3D-LD: a Graphical Wavelet-based Method for Analyzing Scaling Processes

Steve Uhlig[†] Olivier Bonaventure[†*] Chris Rapier[‡]

[†] Infonet group, University of Namur, Belgium.
[‡] Pittsburgh Supercomputing Center, Pittsburgh, USA.

Abstract

This paper proposes a novel graphical method (3D-LD) based on wavelets for studying scaling and non-stationary signals. We show that it allows to better study the timescale-dependent qualitative properties of such signals, compared to the classical techniques. By using the 3D-LD, we study two recent network traffic traces to understand the actual nature of scaling in TCP flow arrivals. We show that the TCP flow arrivals is a complex scaling process, which is non-stationary with respect to its degree of statistical dependence as well as over timescales longer than hours. We show that the application mix in the traffic has a significant impact scaling over timescales between seconds and minutes. This scaling for timescales between seconds and minutes is created by statistical dependence within user sessions. Self-similarity (or long-range dependence) that appears over timescales larger than several minutes on the other hand seem to be an invariant of the flow arrivals and is, in all likelihood, created by the user sessions arrivals. Based on this analysis, we propose a simple model for the TCP flow arrivals process, taking into account the timescales ranging from seconds to hours, and we show that simulating realistic TCP flow arrivals conforming to this model is feasible.

Keywords: scaling processes, wavelet analysis, network traffic modeling, TCP flow arrivals.

1 Introduction

It has been more than a decade since self-similarity was discovered in network traffic [1] [2]. Since then many studies have looked at network traffic variability and shown the ubiquity of its wild behavior; see [3] for a detailed and up to date review of self-similarity in network traffic. While there can be no doubt about the existence of statistical self-similarity in network traffic at many timescales, few studies have tried to disentangle the various causes of this self-similarity, although [4] is probably the most complete study of the qualitative causes of scaling in network traffic.

*Corresponding author. E-mail: bonaventure@acm.org, URL: http://www.infonet.fundp.ac.be.

Among all components of network traffic, the TCP flow volumes have been widely studied in [5], [6], [7], [8] and [9] and in many cases a connection has been made between "long-tails" and self-similarity. The question of whether flow sizes are heavy-tailed in the strict sense or over a given range of scales, as discussed in [9], is not entirely important as long as flow sizes are shown to be partly responsible for the self-similarity. On the other hand, to understand the traffic dynamics as a whole, a proper understanding of the flows carrying the traffic is necessary. For that purpose, we rely on two recent traces to study the "invariant" nature of scaling in TCP flow arrivals. Other traces are analyzed in a longer technical report [10].

In analyzing scaling processes several techniques have been proposed in the literature most of which rely on strong assumptions regarding the nature of the process. These "classical techniques" as we call them [11] require that the studied signal exhibits either statistical dependence or self-similarity (or long-range dependence), but not both. This means that hybrid signals, such as network traffic, that contain components of different nature, cannot be correctly analyzed with these tools. This is the main reason why wavelets have become so widely used, because they allow a finer analysis of the time-scale dependent properties of the signal.

The primary motivation for writing this paper is to understand the complex nature of network traffic, where different behaviors have been identified at different timescales. However, the problem with such a signal is that it is not possible to obtain an accurate picture of its components by relying on a single analysis technique. The need to rely on wavelets for analyzing network traffic was recognized several years ago. The work carried out by many people (see [12]) has allowed wavelet-based tools to disentangle several issues in our understanding of the network traffic dynamics. [13] provides an excellent review of all the wavelet-based techniques applicable for analyzing network traffic data. This paper is another step forward in the use of wavelets for analyzing complex-behaving signals with an emphasis on network traffic and particularly the process of the arrivals of the flows.

The remainder of this paper is structured as follows: Section 2.1 provides the basic definitions used in this

paper. Section 2.2 presents and discusses the classical wavelet-based method for analyzing scaling processes. Section 3 presents the 3D-LD, the graphical wavelet-based method we propose. Section 4 describes the studied traffic traces. Section 5 discusses the related work in TCP flow arrivals characterization. Section 6 studies the scaling of the TCP flow arrivals process in these traces. Finally, in section 7 we propose a model for generating a synthetic TCP flow arrivals process.

2 Definitions

2.1 Scaling, self-similarity, and LRD

Scaling behavior is a broad term connected with the absence of a particular characteristic scale controlling the process[1] under study. Among scaling processes, two are of interest in the context of this paper, namely self-similar processes and long-range dependent (LRD) processes.

