SELFIS: A Tool For Self-Similarity and Long-Range Dependence Analysis

[Extended Abstract]

Thomas Karagiannis
CSE Dept., UC Riverside
Surge Building, University of California
Riverside, CA 92521
tkarag@cs.ucr.edu

Michalis Faloutsos
CSE Dept., UC Riverside
Surge Building, University of California
Riverside, CA 92521
michalis@cs.ucr.edu

ABSTRACT

Over the last few years, the network community has started to rely heavily on the use of novel concepts such as fractals, self-similarity, long-range dependence, power-laws. Especially evidence of fractals, self-similarity and long-range dependence in network traffic have been widely observed. Despite their wide use, there is still much confusion regarding the identification of such phenomena in real network traffic data. 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. In this extended abstract, we make a first step towards a systematic approach in estimating self-similarity and long-range dependence. We present SELFIS, a java-based tool that will automate the self-similarity analysis. To our knowledge, our software tool is the first attempt to create a stand-alone, free, open-source platform to facilitate self-similarity analysis. We show the use of our tool and describe the methodologies that currently incorporates in real Internet data. Finally, we present an intuitive approach to validate the existence of long-range dependence.

1. INTRODUCTION

Real data analysis has become challenging for engineers. Fractals, self-similarity, long-range dependence, power-laws, time-series analysis are used more and more in data analysis. However helpful these tools may be, they have often been a burden for practitioners. Many of the researchers are not familiar with all the possibilities and capabilities the statistical methodology has to offer. First, many of the new notions are fairly complex not only in definition but also in intuition. There is no systematic classification of concepts which results in confusion, partial understanding and misinterpretation of terms. Second, it is not clear yet how all these statistical analysis tools relate to engineering purposes. Finally, their multidisciplinary character leads to a lack of a comprehensive single source that could be the reference point. As a result, researchers are forced to implement similar, if not the same analytical tools. This results in repetition of effort, discrepancies in analyzing findings and difficulties in reproducing and comparing results.

The question this extended abstract addresses is: *How can we estimate long-range dependence and self-similarity efficiently?* Our goal is to facilitate practitioners by providing *SELFIS*, a software tool. We aim to provide a common software platform that unifies the effort of multiple research communities. The benefits will be significant:

- Leverage of expertise from different disciplines.
- Create a common point of reference that will provide repeatable and comparable results.
- Assist in spreading fractals and long-range dependence-analysis by making them easily accessible to and computable by non-experts.

The main characteristics of SELFIS are the following: a) It integrates three classes of functions: Self-similarity and long-range dependence analysis; Fourier and wavelet transforms; data processing and cleansing algorithms. b) It is implemented as an independent software tool, so that users will not need additional commercial software to employ it. c) Modular design allows for other researchers to contribute their source code. SELFIS can save researchers the time that is normally required for collecting, analyzing and programming sophisticated algorithms. Our ambition is to establish SELFIS as the de facto open-source software for time-series analysis. SELFIS has already attracted interest from the networking community. It will be used at ISI by John Heidemann's group. Also, it will be incorporated in Javasim at University of Illinois, Urbana-Champaign.

In addition we describe an intuitive approach to validate long-range dependence. This methodology has been used before [3], but has not received sufficient attention. We call this methodology bucket shuffling. Bucket shuffling is based on decoupling short-term from long-range correlations. This is achieved by shuffling parts of a time-series and the examination of the sample autocorrelation function. Moreover, we present the functionality of our tool by analyzing real data. The data consists of measurements, conducted for various routes inside and outside US. Specifically, we show that packet-loss demonstrates long-range dependent behavior in large time scales (1sec).

Our contributions can be summarized as follows:

- We develop a software tool, *SELFIS*: It is a java-based, portable, expandable, object-oriented, freely distributed as a service to the community. We intend to maintain our tool up to date and integrate more functionality from other developers.
- We present a straightforward, intuitive approach for long-range dependence detection and validation. We call this approach bucket shuffling.
- We show that packet-loss shows long-range dependent behavior in large time scales.

