Long-Range Dependence: Now you see it now you don't!

T. Karagiannis,M. Faloutsos and R. H. Riedi

April 10, 2002

Abstract

Over the last few years, the network community has started to rely heavily on the use of novel concepts such as self-similarity and Long-Range Dependence (LRD). Despite their wide use, there is still much confusion regarding the identification of such phenomena in real network traffic data. In this paper, we show that estimating Long Range Dependence is not straightforward: there is no systematic or definitive methodology. There exist several estimating methodologies, but they can give misleading and conflicting estimates. More specifically, we arrive at several conclusions that could provide quidelines for a systematic approach to LRD. First, long-range dependence may exist even, if the estimators have different estimates in value. Second, long-range dependence is unlikely to exist, if there are several estimators that do not "converge" statistically to a value. Third, we show that periodicity can obscure the analysis of a signal giving partial evidence of long range dependence. Fourth, the Whittle estimator is the most accurate in finding the exact value when LRD exists, but it can be fooled easily by periodicity. As a case-study, we analyze real round-trip time data. We find and remove a periodic component from the signal, before we can identify long-range dependence in the remaining signal.

1 Introduction

Self-similarity and long-range dependence (LRD) have become key concepts in analyzing networking traffic data over the past years. The community recognizes their overwhelming evidence of in multiple facets such as traffic load, and packet arrival times. Simply put, most researchers expect to identify and use LRD in their analysis or simulations. However, there are two important questions related to long-range dependence that have not received as much attention: a) how can we calculate it accurately, b) what does it really mean for network analysis and modeling? In this paper, we focus on the first question, since it is a necessary step to answer the second question.

Surprisingly, despite its ever-increasing use, there does not exist a definitive systematic way to calculate long-range dependence. The question is simple: given a time series does it exhibit long-range dependence? The predominant way to quantify long-range dependence is the value of the *Hurst exponent*, which is a scalar. So, the question becomes how we can calculate the Hurst exponent. It turns out that this is not straighforward. For one, the Hurst exponent can not be calculated in a definitive way, it can only be estimated. Second, there are several different methods to estimate the Hurst exponent, but they often produce conflicting estimates. It is not clear which of the estimators provides the most accurate estimation. As a result, there is no systematic method or a common reference point that would make the use of long range dependence in a reliable and reproducible way. As a consequence, studies can often arrive arbitrary and misleading conclusions.

The goal of this paper is to shed some light in the estimation of long-range dependence motivated by the absence of such a systematic approach. In addition, we also want to draw the attention of the community to this problem. We start with a "reverse engineering" approach: we observe the results of the estimators on a series of artificial and real signals. Our ambition is to be able we can "interpret" the profile of an unkown signal using our library of profiles. Through this work, we also develop guidelines for a systematic approach to the estimation of long-range dependence. More specifically, we test the estimators with three different types of data.

- *Synthetic data with known LRD value (for accuracy).* We find that the values of the estimators can differ significantly.
- *Artificial non-LRD data (for sensitivity).* We find that it is easy to fool several of the estimators. Specifically, we find that periodicity poses a serious threat.
- *Measured round-trip time from the Internet.* We find that the round-trip time is characterized by a strong periodic component[1], and only after this is remover, we can identify long-range dependence.

An additional contribution is the tool, SELFYS, that we developed for the purpose of this analysis. It is a

[1] We have not traced the origin of the periodicity. However, the focus of the paper is on describing effectively real data.

collection of LRD estimators, generators, and time series analysis methodologies. SELFYS is a java-based, open-source, tool provided as a service to the community.

The rest of this paper is organized as follows. Section 2 provides background work and the mathematical definitions of self-similarity and long-range dependence. Section 3 shows the evaluation of long-range dependence estimators and presents cases that can deceive the estimators. Section 4 is a study of long-range dependence in RTT delay in the Internet. Section 5 concludes the paper.