Self-similarity, or more precisely self-affinity [14], with parameter H ($0 < H < 1$), is defined by

$$\{X(t), t \in \mathbb{R}\} \stackrel{d}{=} \{c^H X(t/c), t \in \mathbb{R}\}, \forall c > 0, \quad (1)$$

where $\stackrel{d}{=}$ denotes equivalence in finite distributions. A property of self-similar processes concerns the scaling of their moments:

$$E|X(t)|^q = E|X(1)|^q \, |t|^{qH} . \quad (2)$$

This latter property relates the behavior of moment q (if it exists) as a power law of time. A self-similar process will exhibit the same fluctuations independent of the considered scale, hence cannot be stationary due to this dependency of all moments on time. Recall that a process is said to be stationary if its statistical properties do not depend on time.

On the other hand, long-range dependence (LRD), or long-memory, is associated with stationary processes. A second-order stationary process displays LRD if its autocovariance $r(k)$ behaves like

$$r(k) \sim c_r |k|^{2H-2} \quad as \quad |k| \to 0 \quad (3)$$

when $1/2 < H < 1$. In such a case, the correlations decay so slowly that $\sum_{k=-\infty}^{\infty} r(k) = \infty$. When $H = 1/2$, the process is uncorrelated ($r(k) = 0$). On the other hand, when $0 < H < 1/2$, we speak of short-range dependence, or anti-persistence, due to negative dependence ($r(k) < 0$ for $k \neq 0$) between time lags. The divergence of the correlations when $1/2 < H < 1$ also means that the spectral density behaves like

$$f(\lambda) \sim c_f |\lambda|^{1-2H} \quad as \quad |\lambda| \to 0, \quad (4)$$

hence diverges at the origin ($f(0) = \infty$). But because of stationarity, LRD processes cannot exhibit fluctuations at any arbitrary scale, as do self-similar processes.

An often mistaken relationship between LRD and self-similarity concerns the fact that strict self-similarity for $H > 0.5$ implies LRD. This means that there are in fact two types of LRD behavior: one that is limited to the strict definition in 3 and 4 for a second-order stationary process for which strict self-similarity does not hold, and the simple consequence of LRD that arises for a strictly self-similar process. This distinction is not simply formal but has important consequences in practice on the type of process considered. For example, it is not always possible to distinguish in practice between pure statistical correlations that span more than the length of the studied sample, pure LRD and LRD produced by a self-similar process. Only the knowledge of the physics of the process can make the difference between these processes, or to rely on a longer time-series.

2.2 Wavelet analysis

In this section we discuss the use of wavelets for analyzing scaling processes. For a brief introduction to wavelet analysis, we refer the reader to [12] and [13]. For a complete treatment of wavelet theory and analysis, see [15]. The important features of the wavelet transform for studying scaling processes are as follows:

1. scale invariance: the wavelet basis being built through the dilation operator, the family of analyzing functions has a built-in scaling property.

2. vanishing moments: the mother wavelet (ψ_0) has a number $N \geq 1$ of vanishing moments $\int t^k \psi_0(t) dt \equiv 0, k = 0, ..., N-1$, allowing the wavelet to remove any polynomial trend of order up to $N-1$.

The first advantage of the wavelet domain concerns its reduced bias against non-stationarity and polynomial trends (point (2)), these two features are capable of making a non-scaling signal look like a true scaling process, especially in the case of a highly non-stationary signal like network traffic. The other advantage of the wavelet domain is the built-in scaling (point (1)) of the wavelet basis which allows a multiresolution analysis [15]. This feature is particularly useful for signals containing components with different qualitative behavior at various timescales since multiresolution analysis will split the signal into its component timescales allowing the study of each with great independence.

The wavelet-based estimator of scaling, the *logscale diagram* (LD) [12] consists in plotting $y_j = log_2(\mu_j)$ against j with the confidence intervals, where

$$\mu_j = \frac{1}{n_j} \sum_{k=1}^{n_j} |d(j,k)|^2, \quad (5)$$

[1] We use the terms process, signal, and time-series interchangeably in this paper.

n_j is the number of wavelet coefficients at octave j and $d(j,.)$ denotes the wavelet coefficients (of the forward wavelet transform of the time-series) at octave j. Each octave reduces the data by a factor of 2, so that each subsequent octave represents a timescale two times longer. The *logscale diagram* is a second-order statistics (sample variance) of the wavelet coefficients, hence it does not capture the higher-order properties of the signal. The *logscale diagram* allows the detection of scaling through observation of alignment (linear trend) of the confidence intervals of the y_j within some scaling region (some octave range). If, and only if alignment is detected, can estimation of scaling parameters be performed. Note that we use the term "scaling" not in the strict sense but to express any loose alignment of the LD for some octave range (with strictly positive slope α). This will not pose any problem in our case since we do not estimate the value of the scaling exponent but only deal with qualitative aspects. For further details about the correct interpretation of the LD, we refer the reader to [12].

