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

Thomas Karagiannis
CSE Dept., UC Riverside
tkarag@cs.ucr.edu

Michalis Faloutsos
CSE Dept., UC Riverside
michalis@cs.ucr.edu

Rudolff H. Riedi
ECE Dept., Rice University
riedi@rice.edu

***Abstract*— Over the last few years, the network community has started to make heavy 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 guidelines for a systematic approach to LRD. First, long-range dependence may exist even, if the estimators have different estimates of the Hurst exponent in the interval 0.5-1. Second, long-range dependence is unlikely to exist, if there are several estimators that fail to estimate the Hurst exponent. 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.**

I. 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 in multiple facets such as traffic load and packet arrival times. Simply put, most researchers expect to identify and use LRD in their analysis. 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 straightforward. 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. Limitations and pitfalls in long-range dependence estimation have also been observed [1] [2]. As a result, there is no common reference point that would make the use of long range dependence reliable and reproducible. 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. 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 to "interpret" the profile of an unknown 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 to accurate inference of LRD.
- *Measured round-trip time from the Internet.* We are interested in the performance from an application point of view. We find that the round-trip time is characterized by a strong periodic component, and only after this is removed, 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 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 II provides background work and the mathematical definitions of self-similarity and long-range dependence. Section III shows the evaluation of long-range dependence estimators and presents cases that can deceive the estimators. Section IV is a study of long-range dependence in RTT delay in the Internet. Section V concludes the paper.

This material is based upon work supported by the DARPA award N660001-00-1-8936

II. Definitions - Background

A stationary process X_t has long-memory or is long-range dependent [3], if there exists a real number $\alpha \in (0, 1)$ and a constant $c_p > 0$ such that $\lim_{k\to\infty} \rho(k)/[c_p k^{-\alpha}] = 1$ where $\rho(k)$ the sample correlations. The classical parameter that characterizes long-range dependence is the Hurst exponent (H), where $H = 1 - \alpha/2$. Long-memory occurs when $\frac{1}{2} < H < 1$. Intuitively, events that are far apart are correlated, since the correlations decay very slowly to zero. On the contrary, short-range dependence is characterized by quickly decaying correlations (e.g. ARMA, MARKOV processes).

There are many estimators that are used to estimate the value of the Hurst parameter. An overview of a large number of the estimation methodologies can be found in [4], [3]. In this paper we evaluate the following estimators:

- *Absolute Value method*, where an aggregated series $X^{(m)}$ is defined, using different block sizes m. 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.
- *Variance method*, where the log-log plot of the sample variance versus the aggregation 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 [5] 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 [6], [7], [8], [9], [10]. In addition, the relevance of LRD in network traffic is studied in [11].

III. 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 the estimators can be deceived to identify non long-range dependent signals as long-range dependent. We reach the following main conclusions: a) There is no ultimate estimator that can apply to every case and b) Periodicity, non-stationarity and noise affect the outcome of the estimators.

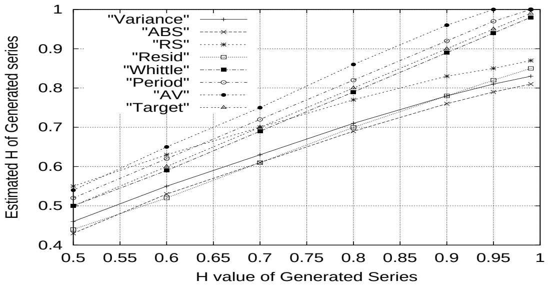

Fig. 1. The performance of the estimators using Paxson's generator. The "Target" line shows an optimal estimation of the FGN data. The Whittle and Periodogram estimators follow best the Target line.

A. 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 first is developed by Paxson [12], while the second is described in [13]. 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 65536 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.

Fig. 1 and Table I summarize our findings for the Paxson generator. In fig. 1, the X axis presents the Hurst exponent value of the FGN series and the Y axis shows the estimation of the corresponding methodology. The "Target" line presents what the optimal estimation of the FGN data for each case would be. In table I, 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. [1]

[1] Throughout this paper, the results presented correspond to correlation coefficients of 97% and 95% confidence intervals.

Observing table I, one can conclude that Whittle is the most robust estimator. The Periodogram also gives satisfying estimations. These conclusions agree with the observations in [4]. 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.

