# A User-Friendly Self-Similarity Analysis Tool [*]

Thomas Karagiannis, Michalis Faloutsos, Mart Molle
{tkarag,michalis,mart}@cs.ucr.edu
Department of Computer Science & Engineering
University of California, Riverside

## ABSTRACT

**The concepts of self-similarity, fractals, and long-range dependence (LRD) have revolutionized network modeling during the last decade. However, despite all the attention these concepts have received, they remain difficult to use by non-experts. This difficulty can be attributed to a relative complexity of the mathematical basis, the absence of a systematic approach to their application and the absence of publicly available software. In this paper, we introduce SELFIS, a comprehensive tool, to facilitate the evaluation of LRD by practitioners. Our goal is to create a stand-alone public tool that can become a reference point for the community. Our tool integrates most of the required functionality for an in-depth LRD analysis, including several LRD estimators. In addition, SELFIS includes a powerful approach to stress-test the existence of LRD. Using our tool, evidence are presented that the widely-used LRD estimators can provide misleading results. It is worth mentioning that 25 researchers have acquired SELFIS within a month of its release, which clearly demonstrates the need for such a tool.**

## 1. INTRODUCTION

Self-similarity, fractals, and long-range dependence (LRD) have emerged as powerful tools for modeling the behavior of real processes and systems. These concepts have been applied to numerous disciplines [25] ranging from molecular biology and genetics [3] [17] to geology [26]. However, in this work we focus on their application to the measured behavior of computer networks.

Following the seminal work of Leland et al. [16], long-range dependence has become a key concept in analyzing networking traffic data over the last decade. The community has observed an overwhelming manifestation of self-similarity in multiple network aspects such as traffic load, packet arrival times and queue sizes. As a result, most researchers expect to identify LRD in their analysis of measurements. Furthermore, realistic simulations require models that exhibit LRD.

Intuitively, the properties of LRD and self-similarity measure the importance of long-term memory in the evolution of a process over time. When applied to networking, we say that a time-dependent process (such as packet arrivals, queue lengths, etc.) is LRD if its current value is strongly correlated with its previous values far into the past. As a result, a sequence of measured values from an LRD process tends to generate similar (rather than independent) values. Thus, as is well known in the case of self-similar behavior, changing time scales through *aggregation* (i.e., taking the sum or average of a series of high resolution measurements to produce a single low resolution measurement) may have little impact on the apparent smoothness of an LRD process.

Recognizing the presence of LRD is important for practitioners, because it can significantly change the behavior of the network. For example, if the

[*] This work was supported by the NSF CAREER grant ANIR 9985195, and DARPA award NMS N660001-00-1-8936, and NSF grant IIS-0208950, and TCS Inc., and DIMI matching fund DIM00-10071.

packet arrival process were LRD, then larger input buffers would be needed to meet a given packet loss-rate specification. Moreover, the LRD property is completely different from ordinary notions of the variance of a random process. High variance only means that individual samples from the process may deviate significantly from the global average value. Nevertheless, if those individual samples are mutually independent (or they exhibit only short-range correlations), then the aggregated process quickly converges to a smooth function that is concentrated around the global average. Conversely, even a process that exhibits low variance could be hiding a significant LRD component, in which case it may continue to exhibit similar variance despite repeated aggregation.

Unfortunately, despite its widespread presence, the evaluation of LRD poses significant difficulties especially for practitioners. We can see several barriers to limit the use of these concepts in the networking community.

- **Complexity:** Many of the concepts are fairly hard to comprehend both from an intuitive and a mathematical perspective. There have been limited efforts to systematize and simplify these concepts, which results in confusion, partial understanding and misinterpretation of terms.
- **Confusion:** There does not exist a straightforward step-by-step approach to quantify LRD.

  A measure of the strength of LRD is the **Hurst exponent (H)**, a scalar. However, the Hurst exponent can only be estimated and not calculated in a definitive way. The several different *estimators* for the Hurst exponent often produce conflicting results [14] [18].
- **Lack of support:** There does not exist a single source of information or tools. As a result, compiling and digesting the LRD literature and developing tools from scratch is a non-trivial effort.

The overarching goal of this work is to demystify LRD and make it accessible to non-experts. Therefore, we have developed *SELFIS, a SELF-similarity analysIS tool*, a first step towards supporting a community-wide reference implementation of the major algorithms used for self-similarity analysis. For this reason, SELFIS is: a) free, b)user-friendly, and c) extensible. The need for such a tool has been demonstrated by the 200 researchers[1] who downloaded the tool. It is implemented in java to avoid the need of costly commercial software. Furthermore, its modular open-source design allows for a collaborative development that can integrate the community expertise. More specifically, the design goals for SELFIS include:

- **Ease of use and accessibility:** Through a straightforward graphical user interface, it enables non-experts to use self-similarity by making the tool valuable both for research and educational purposes. The simplicity of the interface allows for effortless use of the capabilities of SELFIS, while the visualization of the output of LRD test algorithms offers a quick sanity check and educational aid. However, simplicity does not depreciate the power of the analysis. Together with our analysis of LRD algorithms (e.g., section 4, [14] [13]), SELFIS is a powerful yet simple and intellectually accessible tool.
- **Repeatability and Consistency:** Results and observations from different research efforts can be replicated and validated. SELFIS offers a common reference platform for self-similarity analysis so that researchers will not be required to implement sophisticated algorithms from scratch.
- **Evaluation:** Observations can be made about the performance, capabilities and limitations of various long-range dependence estimators. In addition, future versions of SELFIS will incorporate algorithms and heuristics to overcome the statistical limitations to allow for robust estimation.