2 Definitions - Background

Self-similarity is observed when a time series has the same autocorrelation function at different levels of aggregation. That is, a stationary time series X_t is self-similar, if we define the aggregated series $X_k^{(m)}$ using different block sizes m, and X_t has the same autocorrelation function r with $X_k^{(m)}$ for each aggregation level m (where $r(k) = E[(X_t - \mu)(X_{t+k} - \mu)]/\sigma^2)$. Intuitively, this means that a time series presents the same statistical properties at different aggregation levels. If the autocorrelation function follows a power law, that is $r(k) \sim k^{-\beta}$ as $k \to \infty$ then the process is said to have long-range dependence. A metric of self-similarity is the Hurst exponent (H). Long-range dependence is characterized by $0.5 < H < 1$.

There are many estimators that are used to estimate the value of the Hurst exponent. In this paper we evaluate the following estimators:

- *Absolute Value method*, where the log-log plot of the aggregation level versus the absolute first moment of the aggregated series $X^{(m)}$ should be a straight line with slope of H-1, if the data are long-range dependent (where H is the Hurst exponent).

- *Variance method*, where the log-log plot of the sample variance versus the arrgegation level must be a straight line with slope β greater than -1. In this case $H = 1 - \frac{\beta}{2}$.

- *R/S method.* A log-log plot of the R/S statistic versus the number of points of the aggregated series should be a straight line with the slope being an estimation of the Hurst exponent.

- *Periodogram method.* This method plots the logarithm of the spectral density of a time series versus the logarithm of the frequencies. The slope provides an estimate of H.

- *Whittle* estimator. The method is based on the minimization of a likelihood function, which is applied to the periodogram of the time series.

- *Variance of Residuals.* A log-log plot of the aggregation level versus the average of the variance of the residuals of the series should be a straight line with slope of H/2.

- *Abry-Veitch.* Wavelets are used in order to estimate the Hurst exponent.

The ability of self-similarity based modeling to better fit Internet data than traditional methods, has been well documented over the past few years. Willinger and Paxson in [9] present the failure of the Poisson process to capture Internet traffic. Furthermore, different types of network traffic are shown to be dominated by long-range dependence phenomena [3], [10], [7], [1]. The relevance of LRD in network traffic is studied in [4], while in [8] a new wavelet method for synthesizing LRD series is developed.

3 Evaluating the estimators

This section presents an evaluation of the methodologies that are used to estimate the Hurst exponent. In the first part of the section, we use Fractional Gaussian Noise generators in order to generate long-range dependent series and study the behavior of the estimators. In the second part, we show that non long-range dependent signals can be identified as long-range dependent by some of the estimators.

3.1 Evaluating the estimators using Fractional Gaussian Noise

The evaluation of each estimator is achieved through three different Fractional Gaussian Noise (FGN) generators.FGN generators are often used to synthesize long-range dependence series with a specific Hurst value. The description of the first two can be found in [6] and [2]. The third is based in the Durbin-Levinson coefficients. Due to space limitation, we only present results from the generator developed by Paxson. However, findings are similar for the other two generators.

For each of the three generators we produce samples with different levels of long-range dependence. That is we produce samples of length 4096 with Hurst exponent between 0.5 and 1.For each of these samples, we use the methodologies described in the previous section to estimate the Hurst exponent. Table 1 summarizes our findings for the Paxson generator. The first column shows the Hurst exponent value of the generated series, while the rest columns show the corresponding estimation for each estimator. Since the Whittle estimator and the Abry-Veitch estimator produce confidence intervals next to these columns we present the confidence intervals for these two estimators. [2]

Observing table 1, one can conclude that Whittle is the most robust estimator. The Periodogram also gives satisfying estimations. These conclusions agree with the observations in [5]. The Abry-Veitch estimator seems to overestimate H, while the rest cannot provide sufficient estimations with the exception of RSplot when H is less than 0.8.

3.2 Deceiving the estimators

This subsection shows that the estimators are sensitive in various types of signals. Our goal is to identify cases that would confuse the estimators. In particular we apply the estimators in synthesized signals such as cosine functions with noise or signals that show trend. The following cases are considered.

- *Cosine + White Gaussian Noise (WGN).* The series is synthesized by WGN and the following cosine function : $Acos(\alpha x)$. Table 2 presents results for different values of the amplitude (A) of the cos function. In this case $\alpha = 0.005$. On the other hand, table 3 presents results if $A = 1$ and α varies. Both table show only the estimators that produce estimates for the corresponding signal.

