On the Contribution of TCP to the Self-Similarity of Network Traffic *

Biplab Sikdar and Kenneth S. Vastola

Department of ECSE, Rensselaer Polytechnic Institute
Troy, NY 12180 USA
{bsikdar,vastola}@networks.ecse.rpi.edu

Abstract. Recent research has shown the presence of self-similarity in TCP traffic which is unaffected by the application level and human factors. This suggests the presence of protocol level contributions to network traffic self-similarity, at least in certain time scales where the effect of protocol behavior is most prominent. In this paper we show how TCP's retransmission and congestion control mechanism contributes to the self-similarity of aggregate TCP flows. We develop a mathematical formulation which shows that TCP's retransmission and congestion control mechanism results in packet dynamics of a TCP flow being analogous to a number of ON/OFF sources with OFF periods taken from a heavy tailed distribution. Using well known limit theorems, we then show that this contributes to the self-similar nature of TCP traffic. Our model shows a direct correlation of the loss rates to the degree of self-similarity. Measurements on traces collected by us also exhibit this relationship predicted by our model.

1 Introduction

Research on the causes of self-similarity in network traffic have primarily focused on the application level characteristics of high-speed networks and the human factors involved. In [21], the causes of the self-similarity are investigated at the source level. In [1] the authors cite the distribution of file sizes, the effects of caching and human factors like response time and preference as possible causes for the self-similarity in WWW traffic. On the other hand, protocol level causes of self-similarity in network traffic has been investigated in [2] and [13] which showed that closed loop protocols like TCP lead to much richer scaling behavior than open loop protocols like UDP.

In this paper we show that TCP can contribute to the self-similarity of network traffic and its contribution is visible in the time scales ranging from milliseconds to tens of seconds. Thus though TCP may not be able to contribute at higher time scales, the observed self-similarity in these scales can be attributed to application and human level causes which inherently operate at time scales of

* Supported in part by DARPA contract F30602-00-2-0537 and in part by DoD MURI contract F49620-97-1-0382.

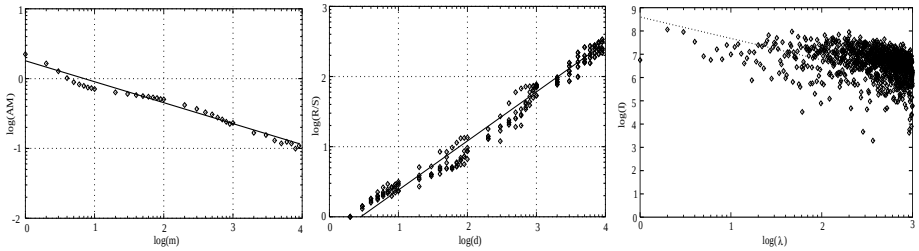
minutes and hours. Also, though in the pure mathematical sense a self-similar process should exhibit the same statistical characteristics over all possible time scales, this is not possible in real systems due to physical limitations. We show that TCP is capable of causing scaling in 3 to 4 time scales (few milliseconds to 10s of seconds) and it is in this sense that we call TCP traffic is self-similar. This range of timescales is generally sufficient for traffic modeling purposes as shown in citeGrBo99, since the range of relevant timescales is determined by the finite buffer sizes of real systems.

In [19], the authors attribute the self-similarity of TCP traffic to the chaotic nature of TCP's congestion control mechanism. The adaptive nature of TCP's congestion control is suggested as the cause for the propagation of self-similarity in the Internet in [20]. The main aim of our paper is to understand the effects of TCP's retransmission and congestion control mechanism on the observed self-similarity of TCP traffic. Our results show that the timeout and exponential backoff mechanisms in TCP play a crucial in inducing self-similarity. We also show that the degree of self-similarity has a direct relationship with the losses experienced by a flow with the traffic no longer self-similar, i.e. $H \approx 0.5$ for very low loss rates. While similar phenomena have been reported recently (after this paper was completed), their models to explain the self-similarity either require unrealistic loss rates to induce self-similarity [7] or are able to show long-range dependence over very small time scales [5]. In this paper, we present a model of TCP based on ON/OFF processes which explains the self-similarity of TCP traffic and validate it using TCP traces collected from the Internet. We also give a mathematical formulation of how TCP's congestion control mechanism leads to self-similarity in the traffic it generates and account for the effects of the network in terms of the loss probabilities and the presence of other flows.

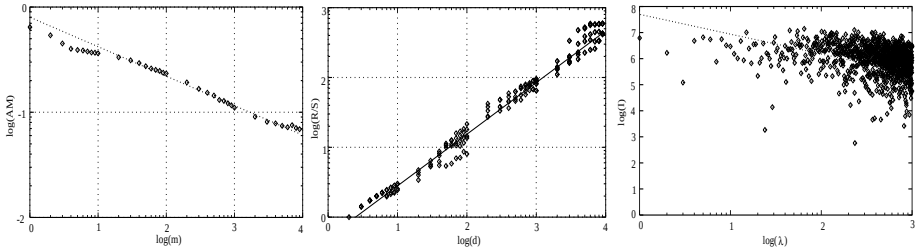
The rest of the paper is organized as follows. In Section 2 we first present the results of tests on traffic traces generated by individual TCP transfers over the Internet showing proof of self-similarity. We then present a model which explains this self-similarity and experimentally validate our model using the same TCP traces. In Section 3 we provide a mathematical foundation for our model and investigate the mechanisms of TCP which contribute to self-similarity in greater detail. Finally, Section 4 presents the discussions and concluding remarks.