In addition to implementing well-known long-range dependence estimation algorithms, SELFIS also incorporates an intuitive approach to verify the existence of long-range dependence. We call this methodology **randomized buckets**. The method allows us to independently control the amount of correlations at different scales in a given dataset. In randomized buckets, the initial time-series is

[1] The researchers span multiple disciplines such as computer science, electrical engineering, mathematics, psychology. Apart from its academic appeal, it has attracted the interest of various research labs around the world such as Telstra Research Laboratories, the Chinese Academy of Science, Ericsson Research, USC/ISI, and Swiss Federal Institute of Technology.

permutated in a controlled fashion, and the statistical properties of the initial and the randomized series are compared. The approach extends and subsumes the methodology used in [7] in 1996.

We extend the initial method in two ways. First, we enable multiple levels of permutations in order to separate user-defined "medium" correlations from long and short. Second, we enable different ways of randomizing the time-series apart from the permutations: we allow sampling with repeatable values.

Finally, to demonstrate the value of SELFIS, we use it to evaluate various long-range dependence estimators in a variety of test cases. Our results show that the estimators have significant limitations and interpretation of results is not trivial. First, several estimators seem to be sensitive to short-range dependence. Using the randomized buckets methodology, we create series without long-range dependencies; however, some of the estimators continue to report the same Hurst exponent values indicating long-range dependence. Second, even in synthesized long-range dependent series, many of the estimators fail significantly to report the correct Hurst exponent value.

The rest of this paper is structured 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 analysis tool. Section 4 is a case study that presents the functionality of SELFIS: randomized buckets and a study of the reliability of long-range dependence estimators. Section 5 concludes our work.

## 2. DEFINITIONS - BACKGROUND

Self-similarity describes the phenomenon where certain properties are preserved irrespective of scaling in space or time. Of interest to network traffic processes is second-order self-similarity. Second-order self-similarity describes the property that the correlation structure of a time-series is preserved irrespective of time aggregation. This correlation is captured by the autocorrelation function (ACF), which measures the similarity between a series $X_t$, and a shifted version of itself, $X_{t+k}$. Simply put, the autocorrelation function of a second-order self-similar time-series is the same across multiple aggregation levels. For detailed description of self-similarity, see [20].

If the ACF decays hyperbolically to zero then the process shows long-range dependence. Long-range dependence measures the memory of a process. Intuitively, distant events in time are correlated. On the contrary, short-range dependence is characterized by quickly decaying correlations (e.g. ARMA processes). The strength of the long-range dependence is quantified by the Hurst exponent (H). A series exhibits LRD when $\frac{1}{2} < H < 1$. Furthermore, the closer $H$ is to 1, the stronger the dependence of the process is.

More rigorously, a stationary process $X_t$ has long-memory or is long-range dependent [5], 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 $\rho(k)$ is the **sample Autocorrelation function (ACF)**:

$$\rho(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}$$

where $\mu, \sigma$ are the sample mean and standard deviation respectively. The definition states that the autocorrelation function of a stationary long-range dependent process, decays to zero with rate approximately $k^{-\alpha}$, where $H = 1 - \frac{\alpha}{2}$ is the Hurst exponent. For traffic modeling purposes stationarity implies that the structure of a time-series does not depend on time.

Another way to characterize long-range dependence is to study the properties of **the aggregated process** $X^{(m)}(k)$ which is defined as follows:

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

To evaluate long-range dependence, the effect of aggregation on various second-order statistics is evaluated. For example, if there is no correlation in the time-series then the variance of the aggregated series should decrease as $\frac{1}{m}$ [5]. Slower decaying variance will imply long-range dependence.

Research dealing with self-similarity and long-range dependence can be classified in two general categories. The first includes studies on the manifestation of such phenomena in networking, their origins and their effects. The second category involves reports on modeling and estimating long-range dependence, as well as showing the complexity and difficulty in its correct use and interpretation.

There has been ample evidence of long-range dependence and scaling phenomena in many different aspects of networking. The first experimental evidence of self-similar characteristics in local area network traffic were presented in [16]. The authors perform a rigorous statistical analysis of Ethernet traffic measurements and were able to establish its self-similar nature. Similar observations were presented for wide area Internet traffic in [22] and World Wide Web traffic in [6] where the underlying distributions of file sizes were shown to be the main cause of self-similarity. In [32], the authors discuss the failure of Poisson modeling in the Internet. Scaling phenomena [10] [29] and the factors that contribute to self-similarity and long-range dependence have been extensively studied [19] [33] [31] [9] [8]. Furthermore, in [11] [24] [12] the relevance and the effects of the self-similarity on various metrics of network performance are examined.

The second major aspect of research dealing with self-similarity and long-range dependence is estimation of the Hurst exponent. An overview of a large number of these estimation methodologies can be found in [27] [5] [28]. Relatively little effort has been devoted to studying the accuracy of the estimation methodologies [27] and pointing out difficulties in long-range dependence estimation [18] [14]. The authors present pitfalls when estimating the intensity of long-range dependence in the presence of trends, non-stationarity, periodicity and noise. Furthermore, the limitations of the variance-time estimator have been analyzed in [15].

Of major importance is also the development of models for simulating long-range dependence. Proposed models like the one in [23] or generators for long-range dependent time-series [21] are hard to evaluate in practice. Thus, there are hardly any studies that assess the various models or compare the different generators. In general, this suggests the need for practical tools and a systematic methodology to estimate, validate and generate long-range dependent time-series.

