

**A SCALE-ORIENTED STUDY OF THE INTERNET
TRAFFIC**

KONSTANTINOS DRAKAKIS

DRAGAN RADULOVIC
INGRID DAUBECHIES

PROGRAM IN APPLIED AND COMPUTATIONAL MATHEMATICS
PRINCETON UNIVERSITY

1. INTRODUCTION

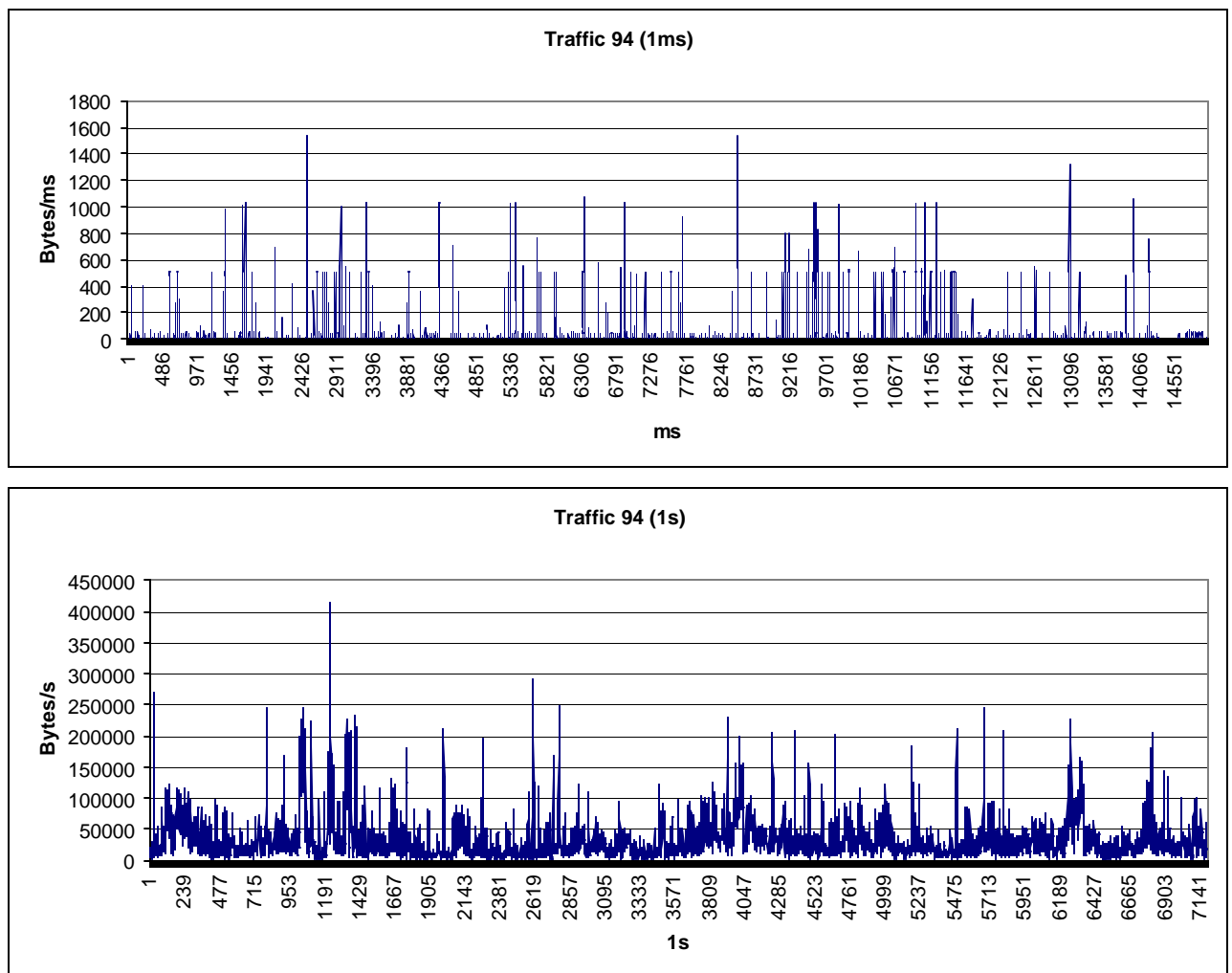
The purpose of this work is to study the Internet traffic as a stochastic process and determine some of the properties it has. Some of the questions that immediately come to mind are the following:

- What is the marginal distribution?
- How much correlated are its samples?
- How “smooth” and “regular” is it?
- Which tools can we use to study it?