The rest of this extended abstract is organized as follows. Section 2 is a brief overview of self-similarity and long-range dependence and summarizes previous findings of self-similarity in network traffic. Section 3 presents SELFIS, our self-similarity tool. Section 4 is a case study that presents the functionality of SELFIS. It is divided in two parts: a) Bucket shuffling, an intuitive approach for long-range dependence detection and b) LRD as a case study in packet-loss. Section 5 concludes the paper.

2. DEFINITIONS - BACKGROUND

A stationary process X_t has long-memory or is long-range dependent [4], 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).

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 [16] 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 [5], [17], [11], [19], [1], [2]. The relevance of LRD in network traffic is studied in [6], while in [12] a new method based on wavelets for synthesizing LRD series is developed.

3. THE SELFIS TOOL

The SELFIS tool is developed to provide all the necessary functionality a network practitioner needs. Our ambition is that SELFIS will be the reference point in time-series analysis. It is a java-based, modular, extendible, freely distributed software tool, that can automate time-series analysis. We chose to develop an independent platform instead of relying on commercial products. Our purpose was to give to the community a ready to use tool, without further obligations of purchasing any software.

The SELFIS tool is a collection of self-similarity and long-range dependence estimation methodologies and time-series processing algorithms. It incorporates various long-range dependence estimators that reveal different characteristics of the analyzed series. Also SELFIS offers data processing methodologies and transforms, such as wavelets, Fourier transform, stationarity tests and smoothing algorithms. In addition, SELFIS provides the possibility of synthesizing long-range dependent time sequences, as it includes fractional Gaussian noise generators. The following subsections present analytically the capabilities SELFIS has to offer.

3.1 Long-Range Dependence Detection

SELFIS implements an intuitive approach for the detection and validation of long-range dependence, `Bucket Shuffling`. Bucket shuffling is based on decoupling short-range form long-range correlations in a series to infer the existence of long-range dependence. This is achieved through shuffling and the examination of the autocorrelation function, Specifically, the time series is divided in buckets of length b. Then two levels of shuffling can be applied:

- External Shuffling: The order of buckets is shuffled, whereas the contents of the buckets remain intact. This can be achieved by creating a new ordered series consisting of bucket ids. Each bucket is given incrementally an id starting from the beginning of the time series. Then we replace each bucket contents after the bucket-id series is shuffled. External shuffling results in preserving the time-series correlations up to the bucket length. Long-range correlations are distorted because of the shuffling. Thus, the autocorrelation function should not exhibit significant correlations beyond the bucket size.
- Internal Shuffling: The order of bucket remains the same as that of the original signal, whereas the contents of each bucket are shuffled. As a result, short-range correlations are distorted, whereas long-range correlations remain relatively unaltered. Hence, if the original signal has long-memory, the autocorrelation function of the internal-shuffled series should still show power-law behavior. Examples of bucket shuffling are presented in the next section.

3.2 Hurst Estimators

Various estimators can be used to provide estimates of self-similarity and long-range dependence. A number of methods, such as RSplot and the Variance method define an aggregated series $X^{(m)}(k)$ given a time series X_i. That is,

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2...., [\frac{N}{m}].$$

On the other hand there are the power spectrum methods like the periodogram estimator. Finally some methods use wavelets and Fourier transform to estimate the Hurst exponent, like the Abry-Veitch wavelet-based estimator in [8]. An overview of a large number of the estimation methodologies can be found in [7], [4]. In our tool the following estimators are included:

- *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 we plot on a log-log plot the sample variance versus the block size of each aggregation. If the series is self-similar with long-range dependence then the plot is a line with slope β greater than -1. The estimation of H is given by $H = 1 + \frac{\beta}{2}$.
- *R/S method.* This method uses the rescaled range statistic (R/S statistic). The R/S statistic is the range of partial sums of deviations of a time-series from its mean, rescaled by its standard deviation. 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. The periodogram is given by

  $$I(\nu) = \frac{1}{2\pi N} |\sum_{j=1}^{N} X(j) e^{ij\nu}|^2$$

  where ν is the frequency, N is the length of the time-series and X is the actual time series.
- *Whittle* estimator. The method is based on the minimization of a likelihood function, which is applied to the periodogram of the time-series. It gives an estimation of H and produces the confidence interval. It does not produce a graphical output.
- *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 for the Hurst exponent to be estimated. The energy of the series in various scales is studied to calculate the Hurst exponent.