## 3. THE SELFIS TOOL

The SELFIS tool [1] (fig. 1) is developed to provide all the necessary functionality for a complete and systematic analysis. Our goal is to establish SELFIS as a reference point in self-similarity 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 currently incorporates all the widely used long-range dependence estimators. 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 different classes of functionality included in SELFIS: a) Hurst exponent estimators, b) randomized buckets, c) transforms, d) data processing and e) fractional Gaussian noise generators.

### 3.1 Hurst Estimators

SELFIS includes most of the existing long-range dependence estimators. These estimators can be classified in two main general categories. In the first, there is a number of time-domain methods, such as RSplot and the Variance method. The second category includes the frequency-based estimators, such as the periodogram, the Whittle and the Abry-Veitch estimators.

The existence of numerous estimators is justified by the asymptotic nature of the Hurst exponent. Intuitively, since the limiting behavior of the process can only be estimated, statistical errors and uncertainty impede reliable and concise calculation of the Hurst exponent. Statistical limitations arise also when applying mathematical definitions in practice, i.e., the estimators assume stationarity which is an elusive concept. Furthermore, each estimator looks at a different property of a given time-series. Thus, it is common that these methodologies produce conflicting estimates for the same time-series. This is true not only for “real-life” time-series where the existence of periodicities, noise or trends has substantial effect on the estimation [14], but also for synthesized LRD series with specific predetermined Hurst exponent value (see section 4.2 for the limitations of the estimators). On the other hand, the estimation methodologies examine specific properties (e.g., variance, power spectrum) at different time scales. At larger time-scales where the behavior at the limit is described, the num-

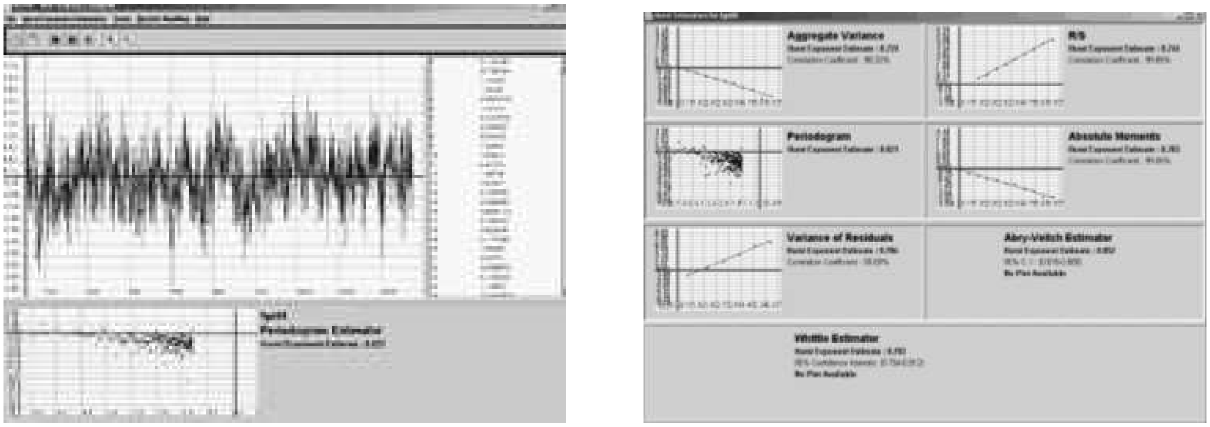

**Figure 1: Two screen dumps of the SELFIS tool.**

ber of samples decreases significantly resulting in statistical uncertainties. Applying all the estimators to a time-series provides with a more complete overall picture of its possible self-similar nature.

For all the aforementioned limitations of the estimation methodologies, SELFIS also reports the statistical significance of each estimation. The correlation coefficient or the confidence intervals where available should always be reported together with the Hurst value. Stressing only the Hurst exponent value is rather meaningless if statistically the value is not significant. More specifically, in our tool the following estimators are included:

**A.Time-domain estimators:** These estimation methodologies are based on investigating the power-law relationship between a specific statistic of the time-series and the aggregation block size $m$.

- *Absolute Value method.* The log-log plot of the aggregation level versus the absolute first moment of the aggregated series $X^{(m)}$ is a straight line with slope of $H-1$, if the time-series is long-range dependent (where $H$ is the Hurst exponent).
- *Variance method.* The method plots in log-log scale the sample variance versus the block size of each aggregation level. If the series is long-range dependent then the plot is a line with slope $\beta$ 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.
- *Variance of Residuals.* The method uses the least-squares method to fit a line to the partial sum of each block $m$. A log-log plot of the aggregation level versus the average of the variance of the residuals after the fitting for each level should be a straight line with slope of $H/2$.

**B.Frequency-domain/wavelet-domain estimators:** These estimators operate in the frequency or the wavelet domain.

- *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

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

  where $\nu$ 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 a confidence interval. It does not produce a graphical output.

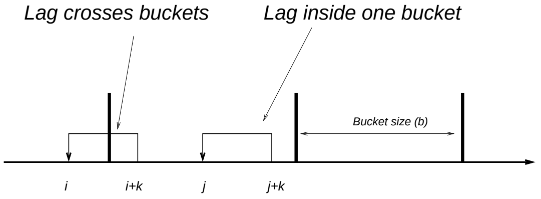

**Figure 2:** **Pairs separated by lag $k$ can belong to the same bucket or not in which case they are inbucket or outbucket respectively.**

- *Abry-Veitch (AV).* The Hurst exponent is estimated by using the wavelet transform of the series [2]. A least-squares fit on the average of the squares of the wavelet coefficients at different scales is an estimate of the Hurst exponent. The method produces both a graphical output and a confidence interval.

