Local and Cumulative Analysis of Self-similar Traffic Traces

Ramírez Pacheco Julio César$^1$ and Torres Román Deni$^{1,2}$
$^1$Department of Electrical Engineering
CINVESTAV-IPN, Guadalajara, México
$^2$Department of Electronics, Systems and Computer Science
ITESO, Guadalajara, México
Phone (33) 31345570           Fax (33) 31345579
cramirez@gdl.cinvestav.mx          dtorres@gdl.cinvestav.mx

Abstract

Internet traffic shows variability in all time scales, which in turn shows statistical self-similarity. This self-similar behaviour has significant implications for QoS since it increments the total delay and packet loss rate. Therefore, we need to test for the degree of self-similarity and use this information for control purposes. For achieving the above-mentioned, the use of traces consisting of several thousands of points and hours of measurement are used. However, there are not enough studies about the number of points required to get an accurate estimation of the Hurst exponent. In this article, we study the local and cumulative behaviour of many real and synthetic self-similar traces. This is done for trying to infer the number of points required for Hurst parameter estimation and for checking dependence of Hurst exponents. We show that local analysis presents self-similarity, and the Hurst exponent tends to be stable in the cumulative case.

Index Terms.- Self-similarity, long-range dependence, long-memory estimators, local LRD analysis, cumulative LRD analysis.

1. Introduction

Traffic in computer networks has been shown to be self-similar[1],[2],[3]. This self-similar behaviour has significant implications for QoS since it increments the total delay and packet loss rate. Due to this reason an accurate estimation of the degree of self-similarity or Hurst exponent is required. From this estimation several control algorithms can be developed to fit the required QoS. When estimating the Hurst exponent we usually require a data trace consisting of several thousands of points that require many hours of measurement, for example in the classic papers[1],[2] the trace BC-pAug89 required more than a million points and more than a day of measurement. In other studies[4] tens of thousands of points of MPEG4 VBR data were required and tens of minutes of measurement. In this article, we study the behaviour of the Hurst exponent in different instants of time and check the dependence of the points. We also try to answer the question of how many points are required in the trace for obtaining an accurate estimate of the Hurst exponent. This can reduce the number of points required for accurate estimation of the degree of self-similarity as well as the time of measurement. For accomplish the above, we make use of local and cumulative analysis of self-similar traffic traces with different time-domain and frequency-domain algorithms. For a complete reference of time-domain and frequency-domain techniques for self-similarity testing see[8],[9]. In the local analysis, the partition of the trace of size $N$ in blocks of size $m$ is first done followed by the computation of the Hurst exponent for each block $m$. In the cumulative case, the trace of size $N$ is first partitioned in blocks of size $m$ and then compute the Hurst exponent for all the blocks of size $im$, where $i = 1, 2, ..., N/m$. This local and cumulative analysis are done for many real and synthetically generated self-similar traffic traces, that contain several thousands of points and diverse characteristics and differs of the local analysis of [7]. The main goals of this paper are:

- To present a local analysis of self-similar network traffic traces.
- To present a cumulative analysis of self-similar network traffic traces.
- To study the dependence of local Hurst values via local analysis.
- To infer the number of points required to get accurate estimates of the Hurst exponent.
The paper is organized as follows. Section II presents the basic definitions of self-similar and long range dependent stochastic processes and also the importance of these for QoS. Section III presents the synthetic and real self-similar traffic traces used in this article and its characteristics. Section IV deals with the local analysis of real and synthetic self-similar traffic traces and the dependence of the points in time. Section V presents the cumulative study of real and synthetic self-similar traffic traces. It also presents a study of the evolution of the Hurst exponent on these traces in time. Finally in Section VI the conclusions and future work are presented.

2. Self-similarity and Long-range dependence

2.1. Self-similarity

Self-similarity in computer networks was first observed in the classic papers[1],[2]. First it was observed in Local Area Network traffic and then extended to WWW traffic[3], VBR traffic[4],[5], etc. It is very different from past views that considered network traffic as a generalization of telephone traffic. In the case of self-similarity, burstiness is present in all time-scales and does not tend to pure white Gaussian noise when aggregated as in the poisson process. Intuitively self-similarity means that the properties or characteristics of an object are maintained independently of scaling in time or space[6]. In computer networking we are concerned with statistical self-similarity, i.e., the behaviour of the autocorrelation function in different time scales. Let \( Y(t) \) be a real stochastic time series, \( Y(t) \) is said to be self-similar if the following condition holds.

\[
Y(t) = \alpha Y(at) \quad (1)
\]