- *FGN series + WGN.* Table 4 presents the results of the estimators when applied to FGN with WGN series. The values in the parenthesis shows the estimation of the raw FGN data. The purpose of this as well as of the next case, is to study the effect of noise and periodicity in LRD signals

- *FGN series + a cosine function.* Table 5 presents the results of the estimators when applied to FGN with periodic components ($cos(0.005x)$). The values in the parenthesis show the estimations if the amplitude of the cosine function is multiplied by three.

- *Trend.* We applied the estimators in various signals that showed a trend. Such signals included combination of WGN and cosine functions with trend. In every case only Whittle gives an estimation for Hurst which is always .99. Also the Periodogram estimates Hurst to be greater than 1.

Summing up, we observed the following:

1. When the data are generated by FGN, Whittle and Perdiodogram seem to give the most accurate estimation for the Hurst exponent.

2. Perdiodicity can mislead the Whittle, the periodogram and the R/S method into falsely reporting LRD. Especially, if the amplitude is large and the period small, then Whittle always estimates Hurst to be 0.99. However, it is interesting to note that Whittle estimates Hurst to be 0.99 even in a plain cosine signal.

3. White noise affects the accuracy of Whittle (by 0.17) more compared to the other estimators (less than 0.04)

4. Trend also misleads Whittle which reported a Hurst value of .99 in every signal with trend.

[2]Throughout this paper, the results presented correspond to confidence coefficients of 97% and 95% confidence intervlas.

4 Long-Range Dependence in Round Trip Time

This section presents a real case study of the Hurst exponent estimators. We apply the estimators in real Internet RTT traces. The set of data includes measurements for one route within the United States, from UCR to CMU. For this route, we measure the Round Trip Time for different packet sizes and different sending rates with the aid of NTP servers. The measurements took place from October 6 to October 9 (Saturday-Monday). The sending rates range from 20msec to 1sec. The packets are sent back-to-back according to the selected sending rate for six minutes every 30 minutes. Hence, for every day there are 48 different six-minute datasets.

To extract the useful information from the raw RTT data, we applied specific time series methodologies like, interpolation to recover from loss (so that our signal would not have discontinuities), removal of outliers and smoothing. Applying the estimators in the RTT signal, resulted in non-consistent estimations, in the sense that some of the estimators showed long-range dependence for some of our datasets. However, further analysis of the signal showed that it is dominated by periodic components. In particular, there was increasing energy in the signal every 5sec. This was true for 85% of our datasets. Removing the periodicity from the signal and applying the Hurst estimators in the new signal reveals long-range dependent behavior. For almost all of our datasets H is found to be between 0.55 and 0.68 by the majority of the estimators. Figures 1 and 2 show a RTT signal, the periodicity and two of the estimators before and after the removal of the periodicity.

5 Conclusions

The goal of this paper is to provide the first steps towards a systematic approach to long-range dependence analysis. We find that this is an essential task, given the increasing interest of the community for long-range dependence. We show that identifying long-range dependence is not straighforward: the estimators have conflicting results. Our work provides some general rules on interpreting these inconsistent results. In addition, we provide a tool that integrates most of the known required functionality for such analysis.

Our work leads to the following conclusions:

- There is no single estimator that can provide a definitive answer. For example, Whittle is the most accurate when LRD exists, but can be mislead in showing LRD by periodic non-LRD data.
- Long-range dependence may exist even, if the estimators have different estimates in value
- Long-range dependence is unlikely to exist, if there are several estimators that do not "converge" statistically to a value
- periodicity can obscure the analysis of a signal giving partial evidence of long range dependence.

We also applied the estimators in real RTT data. RTT is both periodic and long-range dependent. In particular, we showed that RTT is dominated by a periodic component of 5sec. The long-range dependent characteristics of the RTT signals are revealed only after the periodicity is removed.

Finally, our work provides the following tips for practitioners.