3 3D-LD

In this paper, we extend the LD introduced in the previous section to study the scaling properties of signals by computing the LD over fixed-length time intervals. We then graphically compare the evolution over time of these LDs with the LD computed directly over the whole trace. The method we propose consists of a 3-dimensional plot of the evolution over time of LDs computed over constant size time intervals, hence we call these time-varying LDs "3D-LD".

Because we compute the LD over constant-sized blocks for the 3D-LD, an important feature of the mother wavelet is its vanishing moments. This ensures that polynomial variations of the signal up to order $N-1$ (3 in our case) will not bias the LD for a given block. This feature is extremely important because a potentially large bias may come from the signal's level variations (non-stationarity) between the constant size blocks used to compute the LD.

To enhance the readability of the 3D-LDs we compute a LD for every half-block length to smooth the plots instead of computing a LD for non-over-lapping blocks of the entire traces. Again, to enhance readability, we do not show the confidence intervals on the 3D-LDs as it would require plotting three surfaces on the same graph while these confidence intervals mostly depend on the length of the data used for computing the wavelet transform. Hence these confidence intervals do not change much along the 3D-LD since we use constant-sized blocks.

Our method allows to study the scaling properties of signals, without requiring to guess at which timescales scaling occurs. Our approach differs fundamentally from [16] where the goal is to test the constancy of the scaling exponent, based on an a priori guess of the octave range where scaling is due to occur. Indeed, the 3D-LD should allow to directly visualize the octave range of interest as required in [16].

3.1 Toy example

To illustrate the usefulness of the 3D-LD, we provide a "toy" example of a signal whose scaling properties change with time, and compare the LD with the 3D-LD. The signal used in this simulation contains two components. The first of which is a homogeneous Poisson process with a constant mean and $\mu = 100$ (for example representing a flow arrivals process). We then overlay statistical dependence with a time-varying correlation length comprised between 1 and about 130 time ticks[2]. We note that the time unit in the simulation has no particular physical meaning, it is nothing more than a simulation time tick. For creating this dependence, we spread the unitary weight of any "flow" (each unit of the Poisson process) evenly over the required correlation length. This generates dependence across the time period over which we spread the flow's weight. For the first third of the simulation, the correlation length starts at about 130 time units, and it decreases linearly with time. The second third of the simulation keeps the original Poisson arrivals, without creating dependence. Finally, the third part of the simulation uses the same procedure as in the first part, but with a linearly increasing flow duration (or correlation length). This should show a rise of scaling with time therefore increasing the slope in the LD.

Figure 1 illustrates the time evolution of the simulated process, i.e. the time evolution of the number of flows at any given instant. The three periods of the simulation are easily distinguished, with the first being a scaling process with a limited but increasing variability. The second period of the simulation shows the raw Poisson process which presents a higher variability but without any dependence, hence no scaling. Finally, the third period of the simulation shows the decreasing variability of the scaling process but an increasing dependence. The "toy" example clearly exhibits the smaller variability of the correlated process in contrast to the "large" but uncorrelated

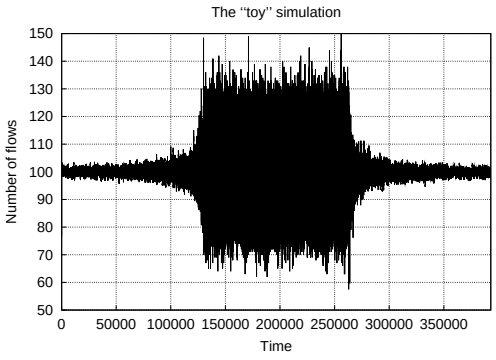

Figure 1: The "toy" simulation.

[2]This choice is purely arbitrary.

variability of the Poisson process. This shows that scaling should not be confused with high variability.

Figure 2 presents the LD for the whole simulation, illustrating the blindness of the LD for studying the time-varying nature of scaling. While the small timescales (octaves < 5) provide indication of no scaling with a flat LD,

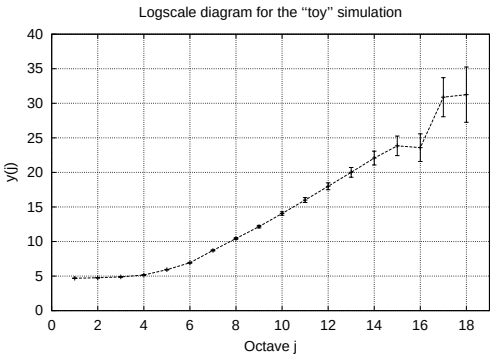

Figure 2: LD for the "toy" simulation.

we cannot identify the Poisson process that occurs for the second third of the simulation. The larger timescales (octaves ≥ 5) tell us that scaling occurs. We know however that true scaling occurs only for timescales from 1 to 7 due to the flows correlation ranging between 1 and 130 time ticks.