2 Self-similarity of TCP Flows

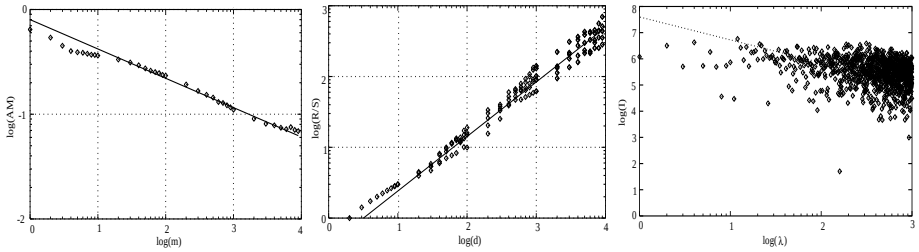
In this section we provide experimental evidence of the self-similarity of individual TCP flows which motivates the investigation of TCP dynamics for causes of self-similarity. In [19] the authors showed that the data sent by an isolated TCP flow from the superposition of a number of TCP flows shows evidence of self-similarity and attribute it to the chaotic nature of TCP's congestion avoidance mechanism. All previous reports of self-similarity in network traffic concentrated on the self-similar characteristics of the aggregated traffic. However, the results in [19] were generated by carrying out experiments using the simulator *ns* which is not an exact reflection of the actual scenarios in the Internet. Hence to dis-



(a) Loss prob = 0.010, $H = 0.70 \pm 0.01$



(b) Loss prob = 0.078, $H = 0.72 \pm 0.05$

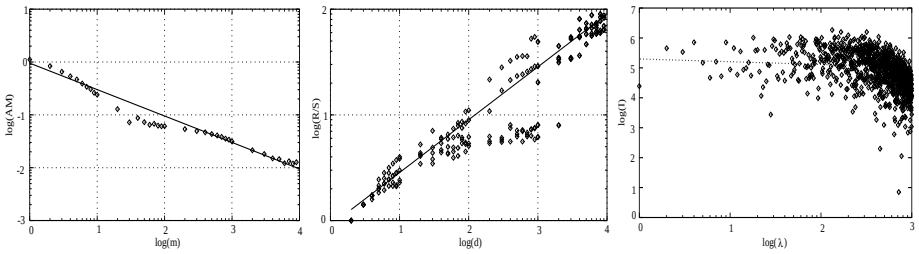


(c) Loss prob = 0.135, $H = 0.76 \pm 0.05$

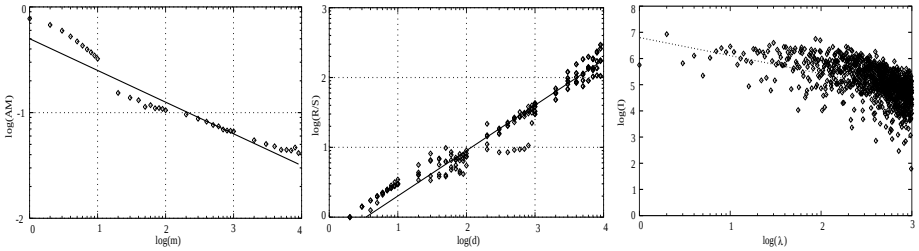
Fig. 1. Tests for self-similarity for the various traces to Columbus, Ohio. For each trace, the figures show the results from the absolute value method (left), R/S statistics method (middle) and the Periodogram method (right).

pel any doubts about the self-similar nature of single TCP microflows, we first present the results from tests for long-range dependence on traces collected from real life TCP connections over the Internet.

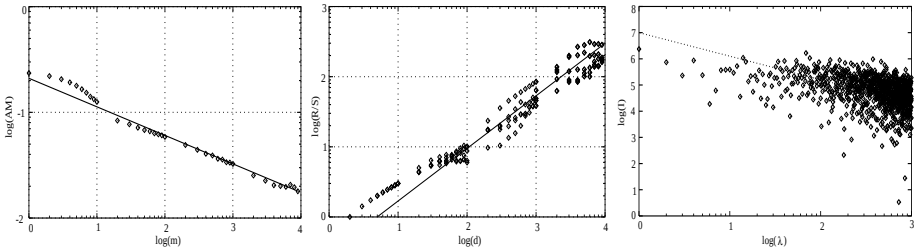
We first give a brief description of the datasets. We collected traces for data transfers originating from a machine running Solaris 2.6 at RPI, Troy, NY. The destinations for the transfers were in Ohio State University, Columbus, OH (HP-UX), University of California, Los Angeles, CA (FreeBSD Cairn-2.5), Massachusetts Institute of Technology, Boston, MA (Linux 2.0.36) and University of Pisa, Pisa, Italy (FreeBSD 3.3). Due to space restrictions, we show results



(a) Loss prob = 0.001, $H = 0.51 \pm 0.01$



(b) Loss prob = 0.006, $H = 0.67 \pm 0.03$

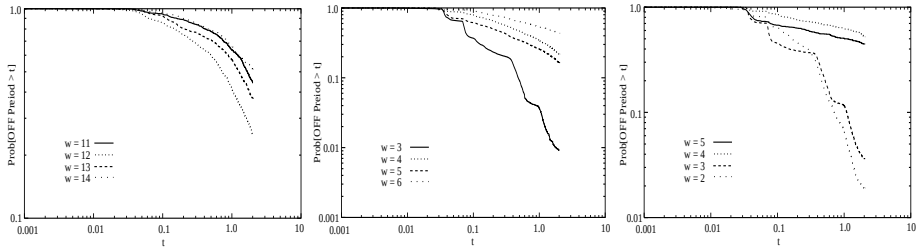


(c) Loss prob = 0.099, $H = 0.73 \pm 0.03$

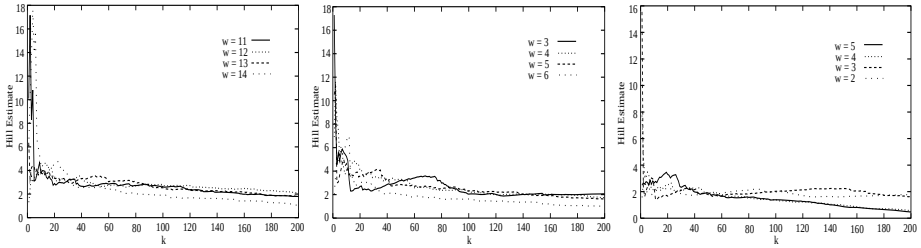
Fig. 2. Tests for self-similarity for the various traces to Pisa, Italy. For each trace, the figures show the results from the absolute value method (left), R/S statistics method (middle) and the Periodogram method (right).

for only the transfers to Ohio and Italy. The results for the others are similar. Each trace is 2000 seconds or around 33 minutes long and was collected using `tcpdump` which did not lose any packets. The transfers were done over periods in 1999 and 2000 at various times of the day and week. Depending on the prevalent network conditions, the loss rates experienced by each flow is different and we use this to classify transfers between a source-destination pair.