- A visual inspection of the signal can be very useful revealing many of its features, like periodicity[3]
- For efficient characterization, it may be necessary to process and decompose the signal.
- Researchers should not rely only in one estimator in deciding the existence of long-range dependence. As we saw, several of the estimators (Whittle, Periodogram) can be overly optimistic in identifying long-range dependence.
- A reporting of the Hurst exponent is meaningful, only if its accompanied by the method that was used, as well as the confidence intervals or correlation coefficient.

References

[1] A. Veres, Z. Kenesi, S. Molnar and G. Vattay. On the Propagation of Long-range Dependency in the Internet. In *SIGCOMM*, 2000.

[2] Edgar E. Peters. *Chaos and Order in the Capital Markets*, page 211. John Wiley & Sons, New York, 1991.

[3] M. E. Crovella, and A. Bestavros. Self-Similarity in World Wide Web Traffic Evidence and Possible Causes. In *IEEE/ACM Transactions on Networking*, 1997.

[4] M. Grossglauser, and J. Bolot. On the Relevance of Long-Range Dependence in Network Traffic. In *IEEE/ACM Transactions on Networking*, 1998.

[5] M. S. Taqqu, and V. Teverovsky . On Estimating the Intensity of Long-Range Dependence in Finite and Infinite Variance Time Series. In R. J. Alder, R. E. Feldman and M.S. Taqqu, editor, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, pages 177–217. Birkhauser, Boston, 1998.

[6] Vern Paxson. Fast approximation of self similar network traffic. Technical Report LBL-36750, 1995.

[7] R. H. Riedi and W. Willinger. *Toward an Improved Understanding of Network Traffic Dynamics*. Self-similar Network Traffic and Performance Evaluation eds. Park and Willinger, (Wiley 2000).

[8] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A Multifractal Wavelet Model with Application to Network Traffic. In *IEEE Special Issue on Information Theory*, pages 992–1018, 1999.

[9] W. Willinger, and V. Paxson. Where Mathematics Meets the Internet. In *Notices of the AMS*, 1998.

[10] W. Willinger, V. Paxson, R. H. Riedi and M. S. Taqqu. Long-range Dependence and Data Network Traffic. In *Long-Range Dependence: Theory and Applications*, 2001.

[3]We recommend plotting the signal at several different scales, since each scale can reveal different characteristics.

Table 2: Estimators predictions for the signal $Acos(0.005x)$. Increasing the amplitude, increases the estimation for the Hurst exponent.

A	*Period*	*R/S*	*Whittle*	*C.I.*
0.3	0.6	0.72	0.55	0.54-0.56
1.3	0.88	0.95	0.72	0.71-0.74
2.3	1	0.98	0.8	0.79-0.82
3.3	1.17	0.98	0.85	0.84-0.87
4.3	1.2	0.96	0.89	0.88-0.91

Table 3: Estimators predictions for the signal $cos(\alpha x)$. Increasing the frequency, increases the Hurst value in Whittle, while decreases in Periodogram and R/S.

α	*Period.*	*R/S*	*Whittle*	*C.I.*	*ABS*	*Variance*
0.01	0.55	0.82	0.7	0.68-0.71	-	-
0.08	0.59	0.56	0.72	0.71-0.74	-	-
0.09	0.55	0.53	0.72	0.71-0.73	-	-
0.1	0.53	0.54	0.72	0.71-0.74	0.35	0.38
0.16	0.43	0.47	0.73	0.71-0.74	0.41	0.44

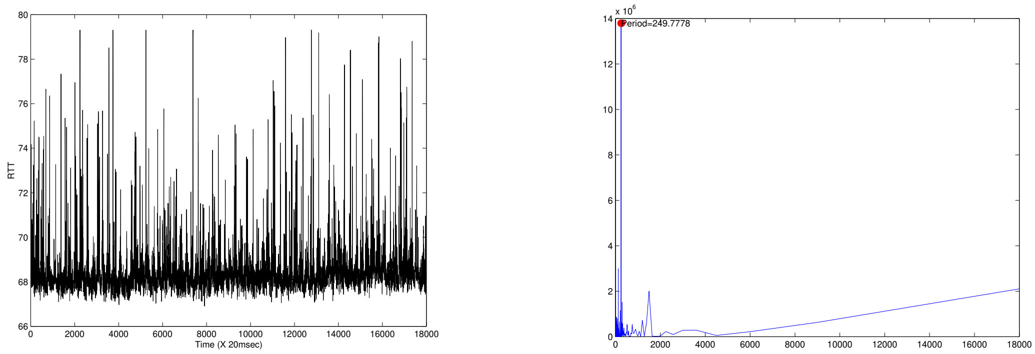

Figure 1: A sample RTT signal and the 5sec (index * 20msec sending rate) periodicity (power vs period)