Figure 3 presents the 3D-LD of the simulated scaling process. We also plotted, only for illustration purposes, the confidence intervals for the 3D-LD, to convince the reader about their strong dependency on the block size and their small dependency on the particular block. The 3D-LD was computed using time blocks of 2^{15} ticks, hence one LD is computed every 2^{14} time ticks of the simulation. The first part of the simulation starts with correlated flow arrivals with a correlation length of 130 or about 2^7. The 3D-LD correctly identifies this component with a positive slope between octaves 1 and 7. The decreasing flow duration makes the initial dependence decrease with time, hence the slope of the 3D-LD which increases towards 0 when the timescale increases towards

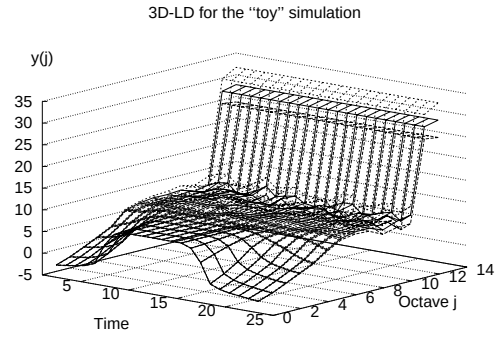

Figure 3: 3D-LD for the "toy" simulation.

one third of the simulation time. The second part of the simulation was a homogeneous Poisson process, hence uncorrelated, which is correctly identified by a slope of 0 in the 3D-LD. Finally, the third part of the process is similar to the first part of the simulation, but with a dependence that increases with time. Therefore, a slope of the 3D-LD which increases with time is seen. Notice the noise present at octave 7 and higher which arise due to the randomness in the simulation, the decrease in the number of the wavelet coefficients, and the increase in the confidence intervals of the LD. Also note that scaling in this simulation is statistical dependence and not self-similarity (or LRD) because it comes from dependence in a stationary process. This simulation illustrates the importance of the definitions found in section 2.1, where we differentiate between correlation-based scaling as seen in this simulation, self-similarity and LRD.

This "toy" example illustrates the advantage of the 3D-LD for studying signals whose scaling nature changes with time. Our example has shown that while the LD only gave an aggregated information about scaling occuring in the signal, the 3D-LD precisely identified when and over which timescales scaling occurs in the signal. We therefore propose the use of the 3D-LD for studying the scaling properties of complex signals for assessing their non-stationarity as well as their scaling nature.

3.2 Interpreting the 3D-LD in practice

There are some useful guidelines to be aware of when interpreting the 3D-LD. The first result seen in the 3D-LD consists of the time variations in the general level (energy) of each LD. A non-stationary signal will therefore provide a 3D-LD with shifts in the level of the surface that follow the non-stationarity of the signal for timescales larger than the size of a block over which any LD is computed. For example, the "toy" simulation has a constant mean, as indicated by the largest octaves of the 3D-LD which are flat. The largest octaves of the 3D-LD provide information about the signal's energy variations with time, at the timescales larger than the block size. Note that obviously the 3D-LD does not allow to study the timescales larger than the block size. The absence of dependence can be identified by a flat region. For example, in the second third of the "toy" simulation octaves 4 and below show no scaling of any sort. Random noise on the other hand, although it does not correspond to dependence, has an intrinsic variability which creates instabilities on the surface of the 3D-LD. This can be seen in the "toy" simulation for octaves 7 and higher. While the random number generators we used can be seen to exhibit some noise at octaves larger than 7 it cannot be mistaken as scaling. Finally, any region with a positive slope (for increasing values of the octaves) of the surface implies some kind of scaling, be it LRD, self-similarity or even bias from noise or components at larger timescales. The nature of scaling cannot be automatically inferred simply from the 3D-LD, it is a

graphical method that helps one to better understand the time-varying nature of the LD. The nature of scaling has to be found in the physical understanding of the process in addition to the information provided by the LD over each block. The interest of the 3D-LD is to better render the time-dependent nature of scaling provided by the LD.

4 Traffic traces

This section describes the two traffic traces used in this paper. A more detailed analysis of longer traffic traces can be found in [10].

4.1 Yucom

This trace consists of all flows between modem clients of a small Belgian ISP, Yucom. Yucom is a commercial ISP that provides Internet access to dialup users through regular modem pools. The trace spans less than 5 days between April 17^{th} and April 21^{st} 2001. It is a Netflow trace [17], not a packet trace. The collecting machine was located within the premises of the ISP and was running *OSU-flow-tools* [18]. Due to the flow timeout policy used by Netflow, we did not rely on the raw flow information as it over-estimates the number of flows. Instead we considered only the incoming flows that had the SYN flag set hence, TCP traffic only. This means that even if we do not have the correct TCP flows information because of the Netflow timeout, we have the same information concerning the new TCP flows (SYN). During the whole period of the trace, there were $59,581,814$ incoming TCP flows that had the SYN flag set. Among the 574 GBytes of the incoming TCP flows in the trace, http and https represents more than 72% of the traffic volume and more than 77% of the flows, chat 16.2% and 7.7%, and e-mail 5% and 6%.