Figure 1 shows the results of the tests for long-range dependence on three traces to Ohio which had loss rates of 0.010, 0.078 and 0.135. Figure 2 shows the results of similar tests on the traces collected from transfers to Pisa which



(a) ccdf: $p = 0.010$ (b) ccdf: $p = 0.078$ (c) ccdf: $p = 0.135$



(a) Hill: $p = 0.010$ (b) Hill: $p = 0.078$ (c) Hill: $p = 0.135$

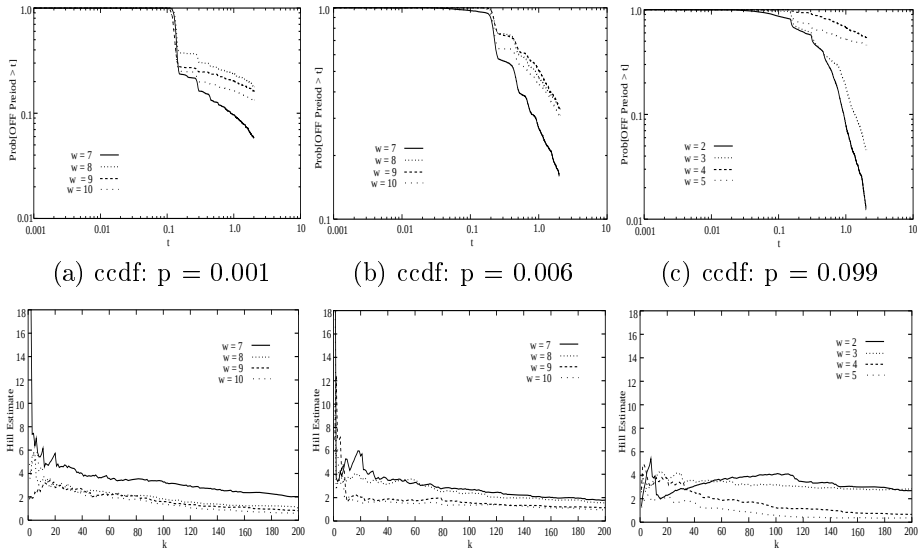
Fig. 3. Tests for heavy-tailed nature of the OFF times for the various traces to Columbus, Ohio. For each trace, the figures show the ccdf plots (top) and the corresponding Hill's estimates (bottom) for various values of w.

had loss rates of 0.001, 0.006 and 0.099. We tested for long-range dependence using three of the widely used methods [18]: the absolute value method, R/S statistics method and the periodogram method. The results clearly show the long-range dependence in the individual TCP flows. Also the degree of long-range dependence, as indicated by the Hurst parameter, is clearly dependent on the loss rate experienced by the flow, with higher loss rates leading to larger values of H. Also note that for extremely low probabilities (less than 0.001) the traffic is no longer self-similar as indicated by the Hurst parameter of approximately 0.5 as shown in section (a) of Fig. 2. We describe this in detail in the following subsection and in Section 4.

This poses the following questions. What are the underlying mechanisms which are responsible for the direct influence of the loss probabilities on the self-similarity of TCP traffic? What role does TCP's fast-retransmit and timeout mechanisms play in all this? In this paper we address these issues and show how TCP's retransmission and congestion avoidance mechanisms contribute to the self-similar nature of network traffic.

2.1 ON/OFF Model Based Explanation and its Validation

TCP follows a window based flow control mechanism and transmits a certain number of packets in each "round". We define a round as in [12]. A round begins



(a) ccdf: p = 0.001 (b) ccdf: p = 0.006 (c) ccdf: p = 0.099

(a) Hill: p = 0.001 (b) Hill: p = 0.006 (c) Hill: p = 0.099

Fig. 4. Tests for heavy-tailed nature of the OFF times for the various traces to Pisa, Italy. For each trace, the figures show the ccdf plots (top) and the corresponding Hill's estimates (bottom) for various values of w.

with the back to back transmission of a window of packets. After these packets are transmitted, no other packet is transmitted till an ACK is received for one of these packets. The receipt of an ACK marks the end of the round.

To give an explanation for TCP's effect on the self-similarity of network traffic, we consider a TCP flow to be composed of the superposition of W_{max} ON/OFF processes. Each process corresponds to each of the possible values that the *cwnd* of the flow might have since W_{max} is the receiver's advertised maximum buffer size and is the upper limit on *cwnd*. A *cwnd* of w, corresponding to the w^{th} ON/OFF process, $1 \leq w \leq W_{max}$, implies a deterministic ON time which is equal to the time to transmit the w packets with the packets generated at a constant rate in during this period. We note that though in practice there might be a small variation in the time between two successive packets in a round, these are generally very small and with high speed networks these variations are negligible when compared to the RTTs. Also, as described after a few paragraphs, the way we demarcate the end of ON periods ensures that the spacing between the packets in the ON period is almost constant.

The OFF period for the w^{th} process, $1 \leq w \leq W_{max}$, corresponds to the time interval between two successive instants where *cwnd* has the value w. Now, if the distribution of these times has a heavy tail, their complementary cumulative

distribution function (ccdf) $F_c(x)$ behaves like

$$F_c x \sim l x^{-\alpha} L(x) \qquad \text{with } 1 < \alpha < 2 \tag{1}$$

where $l > 0$ is a constant, $L(x)$ is a slowly varying function at infinity, i.e., $\lim_{x\to\infty} L(tx)/L(x) = 1, \forall\ t > 0$ and the relation $f(x) \sim g(x)$ implies $\lim_{x\to\infty} f(x)/g(x) = 1$. We can now use the following Theorem from [17] which says that the superposition of a number of these processes converges in the limit to fractional Brownian motion (fBm) and thus exhibit self-similarity.

Consider M independent ON/OFF processes. Let $F_{1c}^{(r)}$ ($F_{2c}^{(r)}$), $\mu_1^{(r)}$ ($\mu_2^{(r)}$) and $\sigma_1^{2(r)}$ ($\sigma_2^{2(r)}$) be the ccdf, the mean duration and the variance of the ON (OFF) period of the ON/OFF process of type r. Now, if $W_M(Tt)$ represents the aggregated packet count in the interval $[0, Tt]$ due to the contribution from all the M sources then

Theorem 1. (Taqqu, Willinger and Sherman) *As $M^{(r)} \to \infty$, $r = 1, \cdots, R$ and $T \to \infty$, the aggregated cumulative packet traffic $\{W_M(Tt), t \geq 0\}$ behaves statistically like*

$$Tt \sum_{r=1}^{R} \frac{M^{(r)} \mu_1^{(r)}}{\mu_1^{(r)} + \mu_2^{(r)}} + \sum_{r=1}^{R} T^{H^{(r)}} \sqrt{L^{(r)}(T) M^{(r)}}\, \sigma_{\lim}^{(r)} B_{H^{(r)}}(t)$$

where the $B_{H^{(r)}}(t)$ are independent fractional Brownian motions and $H^{(r)}$ and $\sigma_{lim}^{(r)}$ are as defined in [17].