B. Deceiving the estimators

We show that the estimators are quite sensitive and can be deceived to report LRD. 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*. The estimators are applied to periodic datasets to study their behavior in non-LRD data. The series is synthesized by White Gaussian Noise and the following cosine function : $Acos(\alpha x)$. Table II presents results for different values of the amplitude (A) of the cos function. In this case $\alpha = 0.005$. On the other hand, table III presents results if $A = 1$ and α varies. Periodicity can mislead the Whittle, the Periodogram, the R/S and the Abry-Veitch methods into falsely reporting LRD. Especially, if the amplitude is large and the period small, then Whittle always estimates Hurst to be 0.99.
- *FGN series + White Gaussian Noise*. The effect of noise in LRD data is studied. In this case the Whittle and the Abry-Veitch estimators are the ones that are affected the most by noise. Table IV presents the results of the estimators when applied to FGN with White Gaussian Noise series. The values in the parenthesis show the estimation of the raw FGN data.
- *FGN series + a cosine function*. The effect of periodicity in LRD data is considered. Our findings show that periodicity affects all estimations. Table V 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*. The definition of LRD assumes stationary signals. In this case, we intend to identify the impact of non-stationarity on the estimators. Thus, we created various signals with slow and fast decaying or increasing trends. Such signals include combination of White Gaussian Noise 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 the section, our main observations are the following:

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

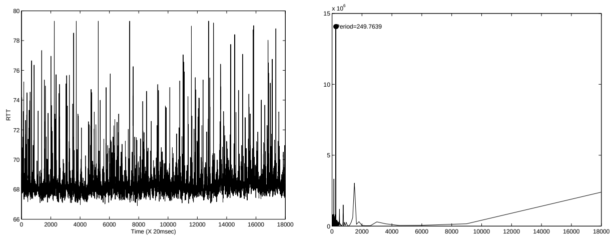

Fig. 2. UP: A sample RTT signal (RTT vs Time). Every value in the X axis represents packets spaced 20msec apart. DOWN: Power spectrum of the same RTT signal using Fourier Transform. Y axis is the power while the X axis is the period (1/frequency). The spike represents period of 5sec (index * 20msec).

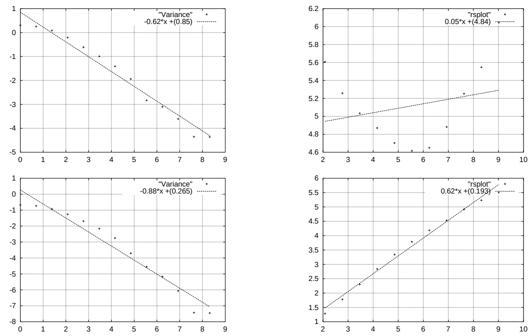

Fig. 3. UP: The Variance method and RSplot before the removal of the dominating periodic components. The estimators do not agree in the existence of long-range dependence. DOWN: The Variance method and RSplot after the removal of the dominating periodic components. Both methods show long-range dependence

2) There is no definite estimator that could be consistently used in every case. Each estimator evaluates different statistics of the signal to infer long-range dependence. Thus, different processes (e.g. noise, periodicity) have different effect on each estimator.
3) Even though the Whittle estimator is considered the most robust, it is the most sensitive of the estimators.

IV. 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. 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 typical 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, we observed a period of 5sec in the signal. This was true for 85% of our datasets for the various packet types or sending rates. However, it is interesting to note that we were able to trace a likely cause of the periodicity to a system maintenance tool in our network. This tool has an approximate period of 5 seconds according to our system administrator. We consider this as a verification of the integrity and effectiveness of our analysis. Note that the end-to-end performance of an application would be affected from such a phenomenon. 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. Fig. 2 and 3 show a RTT signal, the periodicity and two of the estimators before and after the removal of the periodicity.

V. 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 straightforward: 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, provided that the estimates show that $0.5 < H < 1$.
- Long-range dependence is unlikely to exist, if there are several estimators that cannot produce sufficient estimations of the Hurst exponent. (e.g. low confidence intervals).
- 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, we list a set of tips for practitioners, that we realized during our study.

- A reporting of the Hurst exponent is meaningful, only if it is accompanied by the method that was used, as well as the confidence intervals or correlation coefficient.
- Researchers should not rely only on one estimator in deciding the existence of long-range dependence (e.g. [14]). As we saw, several of the estimators (Whittle, Periodogram) can be overly optimistic in identifying long-range dependence.
- For efficient characterization, it may be necessary to process and decompose the signal.
- A visual inspection of the signal can be very useful, providing a qualitative analysis and revealing many of its features, like periodicity.[2]

Putting things in perspective, the overarching goal is to analyze and model the network behavior. And thus, long-range dependence is a powerful tool in this effort. Estimating long-range dependence in a robust and definitive way is an essential step, because only then, we can explore its ability to model effectively real network behavior. We find that there is still a lot of work that needs to be done both in estimating and interpreting long-range dependence.