3.3 Transforms

SELFIS includes the following transforms:

- Fourier Transform. Fourier transform is used to transform a signal from the time domain to the frequency domain. Intuitively, the signal is broken down into sinusoids of different frequencies.
- Wavelets (Haar and D4). Fourier transform cannot present information about the time. Wavelets cover for this inefficiency by combining frequency and time domains.
- Power Spectrum. The power spectrum presents the amount of energy that corresponds to each frequency of the Fourier transform.

3.4 Data Processing

Data processing is an essential element in time-series analysis. Processing reveals the underlying behavior of the series and allows for further analysis. SELFIS currently includes the following data processing methodologies:

- Smoothing Algorithms. Smoothing can be achieved by median, average or exponential smoothing algorithms. Our tool includes the 4253H smoothing algorithm described in [15]. The algorithm has been shown to provide sufficient results for different kinds of data. According to 4253H smoothing the signal is smoothed by successively applying median smoothing with window 4,2,5 and 3 followed by a hanning operation. A hanning operation multiplies the values of a window 3 by 0.25, 0.5 and 0.25 respectively, and sums the results.
- Stationarity tests. Stationarity means intuitively that there is no trend in the series. There is a number of tests that check a series for stationarity. One of the common tests for stationarity is the run test. The test can detect a monotonic trend in the series by evaluating the number of runs. A run is defined as a sequence of identical observations. The number of runs must be a random variable with mean $\frac{N}{2} + 1$ and variance $\frac{N(N-2)}{4(N-1)}$, where N is the length of the series. The number of runs is evaluated from a series $s(i)$, where:

  $$s(i) = 0 \text{ , if } y(i) < median(y), \text{ and}$$

  $$s(i) = 1 \text{ , if } y(i) \geq median(y),$$

  where y(i) is the time series. Stationarity is important when long-range dependence is studied, since estimators fail in non-stationary data. If stationarity is detected, the time series must be differenced successively until stationarity is achieved.

3.5 Fractional Gaussian Noise Generators

Fractional Gaussian Noise (FGN) generators can synthesize series with long-range dependence. The tool includes two generators. The first proposed in [10], is a method based on fast Fourier transform to generate a FGN series. The second generator produces FGN series by using the Durbin-Levinson coefficients. The latter algorithm was implemented in java, based on the source code written by Vadim Teverovsky in S-Plus [14].

4. CASE STUDY

This section presents SELFIS. We present two showcases of the capabilities of the tool. First, an application of bucket shuffling to validate long-range dependence in time-series. To demonstrate the methodology, we use Fractional Gaussian Noise series generated by one of the generators included in SELFIS. Second, initial observations regarding long-range dependence behavior of packet loss in real Internet traffic data are presented.

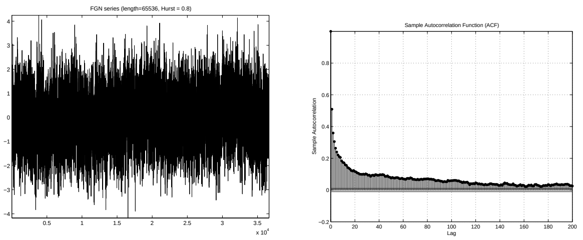

Figure 1: LEFT: FGN series of length 65536 and Hurst 0.8. RIGHT: Autocorrelation function (ACF) of the series up to lag 200. Clearly the ACF shows power-law like behavior.