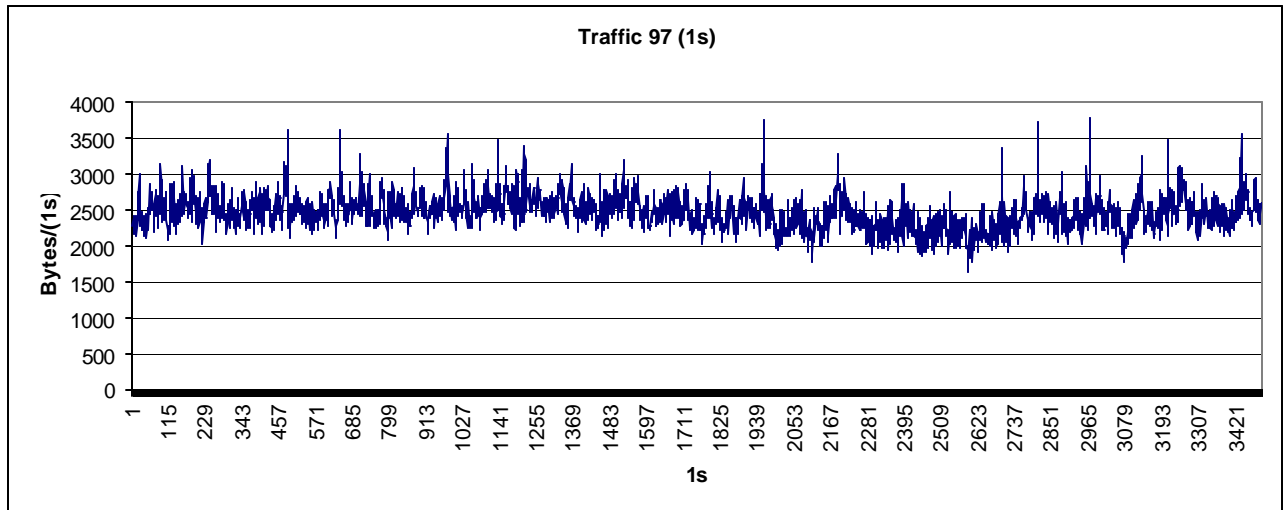
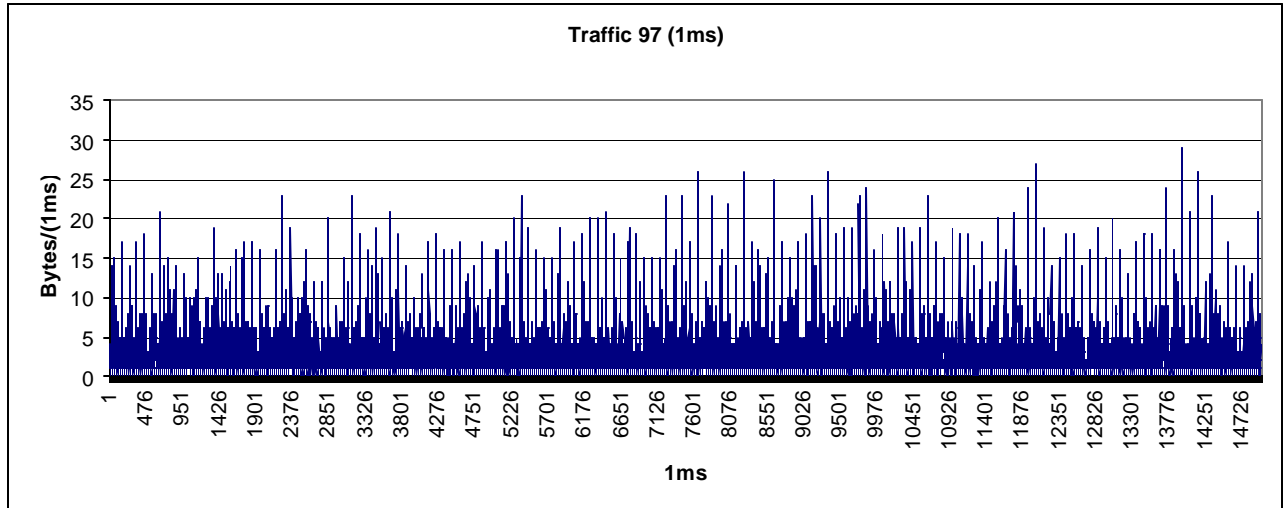
An attempt will be made to answer this and more in the following.

2. THE PROCESS

Let us compare visually two data sets:



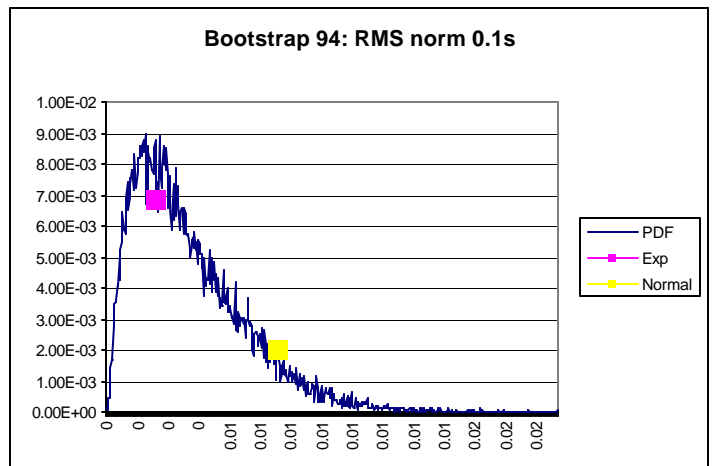
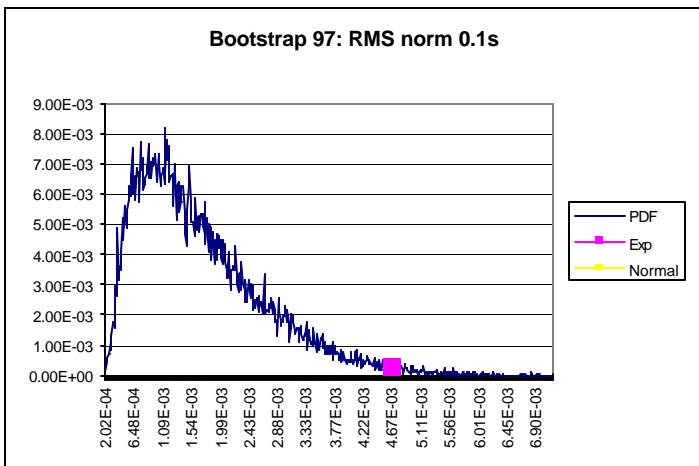
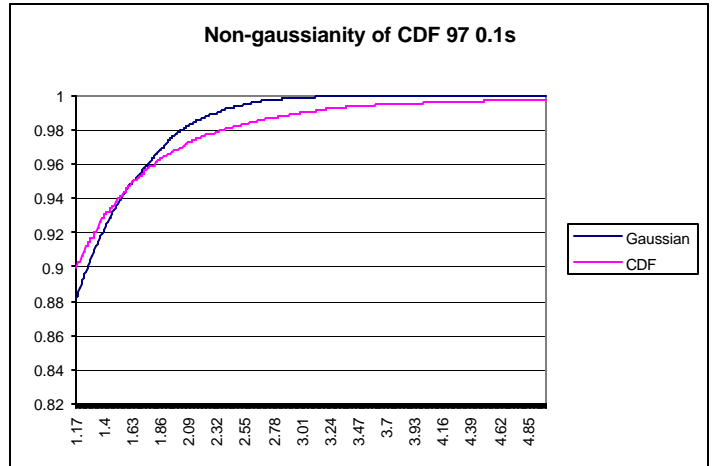
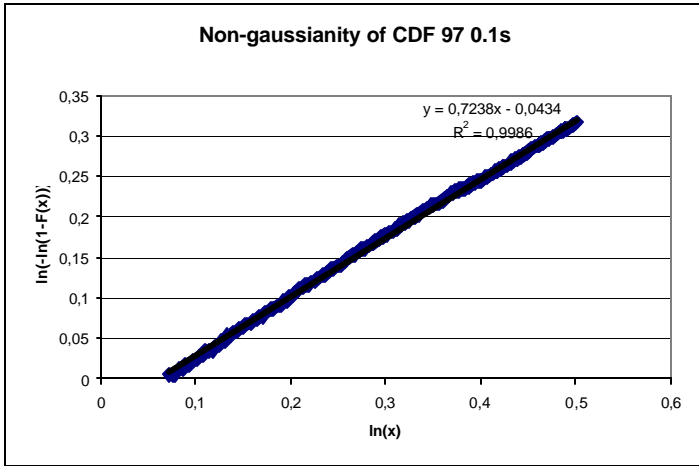
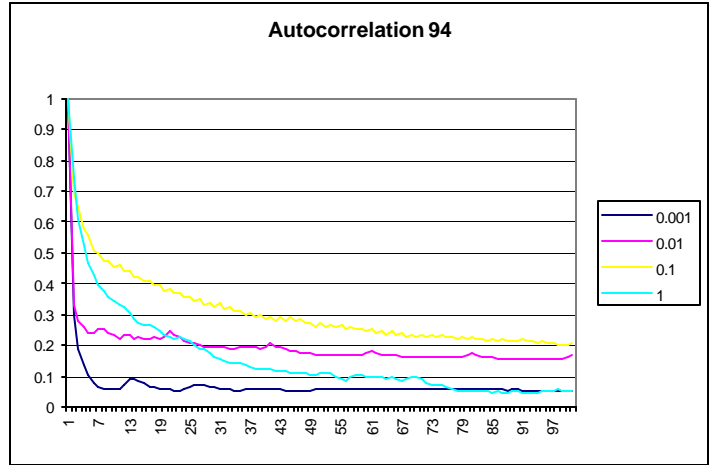
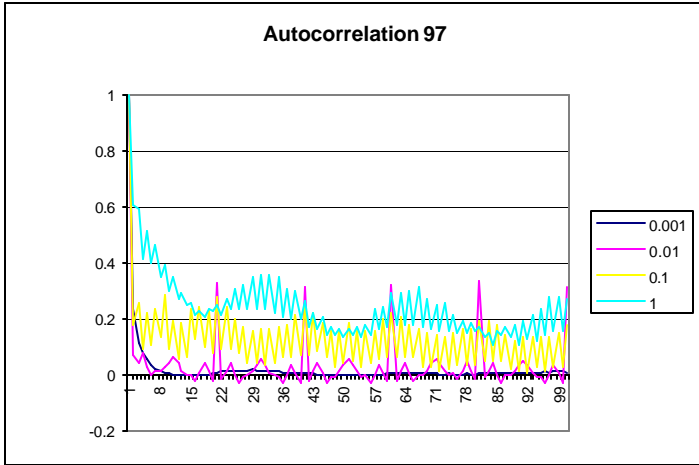
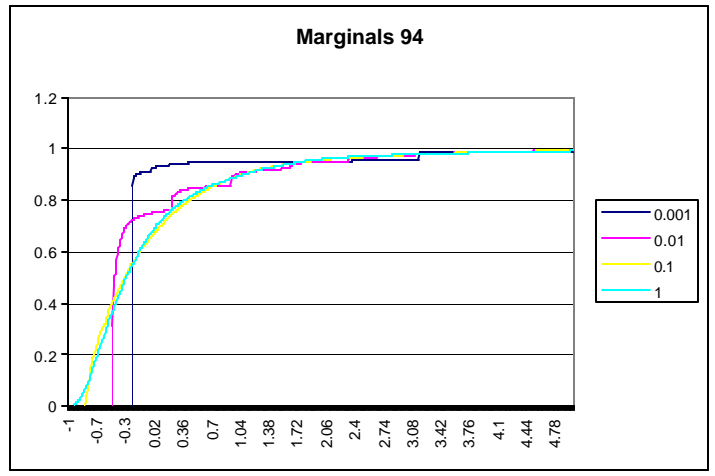
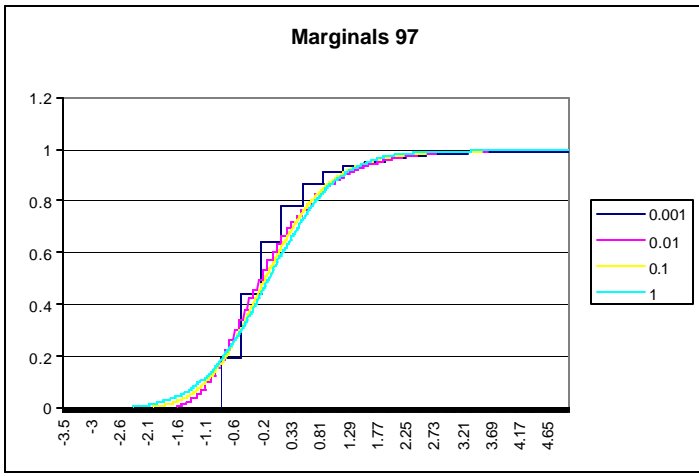
Observe that the second figure consists of averages of 1000 samples of the first and yet it is still very spiky. This appears to violate the CLT.



Here the CLT seems to hold. So, its failure in the first case must be traced in the violation of its assumptions: either the traffic is heavily correlated (exhibits long range dependence) or its distribution is heavy tailed (has infinite variance). In any case, it is clear that the process might have different properties in different time scales. This would justify the use of wavelets.

3. EVIDENCE FOR DEPENDENCE

We show below the autocorrelations for different time scales, and also the marginals. In the case of 94, there is strong evidence that the marginal has exponential tail, and in the case of 97 that it is Weibull. Notice the similarity of the 97 marginal to Gaussian, though. For time scales less than 0.1s the marginal is approximately discrete. In the testing of the statistical hypothesis: “Exponential and not Gaussian marginal” we used the fairly recent technique of **bootstrap**. It is also noteworthy that the autocorrelations for data set 94 initially increase and then decrease as time scale increases. In fine time scales there is no correlation because the low level protocols mix bytes from many different applications more or less independent, and in coarse time scales because different applications, even of the same user, tend to be independent.



Overall, 97 is much more “regular” than 94: we may safely conclude that 97 comes from a LAN whereas 94 comes from a WAN. The rippling that appears in the autocorrelation of 97 can then be accounted for: it is produced by a characteristic RTT that prevails in the LAN and seems to be around 0.2s. Something similar cannot exist in a WAN where users are much more diverse.

A final note on long-range dependence in the case of both 94 and 97 is in order: the CLT predicts that $\frac{1}{n^a} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} Z$, where Z is either Gaussian or p-stable, according to whether $E[X_i^2] < \infty$ or not (assume $E[X_i] = 0$), for some a . But none of these variables exhibits exponential or Weibull tails: this must be the result of data dependence.

4. WAVELET TOOLS AND THE SPECTRUM

Since the previous analysis suggests that different time scales (levels) have different properties, we can try to use wavelets to isolate each of them. We will be using the Haar wavelet, because it is both adequate and simple.