### 3.2 Randomized Buckets

In SELFIS, *Randomized Buckets* is used as an intuitive method for the detection and validation of long-range dependence. We examine numerous ways of randomization, such as moving numbers a constant number of positions in the series or randomizing inside the buckets with replacement. These methodologies will be further commented upon in our future work.

The idea behind randomized buckets is to decouple the short-range from long-range correlations in a series to facilitate the study of the effects of long-range dependence. This is achieved through partitioning the time series into a set of "buckets" of length $b$. Thus, we define the contents of the $u$th bucket to be items $X_{u \cdot b}, \ldots, X_{(u+1) \cdot b-1}$ from the series, and the **home** of item $X_i$ to be bucket $H(i) \equiv \lfloor i/b \rfloor$. Also, we say that two items $(X_i, X_j)$ form an **inbucket** pair if $H(i) = H(j)$; otherwise, they form an **outbucket** pair with an **offset** of $|H(i) - H(j)|$ buckets. Note that this classification depends on the (fixed) locations of the bucket boundaries, and not just the separation between two items in the time series. For example, fig. 2, shows that two items separated by lag $k$ could form either an inbucket or outbucket pair.

Once the series has been partitioned in this way, we can then apply one of the following randomization algorithms to reorder its items:

**External Randomization (EX):** The order of buckets is randomized, whereas the content of each bucket remains intact. This can be achieved by labelling each bucket with a bucket-id between 0 and $\lfloor$Time-SeriesLength$/b\rfloor$, and randomization of the bucket-ids. External randomization preserves all correlations among the inbucket pairs, while equalizing all correlations among the outbucket pairs with different offsets. Thus, if the series is sufficiently long, the ACF should not exhibit significant correlations beyond the bucket size.

**Internal Randomization (IN):** The order of the buckets remains unchanged while the contents of each bucket are randomized. As a result, correlations among the inbucket pairs are equalized, while correlations among the outbucket pairs are preserved, but rounded to a common value for each offset. Thus, if the original signal has long-memory, then the ACF of the internally-randomized series will still show power-law behavior.

**Two-Level Randomization (2L):** Each bucket is further subdivided into a series of "atoms" of size $a$. Thereafter, we apply external randomization to the block of $\lfloor b/a \rfloor$ atoms within each bucket. As a result, both short-range correlations (within each atom) and long-range correlations (across multiple buckets) are preserved, while medium-range correlations (across multiple atoms within the same bucket) are equalized.

### 3.3 Transforms

Transformations are usually applied to reveal information that is not available in the raw time-series. Fourier and wavelet transforms can be useful to reveal periodicities in the series and in general study the frequency components of the time-series. SELFIS includes the following transforms:

- Fourier Transform. Fourier transform is used to transform a series from the time domain to the frequency domain. Intuitively, the signal is transformed into a sum of sinusoids of different frequencies.
- Wavelets (Haar and D4). Wavelet transform is capable of providing the time and frequency information of a time-series simultaneously. 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 applied by median, average or exponential smoothing algorithms. Our tool includes the 4253H smoothing algorithm described in [30]. 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 are a number of tests that check a series for stationarity. One of the common tests for stationarity is the run test [4]. 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, i.e., consecutive equal values in a series. 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. Thus, a run is defined as a sequence of consecutive values that are all above or below the median of the original time-series. Nonstationarity is indicated by a number of runs considerably different than $\frac{N}{2}+1$. Stationarity is important when long-range dependence is studied, since estimators fail in non-stationary data[2].

### 3.5 Fractional Gaussian Noise Generators

Fractional Gaussian noise (fGn) generators can synthesize series with long-range dependence. Our tool includes two generators. The first method is based on fast Fourier transform to generate a fGn series [21]. The second generator produces fGn series by using the Durbin-Levinson coefficients.

## 4. CASE STUDY

This section highlights the capabilities of SELFIS. Two case studies are presented. First, a demonstration of how randomized buckets can be used to stress-test long-range dependence and cancel the effect of short-term correlations. To demonstrate the methodology, we use fractional Gaussian noise series generated by one of the generators included in SELFIS. Second, we demonstrate that long-range dependence estimators have limited capabilities.

These case studies also demonstrate and justify the need for different estimation methodologies. Each estimator has different strengths and weaknesses and thus can be best used at different cases. In addition, understanding the limitations of each estimator allows for sound usage of the SELFIS tool and interpretation of its results.

### 4.1 Randomized Buckets

Randomized buckets (see previous section) is an intuitive, straightforward methodology that validates the existence of long-memory. To show how long-range dependence can be detected using randomized buckets, we synthesized a sample series of fractional Gaussian noise. The series (fig. 3, left plot) has length 65536, Hurst exponent 0.8 and was synthesized using the generator created by Paxson [21]. The middle plot in fig. 3 shows the sample autocorrelation function (ACF) of the series which decays hyperbolically to zero and implies long-range dependence. To ensure that long-range dependence really exists we employ randomized buckets. The right plot in fig. 3 shows the ACF of the fGn series if randomized externally with bucket size 1. This type of randomization removes all correlations by creating a completely random series As expected, the ACF shows that no correlation exists at all time lags. Fig. 4 shows the ACF function after the signal is randomized with three different ways:

- **External Randomization** (using $b = 50$) causes the ACF to drop smoothly from the initial value for the unrandomized sequence to zero as the lag increases, reaching zero at exactly the bucket size! The left plot in fig. 4 shows that all correlations are equalized beyond the bucket size (50).

[2] If stationarity is detected, the time series must be differenced successively until stationarity is achieved.