Where \( t \geq 0, \alpha > 0, =d \) means equality of the finite-dimensional distributions and \( 0 < H < 1 \) represents the Hurst exponent, which quantifies the degree of self-similarity. As can be seen in (1) self-similarity implies that if a process is scaled in time and normalized by \( \alpha H \), it must maintain the distributional properties of the original process or equivalently it must have the same finite-dimensional distribution as the original process. When considering discrete time series the definition is given in terms of the aggregate process. Let \( X(t) \) denote the rate process of a destination host or switch, i.e., the number of packets, bytes, bits, etc. per time unit, we define the aggregate process as:

\[
X^{(m)}(i) = \frac{1}{m} \sum_{t=0}^{m-1} X(t) \quad (2)
\]

Where \( m \) represents the level of aggregation. That is, we obtain a new time series by partitioning the original time series into non-overlapping blocks of size \( m \) and then average each block to obtain the \( i \) values of the new series. From (2) we can apply several algorithms for Hurst exponent estimation[8],[9]. In computer networking, we are interested in asymptotic second order self-similarity since it implies similar distributional behaviour in diverse time scales and not in all time scales as in the strict case. Asymptotic second order self-similarity is defined by the following formula:

\[
\lim_{m \to \infty} \gamma^m(k) = \frac{\sigma^2}{2} (k + 1)^{2H} - 2k^{1+H} + (k - 1)^{2H} \quad (3)
\]

where \( \gamma^m(k) \) is the autocovariance function of the aggregated process and \( H \) represents the Hurst parameter.

2.1. Long-range dependence

Long-range dependence or long memory is intimately related to self-similarity. In long-range dependent processes the autocorrelation function decays hiperbolically. Self-similarity is present in the aggregate process of long-range dependence, LRD, processes and LRD processes can be constructed by the increment process of certain self-similar processes. In long-range dependent processes, the sum of the autocorrelation function is infinity, meaning that decays slowly in time. Let \( X(t) \) be a stationary process, \( X(t) \) is said to be long-range dependent if its autocovariance function, ACF, \( \rho(k) \) has the following asymptotic form:

\[
\rho(k) \sim C_p k^{-\alpha} \quad (4)
\]

Where \( C_p \) is a constant and \( \alpha \in (0,1) \) is a real number, \( \alpha \) is related to the Hurst exponent by the following relation: \( H = 1 - \frac{\alpha}{2} \).

2.3. Importance

The importance of this self-similar behaviour is that it has significant implications in the performance of computer networks. It has been shown that it leads to increased network delays and packet loss rates, then affecting the overall QoS of the network. It is then useful to try to infer the degree of self-similarity as accurate as possible and to be able to generate synthetic traffic with known Hurst parameter for different purposes. For example in computer network simulation, it would be desirable to use synthetic self-similar traffic for protocol design and testing, and also for measuring the performance of a congestion control algorithm in the presence of highly variable traffic.
3. Self-similar traces characteristics

In this section, characteristics of the real and synthetic traces under study are presented. We have used traces of different lengths with known Hurst values. It is well known in the literature that real traces tend to show SRD and LRD. This behaviour have some detrimental effect in the accuracy of the estimators. In this paper it is assumed that real traces exhibit both LRD and SRD. The analysis of pure LRD traces is also done. These are fractional gaussian noise synthetic traffic traces generated by the method proposed in [10]. In the case of the real self-similar traffic traces, several well studied traces are used, we use the famous traces **BC-pAug89, BC-pOct89, BC-pOct89Ext** from Bellcore[11]. It is also used another trace **2002-Apr_08_Mon_2130.7200.sk1.1ms** that can be found in [12] and that was studied in [7]. Table 1. shows the traces we studied.

<table>
<thead>
<tr>
<th>Trace</th>
<th>Type</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC-pAug89</td>
<td>LAN</td>
<td>1000000</td>
</tr>
<tr>
<td>BC-pOct89</td>
<td>LAN</td>
<td>1000000</td>
</tr>
<tr>
<td>BC-pOct89Ext</td>
<td>LAN</td>
<td>1000000</td>
</tr>
<tr>
<td>2002-Apr_08_Mon</td>
<td>LAN</td>
<td>7302269</td>
</tr>
<tr>
<td>fgN070</td>
<td>Synthetic</td>
<td>262144</td>
</tr>
<tr>
<td>fgN080</td>
<td>Synthetic</td>
<td>262144</td>
</tr>
<tr>
<td>fgN090</td>
<td>Synthetic</td>
<td>262144</td>
</tr>
</tbody>
</table>

