov modeling reports good performance prediction for finite buffer systems even when input traffic streams are correlated over many time-scales. [3] shows that there exists a correlation horizon which separates relevant and irrelevant correlation with respect to the performance measure of interest.

Trace generation based on self-similar models must also take into account statistics such as the marginal distribution in order to achieve accurate performance prediction.

V. Paper 4: Variance of Aggregated Web Traffic

A. Introduction

If data traffic were Poisson, increases in the amount of traffic aggregated on a network would rapidly decreases the relative size of bursts. The discovery of pervasive long-range dependence demonstrates that real network traffic is burstier than any possible Poisson model. [4] presents evidence that, despite being non-Poisson, aggregating Web traffic causes it to smooth out as rapidly as Poisson traffic. That is, the relationship between changes in mean bandwidth and changes in variance is the same for Web traffic as it is for Poisson traffic.

[4] considers only Web traffic. The Web traffic from traces of network traffic is self similar. From the traces, [4] shows that the variance of aggregated bandwidth scales almost linearly with the bandwidth. That is, the standard deviation scales with the square root of the total bandwidth. This is the same scaling behavior that Poisson traffic exhibits with respect to aggregation, and suggests that Web traffic is likely to smooth out quickly under increasing aggregation. The explanation is given.

B. Explanations

The traces show that a change in the number of users affects the variance in a way that depends on the correlation between users. Changes in the per-user bandwidth also affect per-user variance, which in turn changes the aggregate variance.

[4] uses some statistics from the traces to show that there is very little correlation among users and the bandwidth distribution of the aggregate is close to normal. Therefore, the variance of aggregate traffic equals the sum of the variances of the individual users. This lack of correlation helps to explain why the variance of aggregated bandwidth changes linearly with the bandwidth, and implies that Web traffic aggregates much like Poisson traffic.

C. Per-User Variance

Traces show that per-user variance also demonstrates an almost linear relationship to per-user mean bandwidth. [4] uses a mathematical model to show this relationship.

First, consider the correlations of ON/OFF/transfer size and the user average bandwidth. The traces show that variation in user bandwidth is mainly caused by different OFF periods and that the average ON periods and transfer sizes tend to have less effect on the bandwidth for most users. Based on this discovery, [4] gets $V = cB$ where V is the variance and B is a user's average bandwidth of a cycle. Therefore, the variance of a user's bandwidth is mostly proportional to the user's average bandwidth.

D. Some Thoughts

[4] considers only cases in which the traffic stream encounters no common bottlenecks. IF there are bottlenecks, I expect that Web traffic aggregates unlike Poisson traffic.

The results in this paper apply only to interactive Web traffic; they are only useful on links engineered to avoid persistent queuing; and they are only useful within the context of a busy hour.

VI. Comparison of Four Papers

[1], [2], [3] show the self-similarity manner in network traffic. However, [4] shows that Web traffic aggregates in a Poisson-like well behaved manner. Actually, [1] also shows that for external traffic during low traffic hours short-range dependent modes seem to describe the data accurately.

[2] uses the variance-time plot and the R/S plot to test and estimate H. [1], [3] uses the variance-time plot, the R/S plot, and the periodogram plot to test and estimate H.

[2] considers the transfer of files whose size is drawn from a heavy-tailed distribution. [3] considers the rate changes whose inter-arrival time is drawn form a heavy-tailed distribution, a truncated Pareto distribution.

[2], [3] starts from the view of the generation of self-similarity, which means a heavy-tailed distribution. [4] starts from the view of the property of self-similarity, the long-range dependence nature.

References

[1] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. *On the Self-Similar Nature of Ethernet Traffic.*IEEE/ACM Trans. Networking, Vol. 2, No. 1, 1994.

[2] K. Park, G.T. Kim, and M.E. Crovella. *On the relationship between file sizes, transport protocols, and self-similar network traffic.* Proc. ICNP, 1996.

[3] M. Crossglauser and J. Bolot. *On the relevance of long range dependence in network traffic.* IEEE Trans. Networking, 1998.

[4] R. Morris and D. Lin. *Variance of aggregated Web traffic.* INFOCOM 2000.