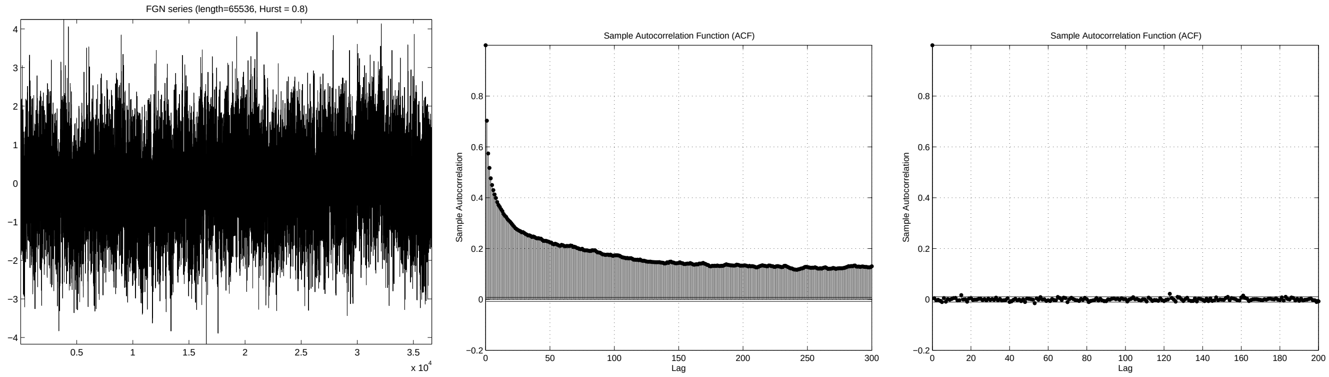

**Figure 3:** **LEFT: fGn series of length 65536 and Hurst 0.8. MIDDLE: Autocorrelation function (ACF) of the series up to lag 300. The ACF shows power-law like behavior. RIGHT: ACF after external randomization with bucket size 1 (full randomization).**

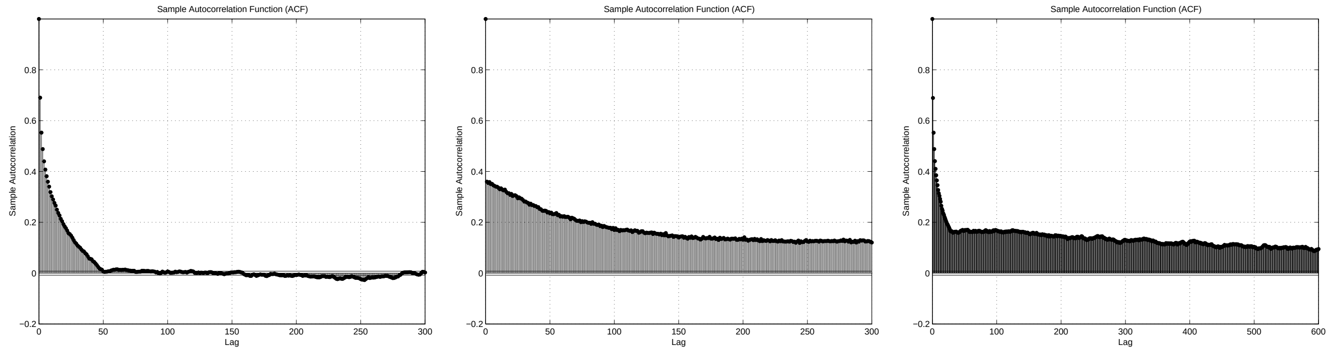

**Figure 4:** **LEFT: External randomization with bucket size 50. After lag 50 all correlations are insignificant. MIDDLE: Internal randomization with bucket size 50. The ACF shows the same power-law behavior like the original series (Fig. 3). RIGHT: Two-level randomization with bucket sizes 300 and 30. Medium-range correlations are distorted.**

- **Internal Randomization** (using $b = 50$) significantly lowers and flattens the ACF at small values of the lag compared to the original (unrandomized) series. However, for large values of the lag, internal randomization has no effect on the ACF. The ACF (fig. 4, middle plot) for large lags (beyond the bucket size) is similar to the ACF of the original series (bucket size 50). Observe that long-range dependence seems to dominate the original series, since the effect of equalizing the inbucket correlations on ACF is minimal.

- **Two-Level Randomization** (using $a = 30$, $b = 300$) exhibits similar behavior to external randomization for small values of the lag (i.e., less than $a$), along with similar behavior to internal randomization for large values of the lag. These two limiting values also match the ACF for the original (unrandomized) series, but for intermediate values the two-level randomization significantly reduces the correlations. The right plot in fig. 4 demonstrates the distortion of medium-range correlations in the ACF after two-level randomization.

To emphasize the effect of the various types of randomization on the correlations of the time-series, we plot in log-log scale the autocorrelation function after various types of randomization (fig. 5). The ACF of the initial (unrandomized) fGn series is a straight line as expected from the definition of long-range dependence (see section 2). The ACF after internal randomization differs from the original ACF for lags smaller than the lag representing the bucket size. On the contrary, ACF after external randomization is similar to the original for small lags only, while the ACF after two-level randomization differs for intermediate lags. Furthermore, in fig. 6 we plot the difference between ACF after the various types of randomization and the initial ACF. The difference is close to zero for