In our case, R corresponds to the maximum window size allowed for any of the flows in the network and the limiting conditions are reached when we have a large number of flows in the network each contributing its ON/OFF processes to the superposition. Now we just need to show that the distribution of the OFF times indeed corresponds to the form of Eqn. 1. In Figs. 3 and 4 we plot the ccdf of the OFF times for various window sizes for the traces for Ohio and Italy and the heavy tailed nature of each is clearly evident. While the ccdf plots often provide solid evidence for or against the presence of heavy tails, an eyeballing method is statistically unsatisfactory and the rough estimates of α obtained from these plots may be unreliable. A statistically more rigorous method for estimating the slope of the tails and thus α is the *Hill's estimator* [8]. The presence of heavy tails is indicated by a straight line behavior of the Hill's estimate $\hat{\alpha}_n$ as the number of samples used in the calculation of the estimate increases while a steadily decreasing pattern is a strong indication of the data being not from a heavy-tailed distribution. Figs. 3 and 4 also plot the Hill's estimates for the OFF time distribution for various window sizes for the Ohio and Italy traces respectively and clearly they are consistent with the form of Eqn. 1. Thus we can conclude that the superposition of such ON/OFF process from a number of TCP flows will converge in the limit to fBm and thus exhibit self-similarity.

It is interesting to note the ccdf and Hill estimate plots for the Italy trace with $p = 0.001$. From Figure 2 the Hurst parameter for this trace can be seen to be around 0.5, i.e. the trace does not exhibit self-similarity. We note from Figure

4 that the Hill estimates for all the ON/OFF process corresponding to this trace are decaying constantly and thus do not have a heavy tailed nature. Thus the ON/OFF processes corresponding to this trace do not satisfy the conditions of Theorem 1 and as a result the trace is not self-similar. In Section 3.3 from our derivation of a lower bound of the ccdf it will be clear why low loss rates fail to give rise to heavy tails.

An important assumption here is the independence of the window sizes of different flows, which need not be the case for *all* the flows in a link. Simulation studies have indicated that the window sizes of TCP flows sharing a common bottleneck link may get synchronized though such synchronization is hard to observe in the Internet [11]. Also, most of the simulation studies focus on very heavily congested bottleneck links while link loads in practice tend to be comparatively much lower. Also, note that the independence requirements fail to be satisfied only when nearly all the flows in a link are correlated. To prove that the independence assumptions of Theorem 2 of [17] are satisfied, we analyzed some of the traces reported in [14]. The results of our statistical tests on these traces to see if the individual TCP flows are indeed independent indicate that amongst the longer flows in the traces, roughly 35-70 % of the flows are mutually independent, providing enough independent flows in the superposition.

An important part in the calculation of the OFF times is what criterion we use to define a OFF period. We define an ON period to be over whenever the distance between two successive packets in the trace exceeds a length δ dependent on the packet transmission time on the link. By keeping δ sufficiently small we can ensure that the spacing between the packets in the ON period is almost constant thus satisfying the requirement of Theorem 1. Also, as in [21], the exact numerical choice of δ does not affect the results and the heavy tailed nature of the ccdf remains an invariant independent of the choice of δ.

3 Investigating the Role of TCP

Having presented a model explaining the self-similarity of TCP traffic we now pinpoint the sources in TCP's retransmission and congestion avoidance mechanism which are responsible for this phenomena. We then derive a lower bound on the tail of the OFF time distribution and show that it decays according to a power law providing a firm mathematical foundation to our model. In this paper we concentrate on TCP Reno as it the most widely deployed variant of TCP. The effect of the other versions of TCP is discussed in Section 4. We assume that the reader is familiar with the basic concepts of TCP like the congestion window *cwnd*, slow start, delayed acknowledgments etc and refer the reader to [16] for details on TCP's algorithms.

3.1 The Impact of Timeouts

From the explanation for the observed self-similarity in TCP traffic given in Section 2 it is obvious that the central aspect of the phenomenon lies in the

infinite variance or the heavy tailed nature of the OFF time distributions. Let us now consider the features of TCP which lead to such a behavior.

In the following we assume an infinite or steady state flow currently in the congestion avoidance mode to make the visualization easier. Consider a TCP flow with a current window size of w, $w < W_{max}$. In every round that follows, the window now increases linearly until it reaches W_{max} and we need a loss for the window to drop back so that we get a window of size w again. Note that if the window reaches a value greater than $2w$ before a loss indication and it results in a fast retransmit, the subsequent congestion avoidance mode will start with a window greater than w leading to even longer times before a window of w is reached. However the occurrence of heavy tails is mainly due to the loss indications which lead to timeouts. This is due to the following reasons. A timeout represents a significant duration when no packets are transmitted and acts as a boundary between ON and OFF periods of the flow as a whole leading to a bursty nature of TCP traffic. The durations of timeouts are generally an order of magnitude greater than the RTT [12] and with coarse TCP timer granularities and variations in the RTT measurements can be quite large. Again, if the retransmitted packet following a timeout is also lost, the silent period is doubled and from the traces reported in [12] the occurrence of multiple consecutive timeouts is frequent. Also, a majority of the losses experienced by TCP flows lead to timeouts which can be attributed to the fact loss that most routers in the Internet deploy droptail queues. Correlated loss models, where all the packets following the first dropped packet in a round are also dropped are an appropriate models for the losses arising from these queues [12]. This coupled with the fact that a single loss in a window less than 4, two or more losses in a window less than 8 and three or more losses for higher windows in TCP Reno will lead to a timeout contributes to the large proportion of timeouts in the observed loss indications. Before moving on to the derivation of the lower bound on the tail of the ccdf, we first derive the probability that a loss in a window of size w leads to a timeout.