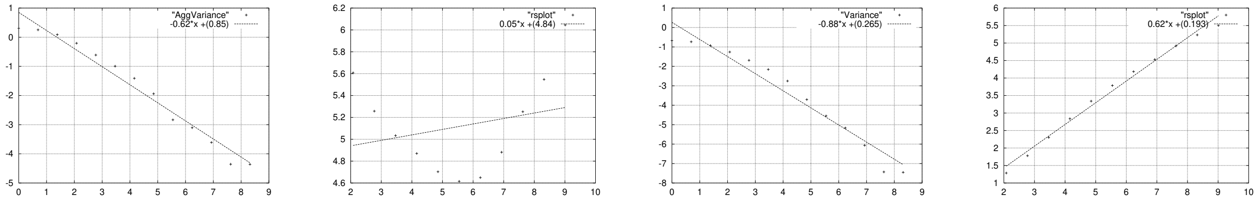

Figure 2: Average method and RSplot before (left two) and after the removal of the dominating periodic components. Both methods show LRD once the periodicity is removed.

Table 1: Estimators results using Paxson's Generator. Whittle and the Periodogram estimate more accurate the generated FGN series

H	*ABS*	*Variance*	*Period*	*Residuals*	*R/S*	*Whittle*	*C.I.*	*Abry-Veitch*	*C.I.*
0.5	0.43	0.46	0.52	0.44	0.55	0.5	0.48-0.52	0.54	0.52-0.57
0.6	0.53	0.55	0.62	0.52	0.63	0.59	0.57-0.61	0.65	0.62-0.67
0.7	0.61	0.63	0.72	0.61	0.7	0.69	0.67-0.71	0.75	0.73-0.78
0.8	0.69	0.71	0.82	0.7	0.77	0.79	0.77-0.81	0.86	0.83-0.88
0.9	0.76	0.78	0.92	0.78	0.83	0.89	0.87-0.91	0.96	0.93-0.98
0.95	0.79	0.81	0.97	0.82	0.85	0.94	0.92-0.96	1	0.98-1
0.99	0.81	0.83	1	0.85	0.87	0.98	0.96-1	1	1-1

Table 4: Estimations for generated FGN series with White Gaussian Noise. The values in the parenthesis show the estimation of the raw FGN data. Noise affects most Whittle

Hurst	*Period*	*R/S*	*Whittle*	*Residuals*	*ABS*	*Variance*
0.5	0.5 (0.48)	0.58 (0.56)	0.5 (0.5)	0.49 (0.44)	0.45 (0.41)	0.48 (0.43)
0.7	0.64 (0.68)	0.69 (0.72)	0.63 (0.7)	0.6 (0.62)	0.59 (0.6)	0.62 (0.61)
0.9	0.86 (0.88)	0.83 (0.85)	0.73 (0.9)	0.76 (0.78)	0.71 (0.74)	0.75 (0.76)

Table 5: Estimations for generated FGN series with a cosine function ($cos(0.05x)$). The values in the parenthesis show the estimation if the amplitude of the cosine function is multiplied by three. All estimations are affected by the periodicity.

Hurst	*Period*	*R/S*	*Whittle*	*Residuals*	*ABS*	*Variance*
0.7	0.7 (0.78)	0.69 (0.59)	0.82 (0.99)	0.63 (0.66)	0.5 (-)	0.54 (-)
0.9	0.9 (0.95)	0.8 (0.66)	0.98 (0.99)	0.78 (0.78)	0.68 (0.52)	0.72 (0.59)