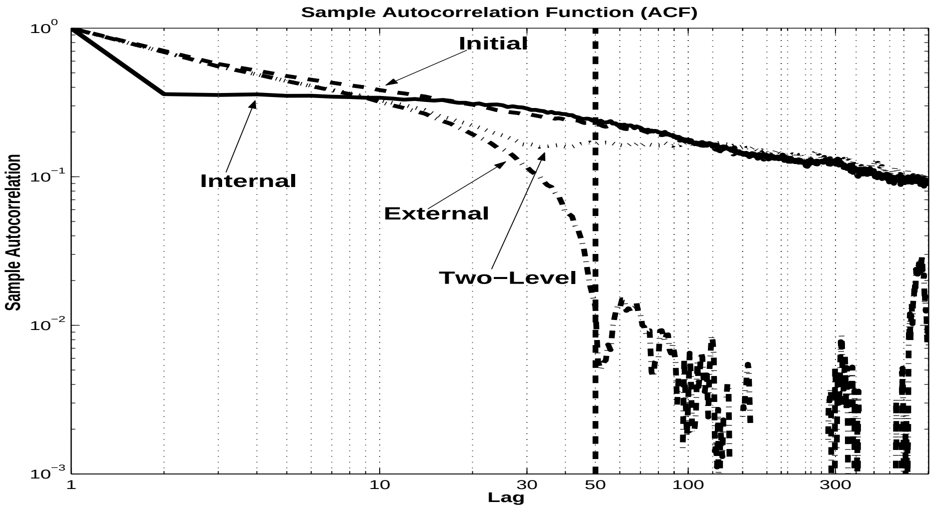

**Figure 5:** **The ACF for various types of randomization. The vertical dashed line shows the bucket size used.**

lags larger than the bucket size for internal and two-level randomization indicating long-range dependence. On the contrary, the difference in the case of external randomization decays with the lag up to the bucket size; beyond that, it starts to grow towards zero. At small lags the ACF after external randomization drops faster to zero than the initial ACF. Thus, the difference between them grows as the lag approaches the bucket size. At the bucket size the difference becomes maximum. After the bucket size the difference becomes smaller as the initial ACF also approaches zero.

## 4.2 Reliability of LRD Estimators

This section is an evaluation of the Hurst exponent estimators. We examine the accuracy and robustness of the estimators using two types of test-cases: a) Synthesized LRD series with known Hurst exponent value to study the accuracy of the estimators. b) Randomized LRD series using randomized buckets to study the effect of short-range correlations on the estimators.

**A. Accuracy on Synthesized LRD Series:** We show that the estimators seldom agree on the value of the Hurst exponent, and often they disagree by significant difference. Each of the estimators was tested against two different types of synthesized long-memory series: a) Autoregressive Fractional Integrated Moving Average processes (ARFIMA) and b) fractional Gaussian noise (fGn) series. For more details, also see [14].

We generate 100 datasets with different seed for each Hurst value from 0.5 to 0.9 with step of 0.1. Fig.7 summarizes our findings for the Paxson generator and the ARFIMA model. For both plots in fig.7, the X axis presents the Hurst exponent value of the fGn series and the Y axis shows the average estimated value of the corresponding methodology. The “Target” line presents what the optimal estimation of the fGn data for each case would be. The 95% confidence intervals are typically within 0.01 of the reported value.

Our findings exhibit the inability of the majority of the estimators to accurately estimate the value of the Hurst exponent. For fGn data, with the exception of the Whittle and Periodogram estimators, all other estimations fail to estimate correctly. We observed similar results in the case of series generated with the ARFIMA model. In the latter case, the Periodogram, Abry-Veitch and R/S estimators produce values closer to the target.

**B. Estimators and Randomized Buckets:** How sensitive are the estimators to short-range correlations? To address this question we employed the randomized buckets methodology. Intuitively, the estimations should not be affected after internal randomization. Note that internal randomization breaks the short-term correlations, while preserving the long-term. On the contrary, external randomization should significantly influence estimations, since long-memory is distorted. Estimates of the Hurst exponent of externally randomized series should be close to 0.5 since long-range dependence has been canceled.

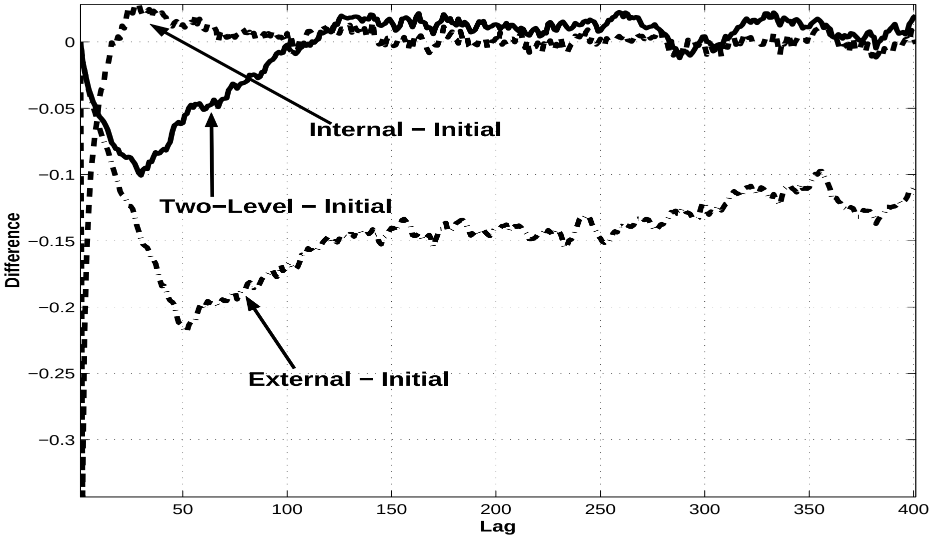

**Figure 6:** **Differences in autocorrelation coefficients of the ACF after randomization. ACF after internal, external and two-level randomization minus the initial ACF.**