4.1 Bucket Shuffling

Bucket shuffling (see precious section) is an intuitive, straightforward methodology that validates the existence of long-memory. To show how long-range dependence can be detected using bucket shuffling , we synthesized a sample series of fractional Gaussian noise. The series (fig. 1) has length 65536, Hurst exponent 0.8 and was synthesized using the generator created by Paxson. The right part of fig. 1 shows the sample autocorrelation function (ACF) of the series which clearly follows a power-law like behavior and implies long-range dependence. To ensure that long-range dependence really exists we employ bucket shuffling. Fig. 2 shows the ACF function after the signal is shuffled with three different ways. First, we shuffle externally with bucket size 1, in order to create a complete randomized signal. As one would expect ACF shows that no correlation exists (Fig. 2 up left). Second, external shuffle cancels the effect of long-range correlations. Clearly the up right part of Fig. 2 shows that there is no correlation beyond the bucket size. Finally, internal shuffle distorts the sort-term correlations, while not affecting the long-range behavior. It is obvious that the ACF presents the same behavior as that of the ACF of the original series. Hence, one can conclude that long-range dependent behavior dominates the original series since there are no short-term correlations.

4.2 LRD in Packet Loss

The set of data includes measurements for various routes inside and outside the United States. Within the United States measurements were conducted for one route, from UCR to CMU. Routes outside the US include measurements from UCR to Greece, Japan and Australia. For these routes, we collect various characteristics of the network such as the Round Trip, packet loss and delay jitter. Measurements are conducted for different packet sizes and different sending rates. The sending rates range from 20msec to 1sec. This section presents only initial results for packet loss, since the scope of this paper is to present the functionality of SELFIS.

Fig. 3 presents a packet-loss time-series. Each data point represents the number of lost packets per second. The sending rate for the specific series is 50msec. Hence, the maximum number of lost packets in a second is 20. The series represents measurements that took place from April 25, 4pm to April 26, 3am. Self-similarity in multiple time scales was observed also in [18] and [9]. Table 1 shows the estimation of Hurst exponent for the time-series. The estimators agree on the existence of long-range dependence, however the estimation of Hurst exponent ranges from 0.61 to 0.83. Variance in Hurst exponent estimation has been observed and analyzed in [13]. Fig. 4 shows the graphical output of two of the estimators, RSplot and the log-log plot of Variance of Residuals. The rest of the traces both for other days as well as for other routes show similar results. The main difference is in the intensity of long-range dependence, namely the value of the Hurst exponent.

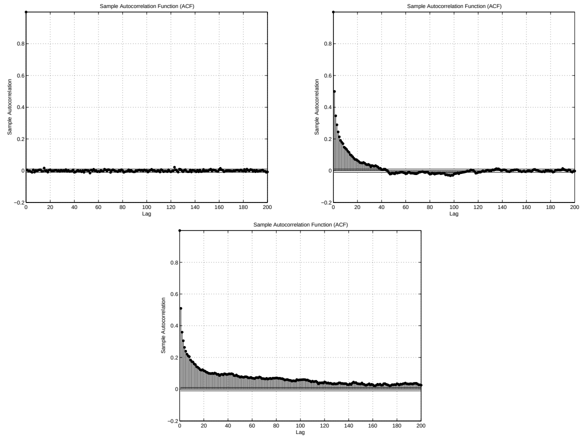

Figure 2: UP LEFT: External bucket shuffling with bucket size 1. Full randomize of the series results in no correlation to appear. UP RIGHT: External bucket shuffling with bucket size 50. After lag 50 all correlations are insignificant. DOWN: Internal bucket shuffling with bucket size 50. The ACF shows the same power-law behavior like the original series (Fig. 1).

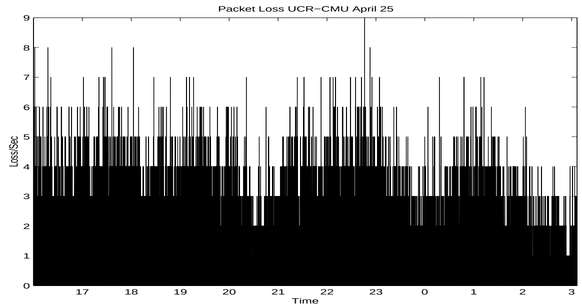

Figure 3: Packet loss per second for the route UCR-CMU (April 25, 4pm - April 26, 3am).