3.2 Probability of Timeouts

Consider a round with window w and let the probability that a loss of any packet in this round will lead to a timeout be denoted by $Q(w)$. We assume that the receiver sends one ACK for every two packets it receives. We assume that all losses are due to packet drops at intermediate queues and that losses due to data corruption are negligible. We also assume droptail queues and the correlated loss model of the previous subsection. Packet losses in a round are assumed to be independent of losses in other rounds and the packet loss probability is denoted by p.

For window sizes less than 4, any packet loss leads to a timeout and thus $Q(w) = 1$ for $1 \leq w \leq 3$. For windows with $4 \leq w \leq 8$ (or $K+1$ to $2(K+1)$) two or more packet losses in a round leads to a timeout. If only one packet is lost in the current round, if we lose any packet in the following round, the flow will eventually timeout. In addition the retransmitted packet must also be

transmitted successfully to avoid a timeout. Thus the probability that a packet loss *does not* lead to a timeout for this range of window values is given by

$$1 - Q(w) = \frac{p(1-p)^{w-1}}{1-(1-p)^w}(1-p)^{w-1}(1-p) \tag{2}$$

The first term corresponds to the probability of exactly one packet loss in a window of w. The second last two terms correspond to the probability that all the $w-1$ packets in the following round and the retransmitted packet are received correctly. Thus

$$Q(w) = 1 - \frac{p(1-p)^{2w-1}}{1-(1-p)^w} \qquad \text{for } 4 \leq w \leq 8 \tag{3}$$

For window sizes greater than 8, three or more losses in a round will lead to a timeout. Also we have to ensure that the retransmitted packet is received successfully along with the fact that none of the packets in the succeeding round are lost. Neglecting the extremely few possibilities in which it is possible to recover a single loss in the succeeding round without going into a timeout, we thus have

$$Q(w) = 1 - \frac{p(2-p)(1-p)^{2w-2}}{1-(1-p)^w} \qquad \text{for } 9 \leq w \leq W_{max}$$

3.3 A Lower Bound on the OFF Time Distribution

We now derive a lower bound on the ccdf by identifying the possible ways in which the time between two successive windows of the same size can exceed a given value. We concentrate on the most likely paths that the *cwnd* is likely to follow while not accounting for the others as their contribution to the ccdf is negligible. In this derivation, we measure time in units of the round trip time.

Let us assume that the current window size is w and we want to find the probability that the time until the next instant where $cndw = w$ is greater than 100. The most obvious possibility is that the flow does not experience any loss for the next 100 rounds so that after some round the *cwnd* stays at W_{max}. However, with higher loss probabilities this event is unlikely and the probability tail based on just this mechanism has an exponential decay. Another possibility could be that after i rounds (when $cwnd > 2w$) the flow experiences a loss which results in a fast retransmit. The flow then transmits the next $100-i$ rounds without any loss. As a variation of this we could have a number of successive fast retransmits without reaching a window of w. Note that each of these possibilities are mutually independent and their individual contribution the tail of the distribution has an exponential decay, each having its own rate. Yet another line of possibilities is timeouts. Let us denote the average duration of a timeout (in terms of RTTs) by $E[TO]$. As the first possibility we could have that there are no losses in the first $100 - E[TO]$ followed by a timeout. We could also have i initial rounds without loss and then n timeouts (with n sufficiently large) before the window

gets a chance to increase to w. Other possibilities include cases where we have timeout periods of length $2E[TO]$, $4E[TO]$ and so on. Again, each of these cases represent independent possibilities whose individual contribution to the tail of the OFF time distribution has an exponential decay, the rate of which depends on the corresponding probability of the loss indications and their effects.

The tail of the OFF time distribution for each window size and the corresponding ON/OFF process can thus be seen as the superposition of a large number independent exponential tails each with its own rate of decay. The mix of these independent exponentials leads to a composite distribution which has a heavy tail over the region of our interest. The following theorem by Bernstein [4] provides the link between the mixture of exponentials and a completely monotone probability density function (pdf).

Theorem 2. (Bernstein) *Every completely monotone pdf f is a mixture of exponential pdfs, i.e.,*

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} dG(\lambda), \qquad t \geq 0 \tag{4}$$

for some proper cdf G.

It can be shown that the commonly used heavy tailed distributions like Pareto and Weibull are completely monotonous. Also, in [3] it is shown that the superposition of a number of properly chosen exponentials can be used to model heavy tailed distributions in the region of primary interest. Having shown the basic construction of how the mix of exponentials lead to heavy tails in the OFF time distribution, we now obtain the probabilities corresponding to each of the possible paths that we described.

Case 1: The no loss case. Let us begin with the simplest case where there are no losses. Consider the w^{th} ON/OFF process which corresponds to a *cwnd* of w, $1 < w < W_{max}$ excluding the special cases with *cwnd*s of 1 and W_{max}. Assume that the current round has a window of size w. The probability that the next window of size w occurs after t units of time (i.e. t RTTs) assuming there are no losses in between is given by

$$P\{T > t\} = (1-p)^{N(t)} \tag{5}$$

where $N(i)$ represents that number of packets that are transmitted in the i rounds following the round with size w and is given by

$$N(i) = \begin{cases} iw + \lceil \frac{i}{2} \rceil \left(i - \lceil \frac{i}{2} \rceil \right) & \text{if } i \leq j \\ jw + \lceil \frac{j}{2} \rceil \left(j - \lceil \frac{j}{2} \rceil \right) + (i-j)W_{max} & \text{else} \end{cases} \tag{6}$$

where $j = 2(W_{max} - w) - 1$ and represents the time it takes for the *cwnd* to reach W_{max}, assuming no losses.