First, we can start forming coarser and coarser levels of traffic:

$$X_{0,i} = X_i, i \in N$$

$$X_{j,i} = \frac{X_{j-1,2i-1} + X_{j-1,2i}}{\sqrt{2}}, D_{j,i} = \frac{X_{j-1,2i-1} - X_{j-1,2i}}{\sqrt{2}}, i \in N, j > 0$$

The $D_{i,j}$'s constitute a measure of the variation within each level. We can interpret then the quantity $E_j = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n D_{i,j}^2$ as the **energy** of the j -th scale, and this is a constant function of j in the case of white Gaussian noise. Closely related to this is the **averaging function** A_j :

$$X_{0,i} = X_i, i \in N$$

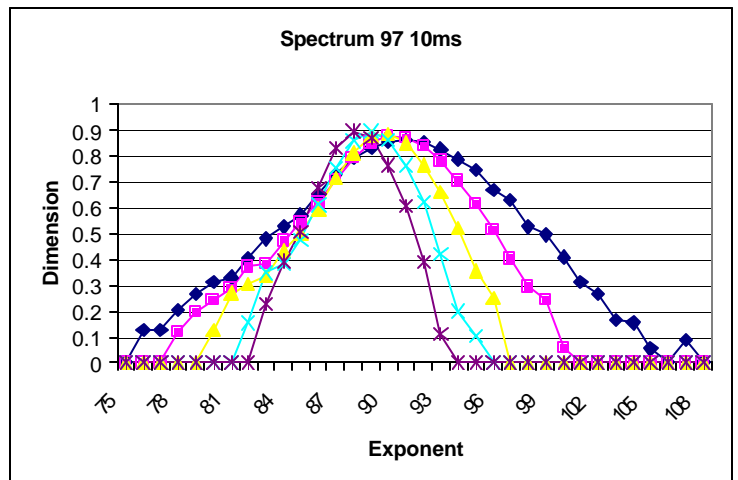
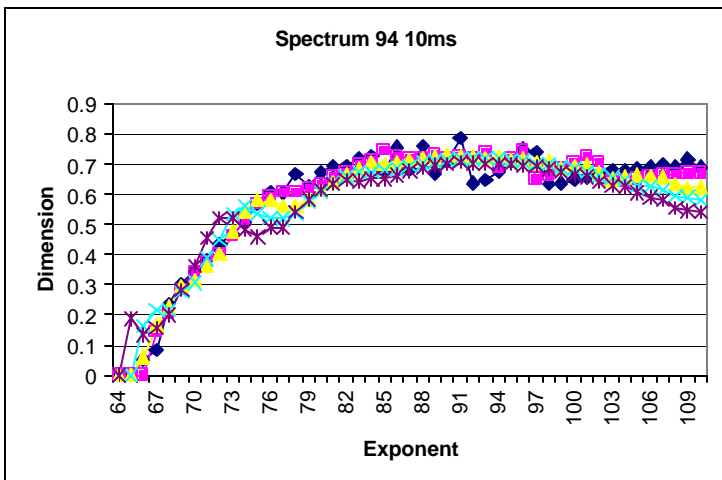
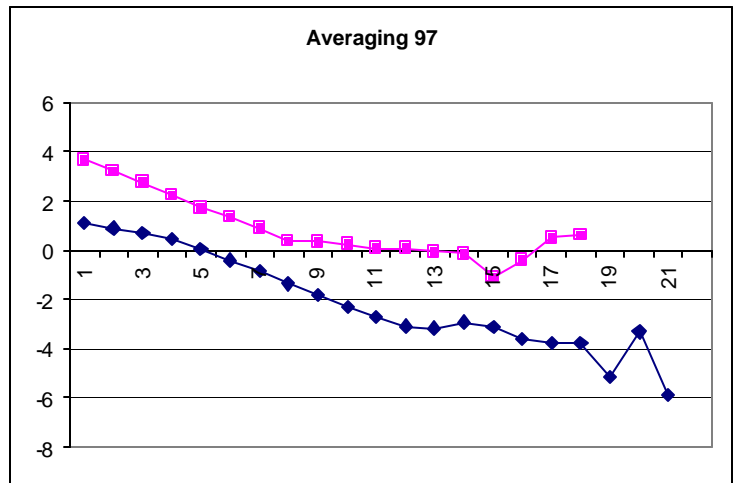
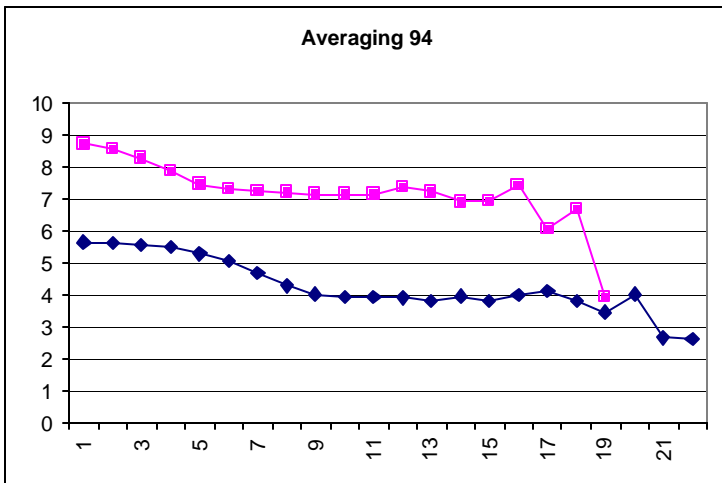
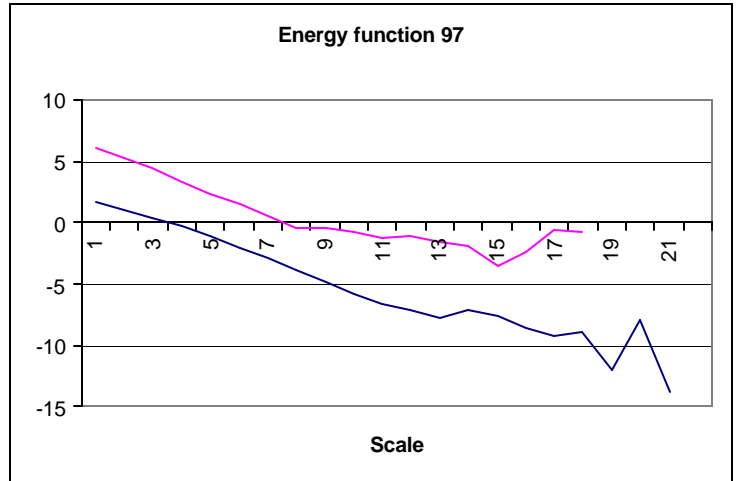
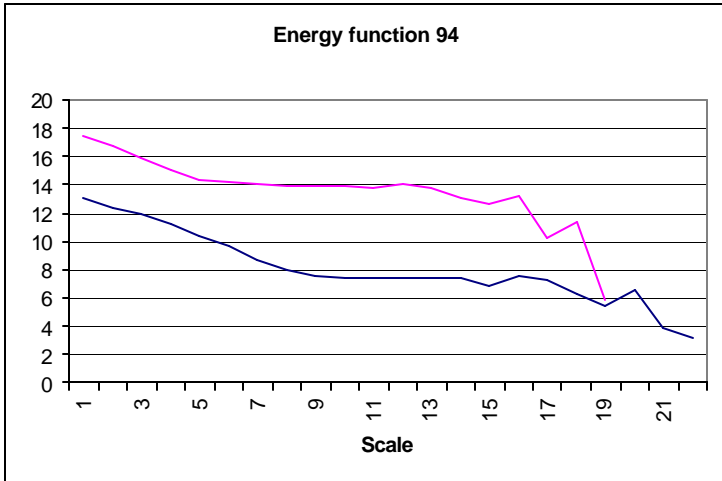
$$X_{j,i} = \frac{X_{j-1,2i-1} + X_{j-1,2i}}{2}, i \in N, j > 0$$