4.2 PSC

This trace consists of all outgoing flows seen on two monitoring points internal to the Pittsburgh Supercomputing Center, an American commodity and Internet2 network provider to organizations in western Pennsylvania. The flows were collected continuously over 24 hours starting from March 12^{th} 2002 16:40 GMT at both monitoring points. These points covered all outbound commodity traffic from non-dormatory hosts at a large university, in addition to all outbound commodity traffic at several smaller organizations. The collection was made using the CoralReef package from CAIDA [19] using typical Netflow settings for flow expiration. During the flow period the network served by one monitoring point experienced considerable congestion, effectively imposing rate limits during peak usage times. We consider the TCP traffic only, but because we had no information about the TCP flags all terminated flows (FIN or CoralReef timeout) were treated as complete flows. There were $39,786,219$ TCP flows in the trace, representing 574 GBytes of IP traffic. The important applications we identified were Gnutella with 22.3% of the bytes and 12.3% of the flows, kazaa with 20.5% and 8.5%, http with 10.6% and 60.1%, so a large fraction of the bytes were generated by P2P applications.

5 Related work

Back in 1995, [6] questioned the Poisson assumption for IP flow arrivals. It showed that IP flows were not strictly Poisson, although different applications exhibited different behaviors. [20] studied in detail the process of the TCP connection arrivals and identified scaling but found that choosing between non-stationary Poisson or Weibull models was not a trivial matter. This paper continues the work from both [6] and [20] in an attempt to understand the nature of observed scaling of TCP flow arrivals, which we define as the TCP flows that are initiated from or arrive at the local network. Note however that contrary to [4], the aim of this paper is not to assess whether TCP flow arrivals are monofractal or characterized by the higher moments of multifractals, but only to study its basic (second-order) scaling properties.

More recently, [21] and [22] have shown that non-stationarity had to be added to long-range dependence and heavy-tailed marginal distributions as a fundamental property of Internet traffic. However, such a non-stationarity can have very different impacts at short timescales because of the artifical dependence it induces in the considered process. For example, an unbiased statistical analysis of the short-time TCP flow arrivals process is quite difficult due to strong variations of the signal's level throughout the trace. This means that relying on linear statistical techniques like classical Fourier analysis and some probabilistic techniques will be strongly biased due to the dependence in the shift of the signal's level between two contiguous time intervals. [23] and [24] both discuss the problem of non-stationarity when testing for long-range dependence.

6 TCP flow arrivals

In section 6.1 we first present the LD for the TCP arrivals process of the two traces and discuss the presence of scaling in them. We then go on to the 3D-LDs in section 6.2 and compare them with the LDs shown in section 6.1.

6.1 Logscale diagrams

Although wavelet coefficients have the desirable property of being almost decorrelated among octaves, the y_j's are computed as the average of the squared value of the coefficients, meaning that non-stationarity will be incorporated in the value of y_j. This implies that a stationary signal, and also some non-stationary signals, will have cor-

rect values of the y_j's independent on the trace length. In the case of some non-stationary signals on the other hand, the y_j's will contain the bias from the non-stationarity at the given octave.

Figure 4 presents the LD for the two traces. As we used a 100 ms resolution the first octave represents the 200 ms timescale. When looking for scaling in the LD, alignment must be present for some range of the considered octaves, but with consideration of the confidence intervals. Therefore, small confidence intervals require a

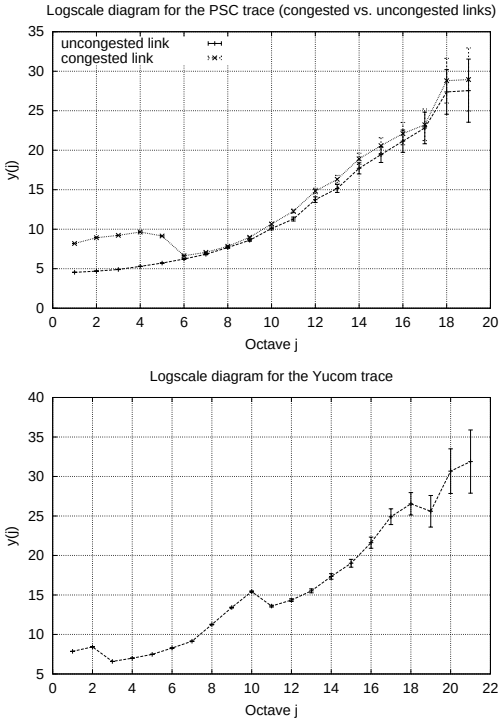

Figure 4: Logscale Diagrams.