5. CONCLUSIONS

Through this work, we wish to facilitate practitioners by providing practical ways for long-range dependence estimation. We develop a software tool, SELFIS that can become a reference point in estimating long-range dependence in time-series. We provide a number of different estimators that capture various features of long-range dependent behavior. SELFIS is designed as a modular, open-source tool that is distributed freely so that it will incorporate more and more functionality in the future. To our knowledge this is the first attempt to collect all the long-range dependent estimators in a common platform without the need of any

Table 1: Estimations for packet loss from UCR-CMU. For each estimator the resulting value of Hurst exponent estimation is shown. All estimations are with 97% correlation coefficients.

Variance	*Resid*	*R/S*	*Whittle*	*AV*	*Period*
0.83	0.79	0.83	0.61	0.61	0.73

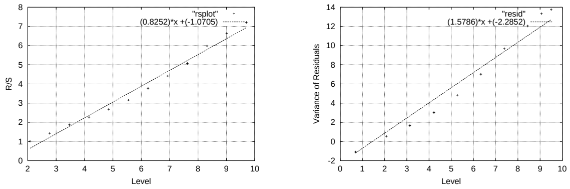

Figure 4: R/Splot and Variance of Residuals log-log plot for packet loss from UCR to CMU

commercial software.

We propose bucket shuffling as the ultimate test to detect and validate long-range dependence. The goal of bucket shuffling is to distinguish short-range from long-range correlations. The decoupling of correlations can show if the behavior of the original series is based on short-term behavior or long-memory. The appropriate selection of fitting and prediction models depends on identification of the length of correlation in the series.

We presented the functionality of SELFIS in a case study. In particular, we studied long-range dependence in packet-loss. We conclude that, packet-loss traces show long-range dependent behavior. We found that this is true in large time scales (1 sec). Various earlier measurements in literature with different datasets emphasize similar findings ([18],[9]). However, it is interesting to note that the estimators do not agree in their estimations and Hurst exponent estimation varies significantly.

SELFIS will be further extended with additional functionality in the future. Calculation of fractal dimensions and forecasting models are some of our priorities. In addition, we are very interested in collaborative development. Interested parties are highly encouraged to contribute code.

Summing up, long-range dependence is identified in increasing aspects of many disciplines such as, networking, databases, economics. Thus, the need for a complete long-range dependence analysis is crucial. SELFIS is a step towards this direction.

6. REFERENCES

[1] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz. The Changing Nature of Network Traffic: Scaling Phenomena. In *ACM Computer Communication Review*, volume 28, pages 5–29, 1998.

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

[3] A. Erramilli, O. Narayan, and W. Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223.

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

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

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

[7] 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.

[8] P. Abry and D. Veitch. Wavelet Analysis of Long-Range Dependence Traffic. In *IEEE Transactions on Information Theory*, 1998.

[9] P. Huang, A. Feldmann and W. Willinger. A Non-Intrusive, Wavelet-Based Approach to Detecting Network Performance Proelms. In *IMW*, 2001.

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

[11] 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).

[12] 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.

[13] T. Karagiannis, M. Faloutsos, R.H. Riedi. Long-Range dependence:Now you see it, now you don't! In *IEEE Global Internet*, 2002 [to appear].

[14] V. Teverovsky. http://math.bu.edu/people/murad/methods/.

[15] Velleman, Paul F., and Hoaglin, David C. *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, MA, 1981.

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

[17] 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.

[18] X. Tian, J. Wu and C. Ji. A Unified Framework for Understanding Network Traffic Using Independent Wavelet Models. In *IEEE INFOCOM*, 2002.

[19] Z. Sahinoglu and S. Tekinay. On Multimedia Netowkrs: Self-similar Traffic and Network Performance. In *IEEE Communications Magazine*, volume 37, pages 48–52, 1999.