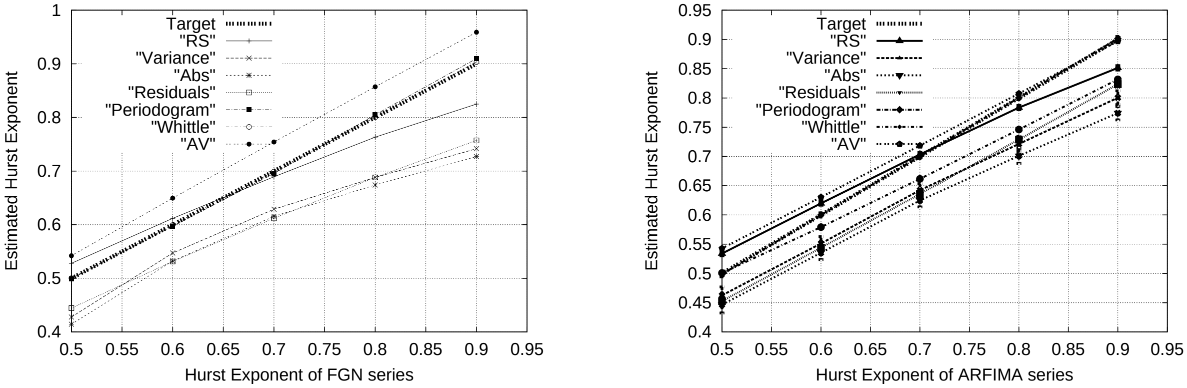

**Figure 7:** **The performance of the estimators using Paxson's generator (left) and ARFIMA (right). The "Target" line shows an optimal estimation of the synthesized data. The Whittle and Periodogram estimators fall exactly on the Target line in the fGn case.**

We synthesized numerous fGn series for various Hurst values between 0.5 and 1. These synthesized series were randomized for different bucket sizes. Fig.8 presents our findings. For both plots, the Y axis shows the estimated Hurst value while the X axis presents the bucket size. The ACF of the initial (unrandomized) time-series is represented in the plots with bucket sizes infinity and one for external and internal randomization respectively. In more detail, this figure shows the average estimations of 100 fGn series with Hurst 0.8 after being externally and internally randomized with bucket size ranging from 10 to 90.

Intuitively, we would expect that especially for small bucket sizes all estimations would be close to 0.5 after external randomization (left plot of fig. 8). This is true for all the estimators except the AV, Whittle and Periodogram estimators who behave counter-intuitively. In particular, AV and Whittle estimators do not seem to be affected by external randomization irrespective of the bucket size. The three "frequency-based" estimators, especially Whittle and AV produce the same estimates as before randomizing (bucket size of infinity in the figure), even though long-range dependence has been eliminated from the series.

Similar counter-intuitive behavior for AV and Whittle holds in the case of internal randomization. Except AV and Whittle, all estimators estimate the same Hurst value before and after internal random-

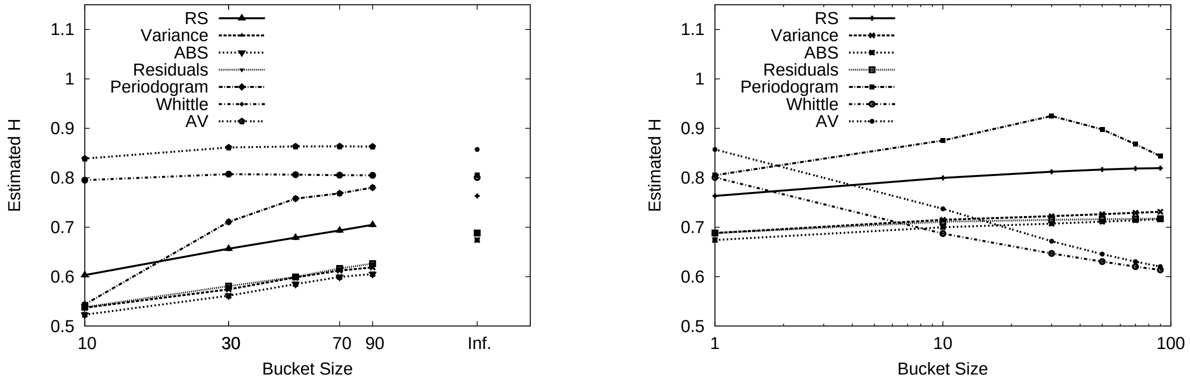

**Figure 8: External (left) and internal (right) randomization with various bucket sizes (10 30 50 70 90). The X axis is presented in log scale. "Frequency based" (AV, Whittle, Periodogram) estimators fail to capture that long-term correlations are distorted in external randomization. Their estimations are the same as before randomizing (the bucket size of the initial ACF can be considered to be infinity in external randomization). In addition, these estimators fail to capture internal randomization. Their estimations after internal randomization deviate significant from the initial estimations (the bucket size of the initial ACF can be considered to be 1 in internal randomization).**

ization as intuitively expected (fig. 8 right plot). On the contrary, estimations of the Hurst value from AV and Whittle estimators drop significantly as the bucket size increases. However, estimations should be unaffected since internal randomization only destroys short-range correlations.

Considering the effect of randomized buckets on long-memory, we can claim that these two estimators —AV and Whittle— seem to depend more on the short-term behavior of the time-series to derive an estimate for the Hurst exponent.

Note, that the estimators that perform the best in synthesized long-range dependence series were the ones most affected by short-term correlations.