very tight alignment. Conversely, when confidence intervals are large a looser alignment will be permitted. Both traces contain a bias from congestion or rate limitation, which induces a higher energy level at the short timescales. This can be seen in octaves 1 and 2 for Yucom and between octaves 1 and 5 for PSC. Other traces analyzed in [10] show that the flow arrivals of congestion-free traffic is approximately Poisson at timescales below a few seconds (octaves ≤ 5). In addition, only one client of the PSC trace undergoes rate limitation, so that the other part of the traffic exhibits a Poisson-like behavior at the short timescales as shown on the top part of figure 4. The medium timescales found between octaves 5 and 10, which extends from seconds to minutes, exhibit scaling although not strict. This phenomenon should be related to the correlation created by the flows within the user sessions, like in our "toy example" in section 3. The top part of figure 4 compares the two internal links from PSC. Here we can see that the rate limited link has this higher energy at small timescales while the other one has arrivals consistent with a Poisson process for timescales below a few seconds.

What can be said based on these LD's is that if scaling occurs in the data, then it is not strict nor over all timescales. In the case of Internet traffic where we know that non-stationarity exists, properly identifying the nature of TCP flow arrivals is important and therefore looking at the time-varying nature of the signal is necessary.

6.2 Time-dependent LDs

The previous section has shown that no scaling occured in the strict sense in the flow arrivals. In this section, we look at the time-varying structure of the TCP flow arrivals to determine whether or not scaling occurs in the signal. In section 6.2.1 we first analyze the small to medium timescales covering seconds to minutes. In section 6.2.2 we then proceed to analyze the larger timescales extending from minutes to hours.

6.2.1 Small to medium timescales

Figure 5 presents the 3D-LD for the uncongested link for PSC and two days of the Yucom trace. First, the Yucom trace clearly shows the dependence of the small timescales (octaves ≤ 5) energy on the time of the day, explaining the non-stationarity described in [21]. This means that the energy of http flows at the small timescales is guided by the level of the signal at the large timescales. The small timescales of PSC on the other hand are much

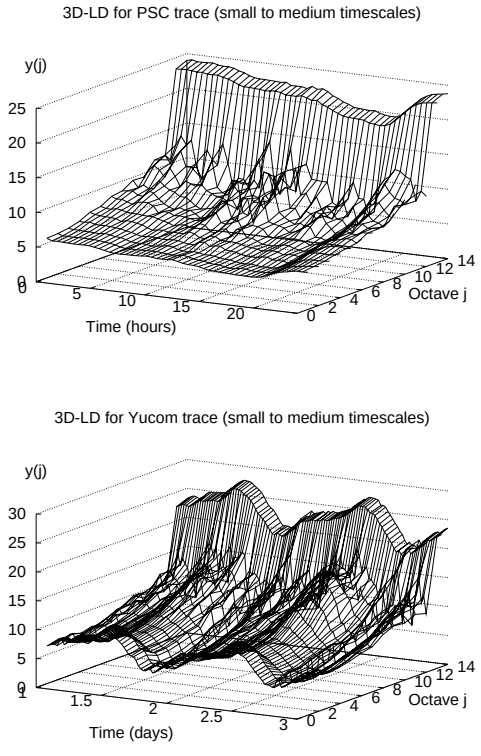

Figure 5: 3D-LDs for small to medium timescales.

less time dependent but the positive slope (both traces) tells that scaling is present. The value of the slope as well as our knowledge of the process both lead us to think that this is statistical dependence. The medium timescales (octaves 6 through 12 inclusive) show another type of non-stationarity; a distinct change in the scaling behavior. Both Yucom and PSC exhibit a changing 3D-LD for the timescales between several seconds and minutes. Yucom has a stronger scaling during peak hours due to more web users, hence more dependence at timescales corresponding to a typical web session, between seconds and minutes in length. PSC does not exhibit this strong dependence at these timescales due to a significantly smaller portion of http traffic in their trace; about 10 % of the flows compared to more than 70 % in the case of Yucom.

To verify our point concerning the influence of intra-session user behavior on scaling at timescales between seconds and minutes, we have split the uncongested part of the PSC trace into http flows (i.e. flows with the source or destination port $= 80$) and non-http traffic (having both source and destination ports larger than 1024) and compared the 3D-LD of these two different flow types. We do not show the corresponding figures only due to space limitations. The 3D-LD for http flows exhibits some dependence for the small timescales ($j \leq 5$) with a positive slope while the non-http flows do not, with a slope of 0 most of the time, hence consistent with an uncorrelated process. This positive slope we have for http flows comes from the fact that for any user surfing on the web, there are several TCP flows that are generated over the duration of his session, creating a statistical dependence similar to the one in our "toy example" of section 3.1. The medium timescales (octaves larger than 6) on the other hand have a non-stationary scaling for all flow types, evidenced by peaks on the 3D-LD; although the peaks of the http flows are partly hidden by the self-similarity or LRD. The origin of these peaks might be the user session arrivals coupled with statistical noise.