References

[1] S.Molnar and T. D. Dang, “Pitfalls in Long Range Dependence Testing and Estimation,” in *GLOBECOM*, 2000.

[2] M. Krunz, “On the limitations of the variance-time test for inference of long-range dependence,” in *IEEE INFOCOM*, 2001, pp. 1254–1260.

[3] J. Beran, *Statistics for Long-memory Processes*, Chapman and Hall, New York, 1994.

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

[5] W. Willinger, and V. Paxson, “Where Mathematics Meets the Internet,” in *Notices of the AMS*, 1998.

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

[7] 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.

[8] 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).

[9] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, “The Changing Nature of Network Traffic: Scaling Phenomena,” in *ACM Computer Communication Review*, 1998, vol. 28, pp. 5–29.

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

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

[12] Vern Paxson, “Fast approximation of self similar network traffic,” Tech. Rep. LBL-36750, 1995.

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

[14] Q. Li,and D.L. Mills, “On the long-range dependence of packet round-trip delays in Internet,” in *IEEE International Conference on Communications*, 1998, pp. 1185–1191.

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

TABLE I

Estimators results using Paxson's Generator. Whittle and the Periodogram estimate more accurately the generated FGN series

H	*ABS*	*Variance*	*Periodogram*	*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 II

Estimators predictions for the signal $Acos(0.005x)$. Increasing the amplitude, increases the estimation for the Hurst exponent. The dashes represent insufficient estimations due to low correlation coefficients.

A	*ABS*	*Variance*	*Periodogram*	*Residuals*	*R/S*	*Whittle*	*C.I.*	*Abry-Veitch*	*C.I.*
0.3	-	-	0.6	-	0.72	0.55	0.54-0.56	0.55	0.53 - 0.58
1.3	-	-	0.88	-	0.95	0.72	0.71-0.74	0.57	0.55 - 0.6
2.3	-	-	1	-	0.98	0.8	0.79-0.82	0.56	0.53 - 0.58
3.3	-	-	1.17	-	0.98	0.85	0.84-0.87	0.58	0.55 - 0.6
4.3	-	-	1.2	-	0.96	0.89	0.88-0.91	0.59	0.58 - 0.62

TABLE III

Estimators predictions for the signal $cos(\alpha x)$. Increasing the frequency, increases the Hurst estimation in Whittle and AV estimators, while decreases the estimation in Periodogram and R/S. The dashes represent insufficient estimations due to low correlation coefficients.

α	*ABS*	*Variance*	*Periodogram*	*Residuals*	*R/S*	*Whittle*	*C.I.*	*Abry-Veitch*	*C.I.*
0.01	-	-	0.55	-	0.82	0.7	0.68-0.71	0.58	0.55-0.6
0.08	-	-	0.59	-	0.56	0.72	0.71-0.74	0.71	0.69-0.74
0.09	-	-	0.55	-	0.53	0.72	0.71-0.73	0.71	0.69-0.74
0.1	0.35	0.38	0.53	-	0.54	0.72	0.71-0.74	0.73	0.7-0.75
0.16	0.41	0.44	0.43	-	0.47	0.73	0.71-0.74	0.73	0.71-0.76

TABLE IV

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 the Whittle and the Abry-Veitch estimators.

Hurst	*ABS*	*Variance*	*Periodogram*	*Residuals*	*R/S*	*Whittle*	*Abry-Veitch*
0.5	0.45 (0.41)	0.48 (0.43)	0.5 (0.48)	0.49 (0.44)	0.58 (0.56)	0.5 (0.5)	0.55 (0.54)
0.7	0.59 (0.6)	0.62 (0.61)	0.64 (0.68)	0.6 (0.62)	0.69 (0.72)	0.63 (0.7)	0.67 (0.75)
0.9	0.71 (0.74)	0.75 (0.76)	0.86 (0.88)	0.76 (0.78)	0.83 (0.85)	0.73 (0.9)	0.77 (0.96)

TABLE V

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. The dashes represent insufficient estimations due to low correlation coefficients.

Hurst	*ABS*	*Variance*	*Periodogram*	*Residuals*	*R/S*	*Whittle*	*Abry-Veitch*
0.7	0.5 (-)	0.54 (-)	0.7 (0.78)	0.63 (0.66)	0.69 (0.59)	0.82 (0.99)	0.85 (0.97)
0.9	0.68 (0.52)	0.72 (0.59)	0.9 (0.95)	0.78 (0.78)	0.8 (0.66)	0.98 (0.99)	1.03 (1.34)