$$A_{j,i} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n |X_{j,2i-1} - X_{j,2i}|$$

The averaging function is a linear function of j with slope -0.5 in the case of white Gaussian noise.

Another tool, not relying on wavelets but that implements a different way of leveling is the **spectrum**: this is the Hausdorff (or the box, more correctly) dimension of the subset of the domain of a function f , in which it is Hölder continuous with exponent a , as a

function of a . We can apply this to the cumulative traffic $\sum_{i=1}^n X_i$, but normalized now so that it is a unit measure on $[0,1]$. In the case of white Gaussian noise we get theoretically only one point at $a=1$ (this is the self-similarity of the Brownian motion). Below the logarithm of the averaging and energy functions is plotted:



Observe that the spectrum makes sense only up to the exponent 1. Then, we get better approximation if we apply the procedure on the derivative. We see that the data set 94 has a much wider spectrum that converges nicely, unlike 97's that shrinks. This suggests multifractality (as suggested by Riedi). For a definition of the spectrum, see R. H. Riedi's "Introduction to multifractals".

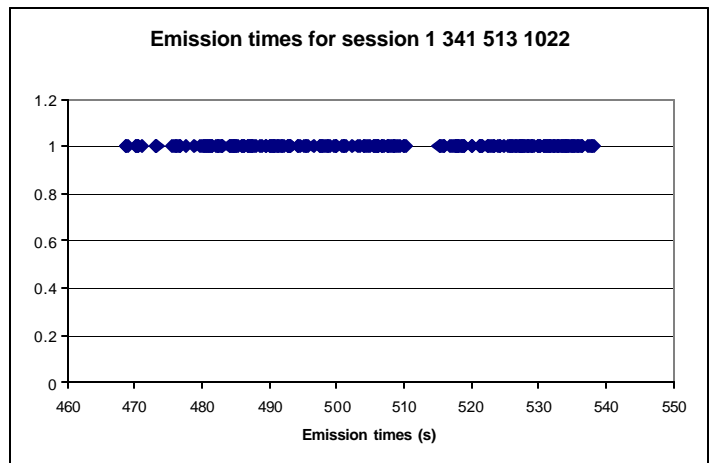
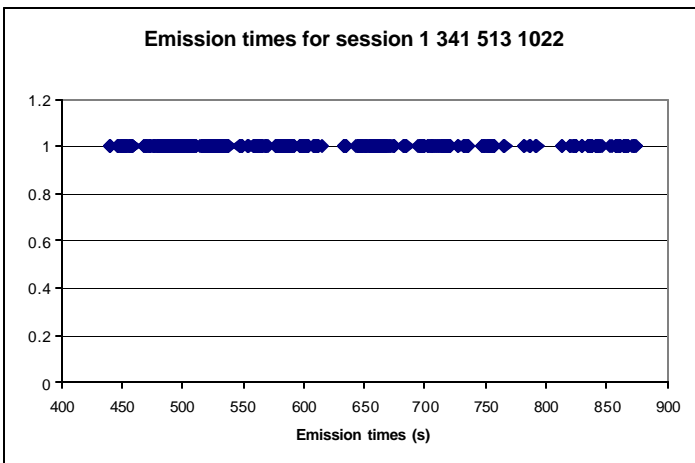
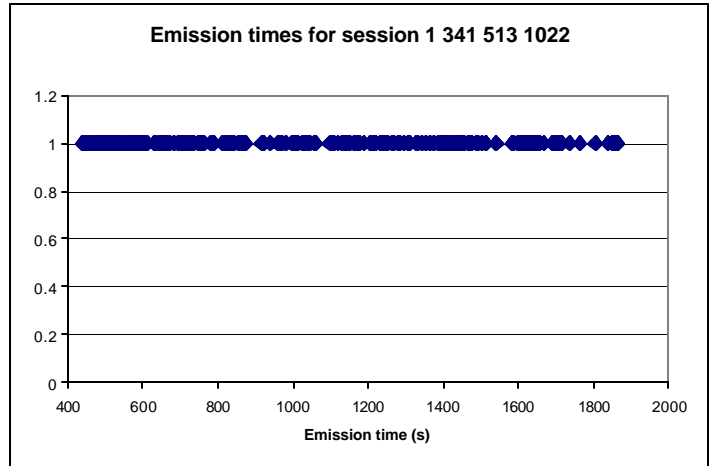
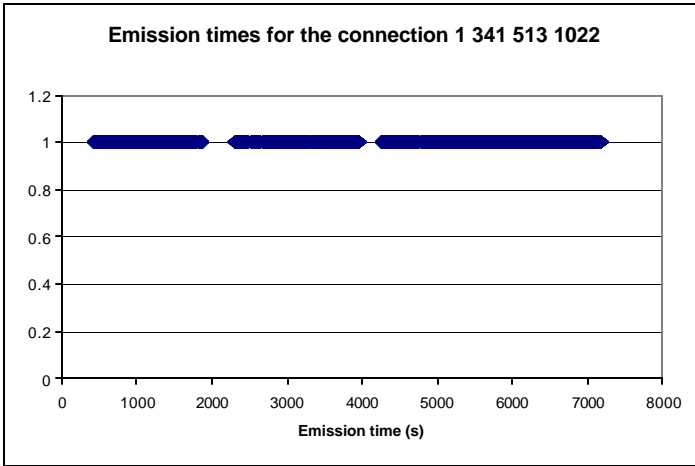
5. THE NOTION OF LEVELS OR SCALES