Case 2: Fast retransmission losses. We now consider the more likely cases where a flow experiences n losses between two successive windows of the same size which are far apart in time. Consider again the w^{th} ON/OFF process, $1 < w < W_{max}$. We can have a OFF time greater than t if we have loss indications at windows greater than $2w$ which result in fast retransmits. For simplicity, we

consider only those cases where the loss occurs in a window of size W_{max}. The flow first transmits packets without loss for the first i rounds during which its window reaches W_{max}. It then experiences a loss which is recovered by a fast retransmit. Since $w < \lceil W_{max}/2 \rceil$ the desired window size is not achieved at the beginning of the congestion avoidance mode. Also, following each loss there are $2(W_{max} - m) - 1$ rounds with $W_{max}(W_{max} - 1) - m(m + 1)$ packets till *cwnd* reaches W_{max} again with $m = \lceil W_{max}/2 \rceil$. Thus there are $t-n-n(2(W_{max}-m)-1)-2(W_{max}-w)+1$ rounds with successfully transmitted windows of W_{max}. The total number of correctly transmitted packets, after algebraic simplifications, is thus

$$\begin{aligned} N_c(w,t) = W_{max}(t - (n+1)W_{max} + 2w + 2nm - 4n) \\ -w(w+1) - nm(m-1) \end{aligned} \tag{7}$$

Now, since there are $M = t - 2nW_{max} + 2w + 2(n-1)m - 2n + 3$ rounds with a *cwnd* of W_{max} with n of them having losses, the probability that the OFF time is greater than t is given by

$$P\{T > t\} = \binom{M}{n} (1 - (1-p)^{W_{max}})^n (1 - Q(W_{max}))^n (1-p)^{N_c(w,t)} \tag{8}$$

Also, since each loss is associated with $2(W_{max}-m)-1$ rounds where the window is not W_{max}, the maximum possible losses in t rounds can be shown to be limited by

$$n_{max} = \left\lfloor \frac{t - 2(W_{max} - w) + 1}{2W_{max} - 2m - 1} \right\rfloor \tag{9}$$

Case 3: Loss indication resulting in a timeout. Let us now consider the case when the TCP flow experiences a single loss indication which results in a timeout. Consider the case when the loss occurs after i rounds from the round with a window of w. The number of packets transmitted in these i rounds, $N(i)$ is given in Eqn. 6 and the value of the *cwnd* in the i^{th} round w_i is given by

$$w_i = \min\{W_{max}, w + \lceil i/2 \rceil\} \tag{10}$$

To find the number of packets transmitted in the slow start phase which follows a timeout, we use the model of [15] which models the window increase pattern more accurately than the commonly used approximation where the window always increases 1.5 times every RTT. From [15], the number of rounds spent in the slow start phase is given by

$$t_{ss}(w_i) = \left\lfloor 2\log_2\left(\frac{2m}{1+\sqrt{2}}\right) \right\rfloor - 1 \tag{11}$$

where $m = \lceil \frac{w_i}{2} \rceil$ and the number of packets transmitted in the slow start phase can be expressed as

$$N_{ss}(w_i) = \left\lfloor 2^{\frac{t_{ss}(w_i)+1}{2}} + 3 \cdot 2^{\frac{4t_{ss}(w_i)-3}{8}} - 2 - \frac{3\sqrt{2}}{2} \right\rfloor \tag{12}$$

If $w > m$ we also have a linear phase where the window increases linearly from m to w. The total time required by the flow to reach a window of w again following the timeout is thus

$$D_{nl}(w, w_i) = \begin{cases} t_{ss}(w) + E[TO] + 1 & \text{if } w \leq m \\ t_{ss}(w_i) + E[TO] + 2(w - m) & \text{else} \end{cases} \tag{13}$$

Now, the probability that we have a loss in a round of size u following the timeout, before the window reaches w, $P_{TO}(u, w_i)$, $1 \leq u < w$, is given by

$$P_{TO}(u, w_i) = \begin{cases} (1-p)^{N_{ss}(2u)}(1 - (1-p)^u) & \text{if } u < m \\ Q(u) & \\ (1-p)^{N_{ss}(w_i)}(1 - (1-p)^{2u}) & \text{else} \\ Q(u)(1-p)^{u(u-1)-m(m-1)} & \end{cases} \tag{14}$$

Note that the $1 - (1-p)^{2u}$ term in the second case has an exponent $2u$ because in the linear phase we have two consecutive rounds with the same window size. Then, the probability that there is another timeout before the window reaches a window of w is given by

$$P_s(w, w_i) = \sum_{u=1}^{w-1} P_{TO}(u, w_i) \tag{15}$$

Note that in the summation above some of the values of $P_{TO}(u, w_i)$ are zero if $u < m$ and *cwnd* skips these values of u due to the exponential increase pattern. After the i^{th} round, on an average 2 more round of packets are sent (where the first couple of losses may be recovered) before the timeout period begins. Thus if $i \geq t - D_{nl}(w, w_i) - E[TO] - 2$, the probability that the off time is greater than t is given by

$$P\{T > t\} = \begin{cases} (1-p)^{N(i)}(1 - (p)^{w_i})Q(w_i) & \text{if } i \geq I_l \\ (1-p)^{N(i)}(1 - (p)^{w_i})Q(w_i)(1 - P_s(I_l - i)) & \text{else} \end{cases} \tag{16}$$

where $I_l = t - E[TO] - 2$. The factor $(1 - P_s(I_l - i))$ in the second case gives the probability that we do not have another loss before the window reaches w. It is absent in first case since $i + E[TO] + 2 \geq t$ and we do not have to consider whether the packets following the timeout period are transmitted correctly or not.

Case 4: When the retransmitted packet is lost. When the first retransmitted packet following a timeout is also lost, the retransmission timer backs off exponentially with a factor of 2 and can thus lead to very large silent periods. The duration of a sequence of n consecutive losses in lengths of $E[TO]$ is given by

$$L_n = \begin{cases} 2^n - 1 & \text{for } n \leq 6 \\ 63 + 64(n - 6) & \text{else} \end{cases} \tag{17}$$

Each of the losses following the initial loss indication occur with probability p. Also, the linear phase of the *cwnd* following the second loss begins after *cwnd*

reaches 2. Now consider the case when the flow experiences n loss indications, $n-1$ of them being losses of retransmitted packets and that the first loss occurred after i rounds. Then, if $i > t - L_n E[TO] - 2(w-2) - 1$ the probability that the off time for window w is greater than t is given by

$$P\{T > t\} = \begin{cases} (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i)p^{n-1} & \text{if } i \geq I_l \\ (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i)p^{n-1}(1-P_s(I_l-i)) & \text{else} \end{cases} \tag{18}$$

where $I_l = t - L_n E[TO] - 2$. The presence of $(1 - P_s(I_l - i))$ in the second case can be explained as before.