4. Local analysis of self-similar traces

In this section, a detailed study of the local analysis of self-similar traces is accomplished. In the local analysis, first a data trace of length \( N \) is partitioned in \( m \) blocks. After this, we compute the Hurst exponent for each block. From the above a new time series \( i=0,1,.. N/m \) is obtained, where each point \( i \) represents the Hurst exponent of each block. The last step is plotting the new series from which a study of the local behaviour of the trace can be performed. Local behaviour can be useful since we have \( m \) Hurst estimates corresponding to different stages of time and from this we can check the dependence of points or Hurst estimates of distant blocks. The local analysis of the traces described in the previous section were done using the following tests for self-similarity:

- Variance Method.
- Variance of Residuals.
- R/S Statistic.
- Periodogram.

Figure 1 shows the local time series of the Bellcore BC-pAug89 LAN trace when using the variance local analysis method. The block size is 1000 points. Therefore, a 1000 points local time series is obtained. As can be seen the local Hurst time series are extremely variable indicating the presence of dependence structure. A test for self-similarity in this time series resulted in a Hurst exponent of 0.6214 when using the R/S-statistic, 0.584 when using the absolute moment method and 0.596 when using the variance method. From the previous values one can infer that there exists a relationship or dependence among these local values.

Figure 2 shows the local time series that resulted when variance local analysis was applied to the trace **BC-pOct89**. As in the previous case, variability is also observed. A test for self-similarity resulted in a Hurst exponent of 0.788 when using the R/S-statistic, 0.86 when using the absolute moment method and finally when applying the variance method a Hurst exponent of 0.83 is obtained.

Figure 3 shows the time series that resulted when variance local analysis was applied to the Bellcore trace **BC-pOct89Ext**. The variability as in the previous cases is also present and when applying a test for long-range dependent behaviour using the R/S-statistic, the Hurst parameter is 1.04. Applying the absolute moment method, the Hurst parameter is 0.914 and when using the variance method it is 0.927.
5. Cumulative analysis of self-similar traces

In this section the attention is now turned to the cumulative analysis of self-similar network traces. By cumulative analysis we mean the division of a trace of length $N$ in blocks of length $m$ and then estimating the Hurst exponent for blocks of size $im$ where $i = 1, 2, ..., N/m$. From this a new time series of $N/m$ points is obtained, from which a cumulative analysis can be done and its behaviour in time can be checked. For example in a cumulative analysis of a 10,000 points time series, the division of the series is first done in blocks of size $m$, say 1000 points from which we will obtain a new series of size $N/m$, in this case a 10 point time series. The first point of the new series represents the Hurst exponent of the first 1000 points of the original time series, the second point will represent the Hurst exponent of the first 2000 points of the original series, the third point will represent the Hurst exponent of the first 3000 points of the original time series and in general the $i$ point will represent the Hurst exponent of the first $im$ points of the original time series until $i \leq N/m$. This type of cumulative analysis is done to several kinds of traces. The objective of cumulative analysis is trying to empirically infer the number of points required for obtaining an accurate estimation of the Hurst exponent. In [1],[2], the traces used consisted of 1,000,000 points and a day of measurements. In [7] the trace analyzed consisted of more than 7 million points and several days of measurement. In other studies[4] tens of minutes of measurement were required to get a final estimation of the Hurst parameter in VBR video traffic. As can be seen the number of points required to obtain a final estimate varies. In this section we try to answer the question of how many points are required to get a final estimation of the Hurst parameter for different types of traces. This is done via the cumulative analysis of self-similar traces. The required number of points for different types of self-similar traces is provided, it means, the inference of the number of points required for accurate estimation of the Hurst parameter for Local Area Network, LAN, variable-bit-rate, VBR, and for Fractional Gaussian Noise, FGN, traces. The cumulative analysis can be done with any time-domain or frequency-domain technique for self-similarity testing. In this paper, we study the following types of tests:

- Cumulative R/S analysis.
- Cumulative periodogram analysis.
- Cumulative variance analysis.
- Cumulative variance of residuals analysis.
We will begin studying some VBR video traces presented in [4]. The characteristics of these series were not presented in section III and we will mentioned each briefly. Figure 5 shows the cumulative R/S analysis of VBR video traces *star wars*, *silence of the lambs* and *mr bean* with block sizes of 1024 points. The first two traces are 89,998 points in length and have a *Hurst* exponent of 0.903 and 1.007 respectively, that was estimated via R/S-statistic. The last one is a 89057 points time series with *Hurst* exponent of 0.933 estimated via R/S. As can be seen from Figure 5, the first 15,000 points of the series does not give us a good estimate of the *Hurst* parameter and from this we can state that for *Hurst* parameter estimation of VBR traces the length of the time series must be greater than 15,000 points. From 15,000 to 30,000 points the *Hurst* parameter shows almost correct Hurst estimates, except that in some cases there exist errors of 0.02 in all three traces, but it is a good estimate and it would be useful in situations when the time is a constraint. From 30,000 to the end of the trace the cumulative *Hurst* values tend to stabilize but again there exists errors in some parts. For the first blocks, the errors were not greater that 0.35, but for the last ones good estimate of the *Hurst* exponent were obtained. From the above-mentioned, we can infer that for VBR video traffic traces the optimum number of points in a trace must be at least 20,000 points. This is very useful since can save a considerable amount of time of measurement.

Now the study of a Local Area Network, LAN, trace is presented, the length of the trace is 1,000,000. This would allow us to see better how the Hurst parameter stabilizes and would allow us to compare the errors of LAN traces against VBR video traces and the pattern of stabilization of each. Figure 6 shows the cumulative R/S analysis for the *BC-pAug89* trace from Bellcore using a block size of 1024 points. The *Hurst* exponent of this trace was reported to be 0.80 when the R/S-statistic was used. As can be seen the first 20,000 points show the same behaviour as the VBR traces studied before but in this case the variability is less which in turn gives smaller errors, again as in the VBR the points from 20,000 to 90,000 tend to have small errors. In contrast to VBR traces this LAN trace show smaller errors. From 150,000 to the end of the trace the *Hurst* exponent behaves as a constant of value 0.8, the *Hurst* exponent of the total time series. From figure 6 can inferred that for obtaining an accurate estimation of the *Hurst* exponent of a LAN trace we must have at least 100,000 points. This would allow us to save a considerable amount of time of measurement since instead of days of measurements we would have hours of measurement. When using a different technique for self-similar testing, similar results are obtained.

Next, a study of the cumulative behaviour of synthetically generated fractional gaussian noise, FGN self-similar traces is presented. We study three sets of traces with length of 262,144 points and with *Hurst* parameters of 0.70, 0.80 and 0.90. These traces were generated by the fast Fourier transform method proposed in [10]. Figure 7 shows the cumulative R/S analysis of the synthetic traces. As can observed, the first 20,000 points tend to show the same behaviour of the previous traces and the *Hurst* exponent in this interval does not correpond to the real *Hurst* exponent of the trace. From 20,000 to 60,000 points the *Hurst* exponent tends to stabilize but again showing an error of 0.03 from the real *Hurst* exponent of the original time series. The real and synthetic traces shows both negative and positive errors. This alternating error or behaviour of real traces could be due to the additional degree of SRD in the trace, since it is known that SRD in the trace affects the efficiency or accuracy of the estimators. From 60,000 to the end of the trace the *Hurst* exponent tends to be stable and showing an error that is typically of 0.02 from the real *Hurst* exponent. From this we can infer that when a trace is purely LRD as in the synthetic traces we need at least 20,000 points.
6. Conclusions and future work

In this paper we have presented a local and cumulative analysis of several real and synthetic self-similar traffic traces. We have found that the points of local self-similarity analysis, i.e. the H-values, are dependent, it means they show self-similarity. This in turn demonstrates that it is not possible to infer the degree of self-similarity in a trace by just selecting small portions of a trace, estimate its Hurst exponent and then computing for example a mean. We have also studied the cumulative analysis of several real and synthetic self-similar traffic traces that allow us to infer the number of points required to obtain an accurate estimation of the Hurst parameter. We have found that the number of points required for accurate self-similarity testing depends of the type of trace, it means, that the number of points in a LAN trace is different of the number of points in a VBR trace. From the cumulative analysis we can conclude that:

- For VBR video traces the number must be at least 20,000 points.
- For LAN traces the number of points must be at least 100,000 points.
- For fractional Gaussian noise traces the number of points must be at least 20,000 points.

Additionally, we have noted that the error of the Hurst exponent in real traces alternate between negative and positive numbers at the beginning of the trace in the cumulative analysis. This alternating behaviour could be due to the SRD present in real traces since it has been shown that SRD affects the accuracy of the self-similar estimators Unfortunately, we have only studied the local and cumulative behaviour for a small number of traffic types, specifically LAN traffic and VBR video traffic. In the future, we expect to complement this work with a local and cumulative analysis of traffic such as WWW traffic and Wireless traffic and compute the dependence and the number of points required for accurate estimation of Hurst exponent. Additionally we expect to show mathematically the behaviour just observed.

7. References