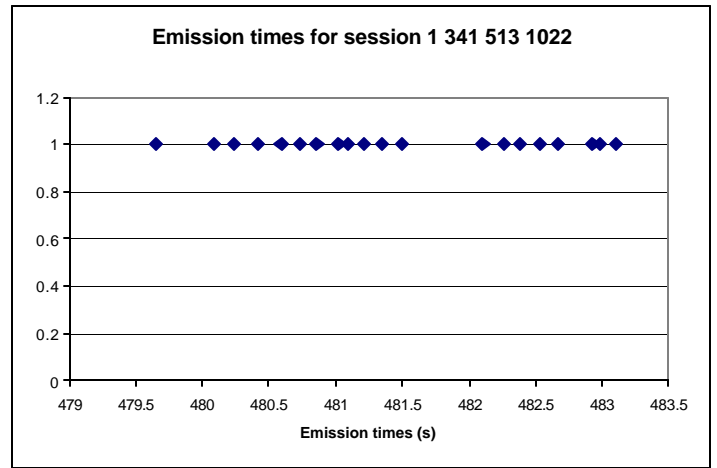
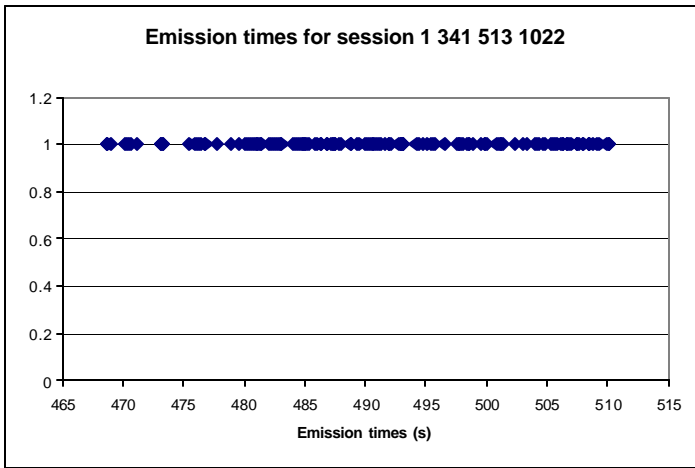
Consider the function $f(t) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} a_{ji} P(2^{-j}t - i)$, $a_{ji} = \pm 1$, where the $\{a_{ji}\}$ are r.v.'s and

$P(t) = \begin{cases} 1, & t \in [0,1] \\ 0, & \text{elsewhere} \end{cases}$. It is then obvious that the energy function at scale j will depend only

on the function $f_j(t) = \sum_{i=0}^{\infty} a_{ji} P(2^{-j}t - i)$, since finer scales will have been averaged out and coarser scales will be "invisible".

In real networks, of course, things will not be so regular, but we will provide evidence in favor of the existence of levels. In the following we depict the structure of a particular connection from data set 94:

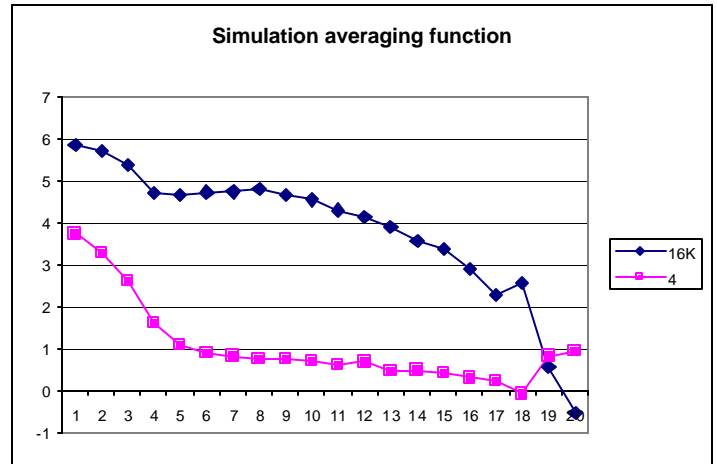
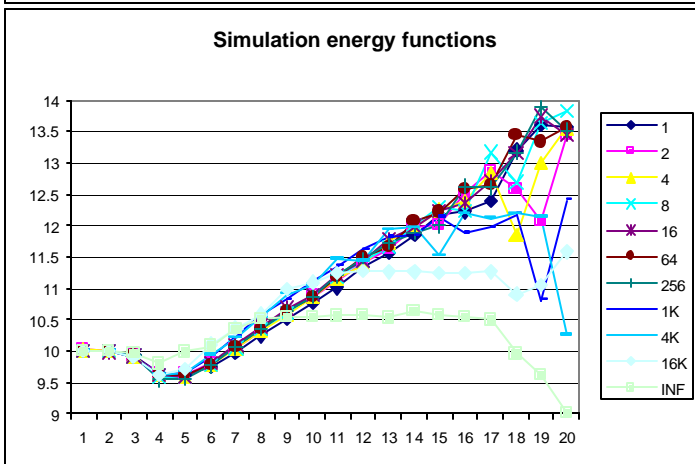
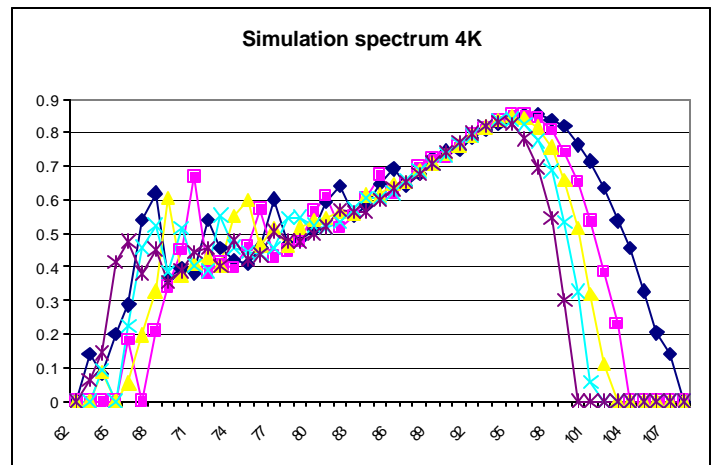
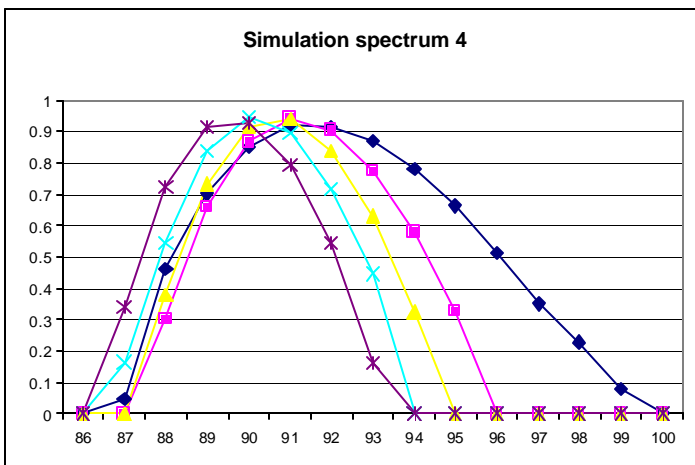