Case 5: n isolated timeouts. Let us now consider the case where there are n isolated timeouts each of length $E[TO]$. After the first loss after i rounds, the slow start phase lasts till *cwnd* reaches $m = \lceil \frac{w_i}{2} \rceil$. The second loss occurs before *cwnd* reaches a values of w. The expected duration between the first and the second loss indications is given by

$$D_l(w_i) = \begin{cases} E[TO] + 2 + \frac{1}{1-P_s(w-1)} \left(\sum_{u=2}^{w-1} u P_{TO}(u) \right) & \text{if } w < m \\ E[TO] + 2 + \frac{1}{1-P_s(w-1)} \left(\sum_{u=2}^{m-1} u P_{TO}(u) \right. & \text{else} \\ \left. + \sum_{u=m+1}^{w-1} (u + 2(u-m) - 0.5) P_{TO}(u) \right) & \end{cases} \tag{19}$$

In the above expression, the second summation in the second case corresponds to the linear increase phase where we have two consecutive windows with the same size. After the initial loss indication, each of the succeeding loss indications can occur at a window between 1 and $w-1$. For each of these, we model the average duration between two successive losses by $D_l(w)$. Also, the probability that there is another loss following first loss (before the window reaches w) leading to a timeout is given by $P_s(w, w_i)$. Correspondingly, we model the same probability for all losses after the second loss by $P_s(w, w)$. Also, after the last loss, it takes $t_{ss}(w-1) + 2(w - \lceil \frac{w-1}{2} \rceil) - 1$ rounds for the window to reach a size of w. Since $t - D_l(w_i) - i$ rounds comprise the duration for the rest of the losses following the first loss indication, we need at least

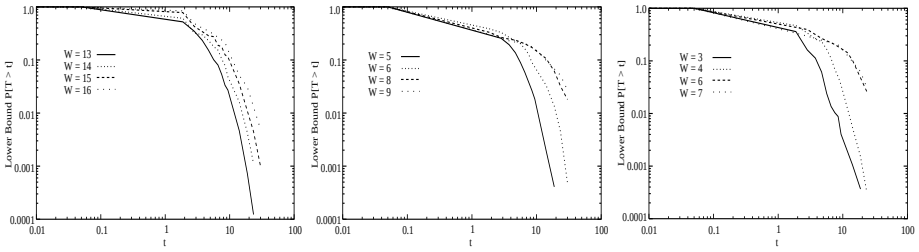
$$n = \left\lceil \frac{t - D_l(w_i) - i}{D_l(w)} \right\rceil + 1 \tag{20}$$

losses for the off time to exceed t. Then if $n > 1$ (the case $n = 1$ had already been considered) the probability that the off time is greater than t is given by

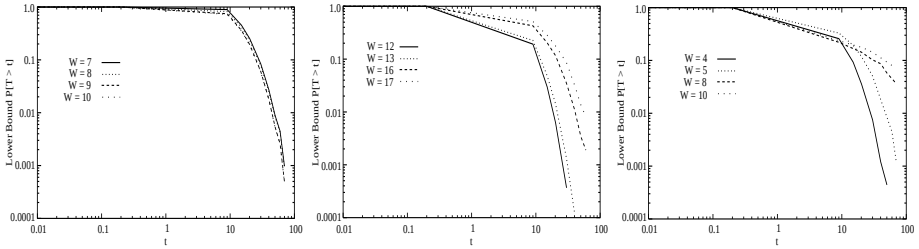
$$P\{T > t\} = \begin{cases} (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i) & \text{if } i \geq I_l \\ P_s(w, w_i)(P_s(w, w))^{n-2} & \\ (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i) & \text{else} \\ P_s(w, w_i)(P_s(w, w))^{n-2}(1 - P_s(I_l - i)) & \end{cases} \tag{21}$$

where $I_l = t - D_l(w_i) - (n-2)D_l(w) - E[TO] - 2$.

Case 6: Multiple consecutive losses. We now consider the cases where there are n losses which are successfully recovered using a single timeout and l



(a) Ohio: p = 0.010 (b) Ohio: p = 0.078 (c) Ohio: p = 0.135



(a) Pisa: p = 0.001 (b) Pisa: p = 0.006 (c) Pisa: p = 0.099

Fig. 5. The lower bound on the tails of the ccdf for the Ohio and Italy traces for various values of w. The time t is in seconds.

losses in which the retransmitted packet is also lost resulting in silent periods which are multiples of $E[TO]$. Let the l periods of consecutive timeouts be all due to j consecutive losses. The probability of each of these n periods is $P_s(w,w)p^{j-1}$ and the probability of the single loss indications is $P_s(w,w_i)$ and $P_s(w,w)$ for the first and the rest of the $n-1$ losses respectively. For a given n and l we can have a sequence corresponding of $n+l$ losses in t rounds only if $t - D_l(wi) - (n+l-1)D_l(w) - l(2^j-2)E[TO] \leq i < t - D_l(wi) - (n+l-2)D_l(w) - (l-1)(2^j-2)E[TO]$. For the values of i falling in this range, the probability that the off time is greater than t is given by

$$P\{T > t\} = \begin{cases} (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i)P_s(w,w_i) \text{ if } i \geq I_l \\ (P_s(w,w))^{n+l-2}p^{l(j-1)} \\ (1-p)^{N(i)}(1-(1-p)^{w_i})Q(w_i)P_s(w,w_i) \text{ else} \\ (P_s(w,w))^{n+l-2}p^{l(j-1)}(1-P_s(I_l-i)) \end{cases} \tag{22}$$

where $I_l = t - D_l(w_i) - (n+l-2)D_l(w) - l(2^j-2)E[TO] - E[TO] - 2$.

3.4 Numerical Results

We now present the numerical evaluation for the lower bounds for the parameters from all the Ohio and the Pisa traces considered in Section 2. In Fig. 5 we show the ccdf for the various window sizes for both destinations. The heavy tailed

Type of Loss	$p = 0.100$		$p = 0.001$	
	prob	ccdf	prob	ccdf
Case 1	0.0000	0.0000	0.0304	0.0304
Case 2	0.0000	0.0000	0.0000	0.0304
Case 3	0.0000	0.0000	0.0123	0.0428
Case 4	3.99E-4	3.99E-4	3.18E-6	0.0428
Case 5	0.0116	0.0120	0.0156	0.0584
Case 6	0.1306	0.1426	3.57E-5	0.0584

Table 1. Contribution of various losses to the ccdf. $t = 200RTTs$, $w = 10$, $W_{max} = 18$.

nature of the tails is evident and as expected, the rate of decay reduces with increasing loss probabilities. Also, to see the impact of timeouts on the tails of the ccdf, in Table 1 we show the contribution to the tails by the various cases involving timeouts that we considered in the previous subsection. As expected, the contribution from the timeouts have a large contribution to the tails, specially higher loss probabilities. For very low loss rates, the contribution due to multiple losses is negligible and the tail is made of just 3-4 exponentials. For higher losses, the probability of multiple timeouts increases and we have a large number of exponentials with different rates the superposition of which leads to a heavy tailed distribution.