Summing up our study of the accuracy and robustness of the Hurst exponent estimators, we reach the following main conclusions:

- When the data are generated by fractional Gaussian noise (fGn), Whittle, Periodogram seem to give the most accurate estimation for the Hurst exponent.
- When the data are synthesized using the ARFIMA model, AV, Periodogram and R/S have the best performance.
- Even though the Whittle estimator is considered the most robust, it is the most sensitive of the estimators.
- AV and Whittle estimators seem to depend mainly on short-range correlation to derive the Hurst exponent estimate.

In general, there is no definite estimator that could be consistently used in every case. Each estimator evaluates different statistics of the time-series to estimate the Hurst exponent. Thus, different processes may have different effect on each estimator.

## 5. CONCLUSIONS

The main contribution of this work is the development of the SELFIS software tool. We believe that SELFIS presents a long-overdue first step towards a widely-used reference platform to facilitate self-similarity analysis. Through the introduction of this tool, we hope to encourage greater use of these techniques. Despite the wide interest in self-similarity and long-range dependence within the community, a common tool not yet emerged. As a result, this impeded the use and comparability of results.

In more detail, SELFIS provides the following direct benefits.

- *Accessibility:* Anyone will be able to use long-range dependence analysis, even non-experts. In this sense, SELFIS will help spreading the use of LRD concept for research and educational purposes.
- *User friendliness:* The interface of the tool is

straightforward making its use effortless, while visualization offers a fast sanity check of resutls.

- *Robustness:* It offers multiple estimators for reliable results and presents their statistical significance.
- *Repeatability:* Results can be replicated and verified.
- *Open source collaborative development:* The community can cooperate on enriching the capabilities of the tool therefore leveraging from each others work. Different groups are welcome to contribute their expertise. We actively solicit contributions.
- *Free:* The use of SELIFS comes with no monetary cost. SELFIS is publicly available and no extra commercial software is needed.

An additional contribution is an implementation of a powerful tool for stress-testing long-range dependence. Randomized buckets can isolate the effect of short, long and medium correlations on a time-series. Although this idea appeared in earlier work it has been neglected. We would like to revive and develop this idea to its full potential.

Our study on the estimation of long-range dependence and our experience with the estimators allow us to highlight a few tips for practitioners. First, 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. In addition, researchers should not rely only on one estimator in deciding the existence of long-range dependence. Several of the estimators can be overly optimistic in identifying long-range dependence. Furthermore, for efficient characterization, it may be necessary to process and decompose the time-series. Finally, a visual inspection of the time-series can be very useful, providing a qualitative analysis and revealing many of its features, like periodicity.

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.

Finally, the algorithms within SELFIS are not restricted to the domain of time-series networking data. Thus, we hope that our work may be applied to the analysis of long-range dependence data sets in other disciplines, such as computer science, economics, sociology, psychology, etc.

## 6. REFERENCES

[1] The SELFIS Tool. http://www.cs.ucr.edu/∼tkarag.

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

[3] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Long-Range Correlations Between DNA Bending Sites. In *Journal of Molecular Biology*, volume 4, pages 903–918, 2002.

[4] J. S. Bendat and A. G. Persol. *Random Data - Analysis and Measurement Procedures.* John Wiley & Sons, NY , 1986.

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

[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] 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, 1996.

[8] A. Feldmann, A.C.Gilbert, and W.Willinger. Data network as cascades: Investigating the multifractal nature of the Internet WAN Traffic. In *Computer Communications Review*, 1998.

[9] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger. Dynamics of IP Traffic: A Study of the Role of Variability and The Impact of Control. In *SIGCOMM*, pages 301–313, 1999.

[10] 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.

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

[12] G. K. K. Park and M. E. Crovella. On the Effect of Traffic Self-Similarity on Network Performance. In *Proceedings of SPIE International Conference on Performance and Control of Network Systems*, 1997.

[13] T. Karagiannis and M. Faloutsos. SELFIS: A Tool For Self-Similarity and Long-Range Dependence Analysis. In *1st Workshop on Fractals and Self-Similarity in Data Mining: Issues and Approaches (in KDD)* , Edmonton, Canada, July 23, 2002.

[14] T. Karagiannis, M. Faloutsos, and R. Riedi. Long-Range dependence:Now you see it, now you don't! In *IEEE GLOBECOM, Global Internet Symposium*, 2002.

[15] M. Krunz. On the limitations of the variance-time test for inference of long-range dependence. In *IEEE INFOCOM*, pages 1254–1260, 2001.

[16] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar nature of Ethernet Traffic. In *IEEE/ACM Transactions on Networking*, 1994.

[17] X. Lu, Z. Sun, H. Chen, and Y. Li. Characterizing Self-Similarity in Bacteria DNA Sequences. In *Physical Review E*, volume 58, pages 3578–3584, 1998.

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

[19] K. Park, G. Kim, and M.E.Crovella. On the Relationship Between File Sizes Transport Protocols, and Self-Similar Network Traffic. In *International Conference on Network Protocols*, pages 171–180, Oct 1996.

[20] K. Park and W. Willinger. Self-similar network traffic: An overview. In *Self-Similar Network Traffic and Performance Evaluation.* Wiley-Interscience, 2000.

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

[22] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 1995.

[23] 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.

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

[25] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* W. H. Freeman & Co., 1992.

[26] R. V. Sole, S. C. Manrubia, M. Benton, and P. Bak. Self-Similarity of Extinction Statistics in the Fossil Record. In *Nature*, volume 388, pages 764–767. Macmillan Publishers Ltd, 1997.

[27] 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.

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

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

[30] Velleman, P. F., and D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, MA, 1981.

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

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

[33] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, 1997.