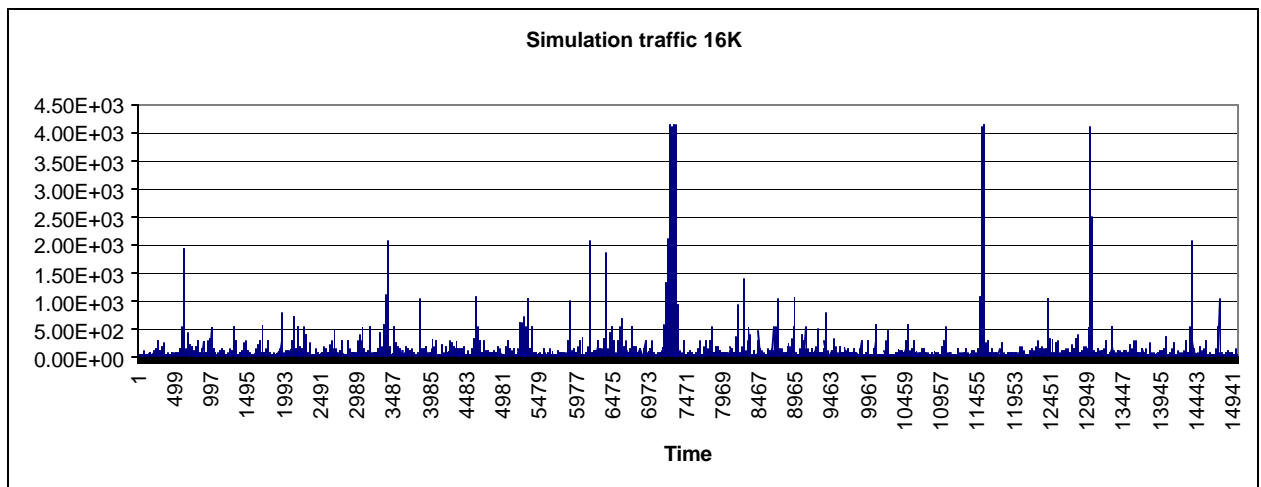
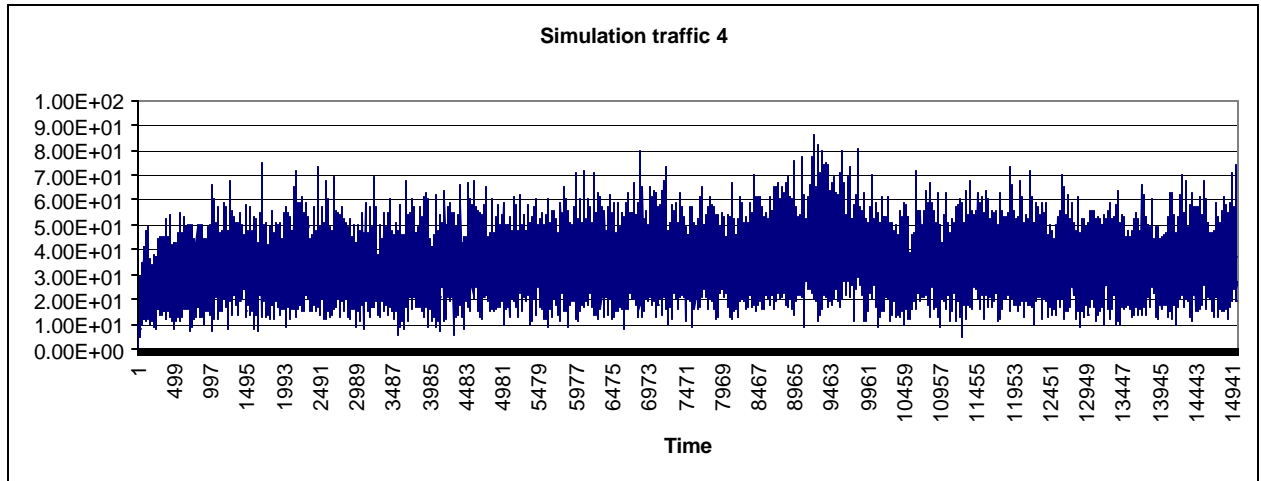
Not all sessions, of course, possess the same number or the same kind of levels. But it is clear that the distribution of inter-emission times in different levels will vary as a result of different mechanisms of the network influencing different levels (in our example, the distributions of the vectors $\{a_{ji}\}$ will vary with j). A level will influence only its neighboring levels, when the energy function is computed. Thus, it is more than plausible that the curving of the energy function is caused by (and indicates) multiple levels.

In practice now, levels are difficult to find algorithmically in all cases, but there is at least one simple case, in which they appear clearly: the case of FTP and other connections in which it is the TCP protocol mechanisms that dominate.

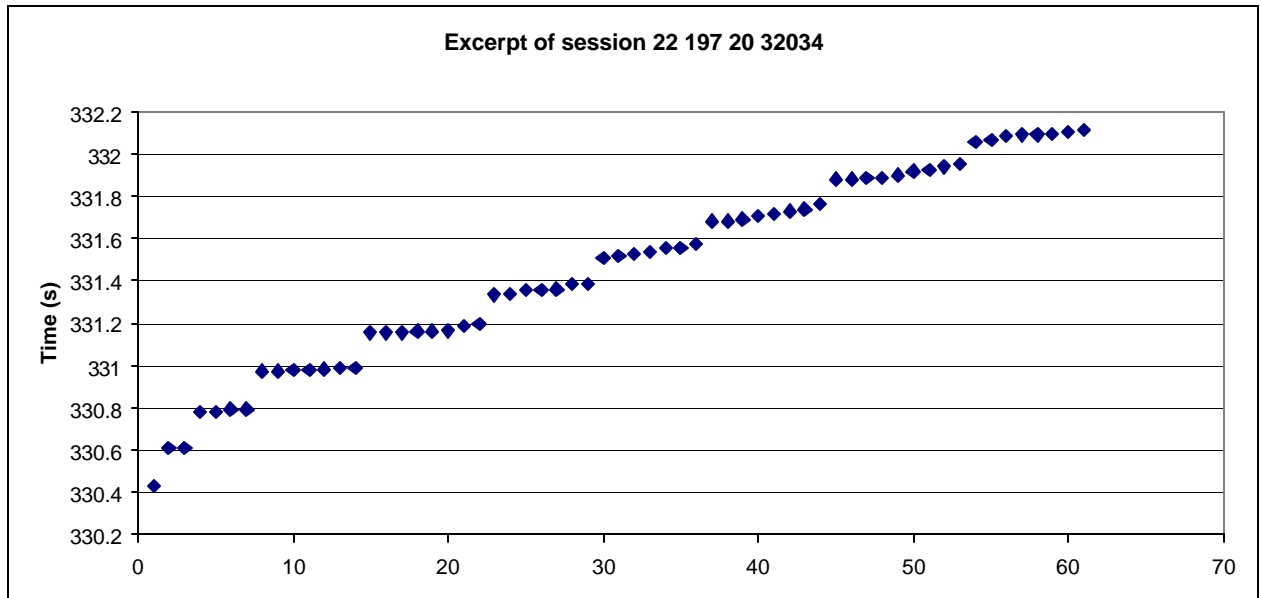
TCP has the “slow start” feature that turns out to be a very important factor of shaping for the traffic, and accounts, partially at least, for its wildness. This we verified by simulation of network traffic. It was this “slow start” that changed the picture:



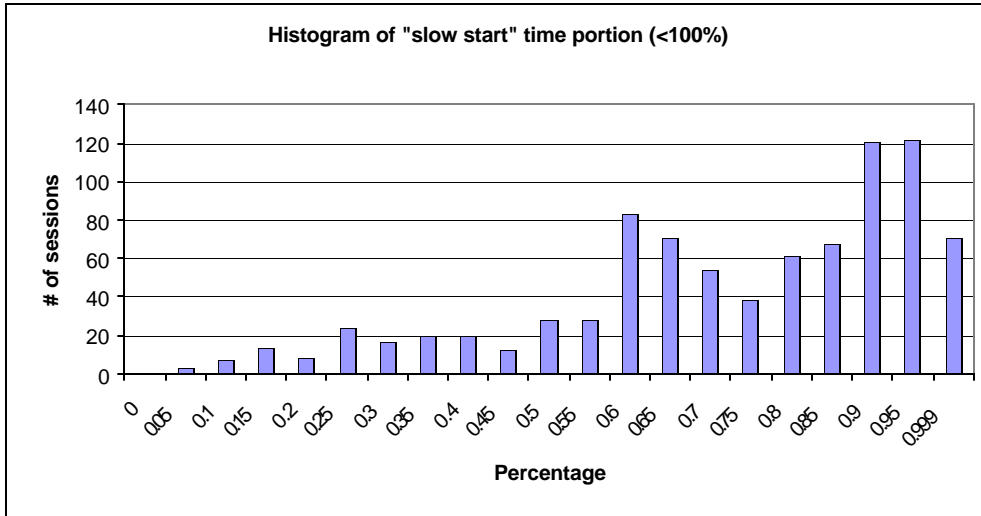
It appears also that the energy and averaging functions are rather insensitive to “slow start” parameters and rather depend on time parameters (such as RTT distributions, application running times etc.). They might also depict the dependence of data.



In data set 94, the “slow start” in some connections is particularly prominent:



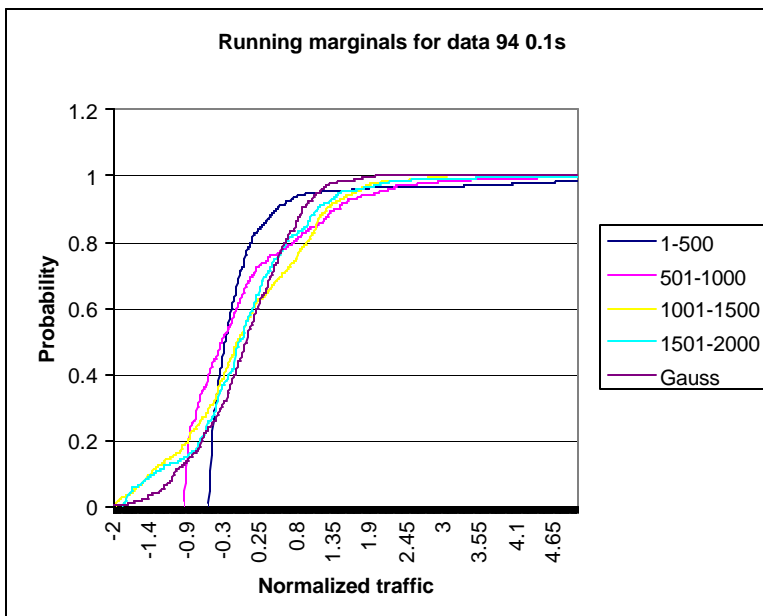
Taking then all sessions in ports 20,53,70 and 80, which make up 31% of the total traffic, we determined the time portion in which they were in “slow start” mode. Out of 1691 sessions, 822 are for all practical purposes 100% in “slow start” (mainly because they are small). The rest have the following histogram:



So, in general “slow start” plays a very important role (“slow start” means roughly sending exponentially increasing number of packets up to a limit).

6. THE “RUNNING MARGINALS”

How does the profile of the traffic change with time? We take marginals of consecutive non-overlapping groups of 500 samples and compare them:



We see that the marginals oscillate between almost Gaussian and something else (maybe exponential-like). The correlation of the measure of the difference between the Gaussian and a marginal and the traffic is 39%. So, Gaussian can be associated with periods of smooth traffic and the rest with periods of some extent of burstiness.

7. CONCLUSIONS AND EXTENSIONS

- Internet traffic can have many different forms and behaviors, not always rich in structure. Sometimes CLT can account for it (LAN), sometimes it cannot (WAN).
 - Our experiments as well as the mechanisms of the network suggest the study of traffic at different levels that will have different properties, because they are dominated by different phenomena. The traffic is little autocorrelated in the finest and coarsest levels, because of protocols and applications, respectively.
 - Wavelets can generate tools for the study of Internet traffic: these tools seem to be sensitive more in the time distribution of packets and the number of levels than in the size of packets. Together with the spectrum, they probably offer us a quite complete picture of the traffic.
 - The “slow start” mechanism that lies at the heart of the TCP protocol is largely responsible for the form the traffic has.
 - The marginal of the traffic evolves with time and oscillates between the classical Gaussian and something else. This evolution is definitely correlated to the burstiness of the traffic.
-
- All the pieces of evidence that have been gathered must be crosschecked, so that deeper understanding of their causes is gained
 - We will have to prove theorems that will confirm and predict various aspects of the traffic behavior.
 - The results will have to be compared with the various protocol specifications (as was already done with “slow start”).
 - More data sets will be needed to further confirm our results.