6.2.2 Medium to large timescales

This section studies medium to large timescales, which cover the range of minutes to hours. We do not study timescales larger than days because they exhibit an extremely regular behavior which provides no means for scaling to appear. Additionally, the PSC trace is too short to study such long timescales so we do not discuss this trace here. Instead, we refer the reader to [10] for a complete study of the TCP flows arrivals using longer traces.

When the different scaling regions of the signal overlap, the 3D-LD does not, as such, allow the identification of the actual nature of scaling at these timescales, but we can still assess whether the two regions above and below octave 15 have a different nature. Statistical dependence should be logically ruled out from the plausible nature of the process at these large timescales. No dependence between the flow arrivals for a given user or a group of users should arise at timescales of minutes to hours, instead, a self-similar or LRD process should appear. To verify that self-similarity or LRD really occurs at these timescales and is not a bias from the larger timescales we rely on the R/S statistic [11], which has been shown robust against departure from Gaussian marginals and also useful, but less robust, against sinusoidal components [25]. Figure 6 gives the R/S plot for the Yucom trace with timescales between one minute and one day, along with the two slopes corresponding to H equal 0.5 and 1. The R/S plot shows that for timescales between 1 minute and several hundred minutes self-similarity is present and that the estimated

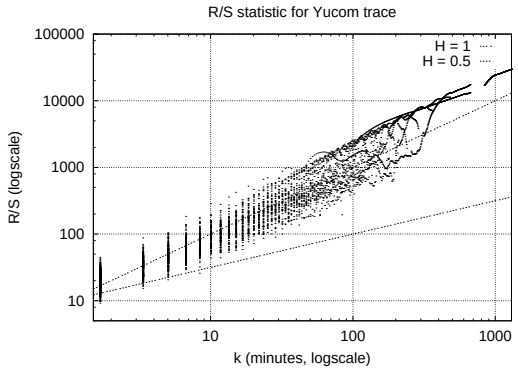

Figure 6: R/S statistic for medium to large timescales.

value of H is well above 0.5. On the other hand, for timescales between several hundred minutes and a day we detect the mark of seasonal components. This has the effect of changing the scatter of the dots, so that the stable dispersion of the dots for a given value of k (time aggregation) is reduced to almost one point and creates "lobes" in the plot. These lobes are the effect of the subharmonics of the detected cyclic components [25]. This would justify the modeling of timescales longer than a few hours with seasonal components [22] and the timescales between several minutes and hours by a self-similar or LRD process. This latter result is consistent with [26] that found self-similarity in the traffic volume for the same timescales.

7 Toward a model for TCP flow arrivals

Based on the results of section 6 and [10], we present in this section a simulated time-series based on the four components we identified:

1. the "time of the day" and "day of the week" seasonalities for timescales longer than hours,
2. a self-similar process for timescales between minutes and hours,
3. a process with non-stationary correlations for timescales between seconds and minutes,

4. a non-stationary Poisson process following the first two components.

The simulated time-series has no other purpose than illustrating the feasibility of generating a TCP flow arrivals process made of these four components. We do not wish to reproduce a signal having exactly the same 3D-LD as on figure 5. For instance, we could have worked directly in the wavelet domain as proposed in [27]. Generating a time-series in this way relies on generating the wavelet coefficients of the signal, scaling their statistical properties at each octave, and then applying the inverse wavelet transform to obtain the time-series. While this method will obviously match the wavelet estimates more closely than any other method (and be more efficient computationally), it requires that the wavelet domain accurately describes the properties of the signal. We saw this not to be true in the case of such a strongly non-stationary signal like the TCP flow arrivals. Therefore, we preferred to work in the time domain and stay close to the way TCP flow arrivals are likely to be generated. A pseudo-code description as well as scripts used for this simulation are presented in [10].

Like both traces analyzed in this paper we use a time-granularity of 100 ms (the time tick). The basis of the model is the "daily periodicity", the first kind of non-stationarity exhibited in network traffic. This part of the signal can be simulated by a trigonometric series matching both the "day of the week" and "hour of the day" periodicities. The issue with this component is that if applied directly on the smallest timescales it would bias them and generate false scaling at all timescales. Hence, we divided the signal into blocks of length 2^{16}. For this seasonal component, we simply used a sinusoid of period equal to one day (864000 time ticks), with a mean of 500 and an amplitude of 150. The fit of the real behavior could be improved by using a more complex combination of sinusoids.