4 Conclusions and Discussions

In this paper we provided an explanation of how TCP can cause self-similarity in network traffic. Using traces of actual TCP transfers over the Internet, we showed that individual TCP flows, isolated from the aggregate flow on the link also have a self-similar nature. Our results also showed that the degree of self-similarity is dependent on the loss rates experienced by the flow and increases with increasing loss rates with the traffic no longer self-similar at very low loss rates. We then proposed a model explaining the contribution of TCP to traffic self-similarity. The model is based on considering each TCP flow as the superposition of a number of ON/OFF processes where the OFF times have a heavy tailed distribution. We verified the model empirically and then provided a firm mathematical basis to the empirical observations of heavy-tailed distributions in the OFF times by deriving a heavy tailed lower bound on the ccdf.

The loss rate experienced by a TCP flow is an important indicator of the degree of self-similarity in the network traffic. A natural construction of the extremely bursty nature of TCP traffic comes from timeouts which represent "silent" periods and separate periods of activity. Since a majority of loss indications under current Internet scenarios lead to timeouts, losses increase the burstiness and the heavy tails in the OFF times. The degree of self-similarity or H being dominated by the heaviest tail in the superposition, higher loss rates

thus lead to higher values of H. In contrast when the loss rate is extremely low, TCP transmits W_{max} packets in every round and behaves like a CBR source. Thus the bursty nature is absent at low loss rates and consequently the OFF times have an exponential tail with the traffic no longer being self-similar. This explains the observations in Section 2 where flows with loss rates less than 0.001 had a Hurst parameter of approximately 0.5. Our findings show that the loss probability is a faithful indicator of the "network's effect" on TCP traffic in terms of both the effects of superposition with other flows and the degree of self-similarity of the traffic.

While TCP Reno is the most widely implemented version of TCP, other versions of TCP are currently under research, the most notable amongst them being TCP SACK. TCP SACK provides robustness against multiple packet losses in a single window and recovers them without resorting to timeouts. However, it does not completely eliminate timeouts since it requires the receipt of K (usually 3) duplicate ACKs before the retransmission mechanism kicks in. Thus timeouts are inevitable for small windows and will be present even for larger windows for correlated losses. Consequently we expect self-similarity to be present in TCP SACK traces also, though the loss rates at which $H > 0.5$ will be greater than those for TCP Reno.

References

1. Crovella, M., Bestavros, A.: Self-similarity in World Wide Web traffic: Evidence and possible causes. IEEE/ACM Trans. on Networking. **5** (1997) 835-846
2. Feldmann, A., Gilbert, A. C., Willinger, W.: Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic. Computer Communications Review **28** (1998) 42-58
3. Feldmann, A., Whitt, W.: Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Proceedings of IEEE INFOCOM. (1997) 1096-1104
4. Feller, W. E.: An introduction to probability theory and its application. Wiley, New York (1971)
5. Figueiredo, D. R., Liu, B., Misra, V., Towsley, D.: On the autocorrelation structure of TCP traffic. Technical Report TR 00-55, Computer Science Department, University of Massachusetts. (2000)
6. Grossglauser, M., Bolot, J-C.: On the relevance of long-range dependence in network traffic. IEEE/ACM Trans. on Networking. **7** (1999) 629-640
7. Guo, L., Crovella, M., Matta, I.: TCP congestion control and heavy tails. Technical Report BU-CS-2000-017, Computer Science Department, Boston University. (2000)
8. Hill, B. M.: A simple general approach to inference about the tail of a distribution. Annals of Statistics. **3** (1975) 1163-1174
9. Jacobson, V.: Congestion avoidance and control. Proceedings of ACM SIGCOMM. (1988) 314-329
10. Kumar, A.: Comparative Performance Analysis of Versions of TCP in a Local Network with a Lossy Link. IEEE/ACM Trans. on Networking. **6** (1998) 485-498
11. May, M., Bonald, T., Bolot, J-C.: Analytic evaluation of RED performance. Proceedings of IEEE INFOCOM. (2000) 1415-1424

12. Padhye, J., Firoiu, V., Towsley D., Kurose, J.: Modeling TCP Reno performance: A simple model and its empirical validation. IEEE/ACM Trans. on Networking. **8** (2000) 133-145
13. Park, K., Kim, G., Crovella, M.: On the relationship between file sizes, transport protocols, and self-similar network traffic. Proceedings of International Conference on Network Protocols, (1996) 171-180
14. Paxson V., Floyd, S.: Wide area traffic: The failure of Poisson modeling. IEEE/ACM Trans. on Networking. **3** (1995) 226-244
15. Sikdar, B., Kalyanaraman, S., Vastola, K. S.: TCP Reno with random losses: Latency, throughput and sensitivity analysis. Proceedings of IEEE IPCCC, (2001) 188-195
16. Stevens, W. R.: TCP/IP illustrated. vol. 1. Addison Wesley (1994)
17. Taqqu, M. S., Willinger, W., Sherman, R.: Proof of a fundamental result in self-similar traffic modeling. Computer Communication Review. **27** (1997) 5-23
18. Taqqu, M. S., Teverovsky, V.: On estimating long-range dependence in finite and infinite variance series. In: Adler, R. J., Feldman, R. E., Taqqu, M. S., (eds.): A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Birkhauser, Boston, (1998) 177-217
19. Veres, A., Boda, M.: The chaotic nature of TCP congestion control. Proceedings of IEEE INFOCOM. (2000) 1715-1723
20. Veres, A., Kenesi, Z., Molnar, S., Vattay, G.: On the Propagation of Long-Range Dependence in the Internet. Proceedings of ACM SIGCOMM. (2000) 243-254
21. Willinger, W., Taqqu, M. S., Sherman R., Wilson, D. V.: Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM Trans. on Networking. **5** (1997) 71-86