Proceeding to smaller timescales, from minutes to several hours, we have the self-similar or LRD component which was generated via a self-similar process with Gaussian increments. Again we divided the sinusoidal seasonality into blocks of 2048 time ticks. To each block a shift equal to the cumulative sequence of the Gaussian increments was added. Figure 7 presents the results of these two steps, where we show the 3D-LD computed over intervals of size 2^{17}. An important issue concerning the simulation of a self-similar signal based on iid increments comes from the limited control one has over the non-stationarity of the generated process. It should be controlled through truncation of the large values generated by the random numbers generator. Working in the spectral domain is another option to better match a particular value of H [28, 29]. The largest octaves of figure 7 show that the sinusoidal component has been modified by the non-stationarity of the self-similar component. Nevertheless, matching the behavior of a real trace requires some adjustments in the self-similar component in order

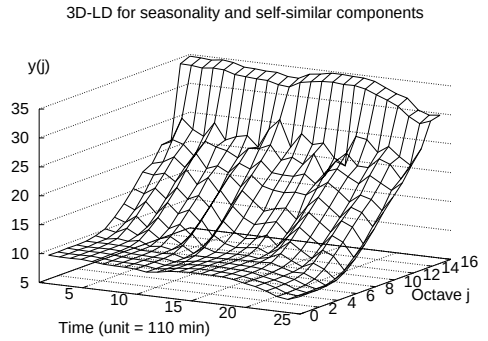

Figure 7: 3D-LD for the simulated seasonality and self-similar components.

to properly simulate the average evolution of the process.

The remaining two components are the correlated process between 2^6 and 2^{10} time ticks and the Poisson process. We first generated a Poisson process with mean equal to the series made of the seasonal component and the self-similar process found at the largest timescale. We then simulated the correlated process by generating several flows for some Poisson flows over these timescales by creating one additional flow arrival every n_{min} time intervals, n being the smallest timescale at which this LRD should appear. On the other end a maximum time length n_{max}, which corresponds to the larger timescale where this LRD should appear. We implemented the LRD non-stationarity as seen in the traces by activating the creation of the correlated flows to generate LRD for some blocks and not for others. Additionally, we activated the LRD process for the time blocks where the sinusoid had a value superior to 550. During these peak hours we allowed a fixed percentage (whose value depended on the sinusoidal component) of the total flows to exhibit correlation at the required timescales. The remaining flows were left unchanged.

Figure 8 shows the 3D-LD for the whole simulation with all components of the signal using blocks of size 2^{17}. Each LD of the 3D-LDs in this section represents roughly

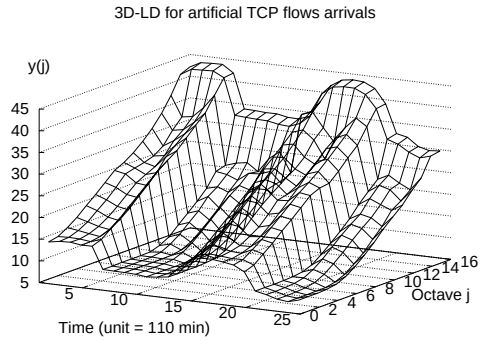

Figure 8: 3D-LD for the simulated TCP flow arrivals.

110 minutes of simulated time. What appears first in this 3D-LD is the increase in total energy of the signal due to the correlated process over the medium timescales. By adding correlated flows between octaves 6 and 10 we also increased the variability of the process, hence the important difference between the peak hours level of each LD and other periods. Although it is difficult to distinguish between the self-similar component and the correlated component, we can see that scaling for the medium timescales, around octave 6, only appear during the peak hours. These occur in the early hours of the simulation and also one day later. The LD allows for easier identification of the three components in this simulation as the variability generated by both the correlated component and the self-similar one modify and partly hide each other.

Finally, figure 9 presents the LD for the simulated TCP flow arrivals. In this figure we can easily identify the three main components, namely: the Poisson process over the smallest octaves, the correlated component up to octave 10, and the self-similar component between octaves 11 and 15. The LD clearly shows that the correlated component does not have exact scaling, and so cannot correspond to a strictly self-similar process. Octaves between 10 and 17, on the other hand, show an approximate scaling illustrating the spread of self-similarity (or LRD) to timescales where it has not been produced. A bias from the self-similar component is responsible for this phenomenon. This illustrates the difficulty of generating (and analyzing) a signal with different scaling occuring at different timescales. Therefore, even if the 3D-LD can reduce the bias of non-stationarity at large timescales, a self-similar component in the signal may still affect timescales where scaling does not, in fact, exist.

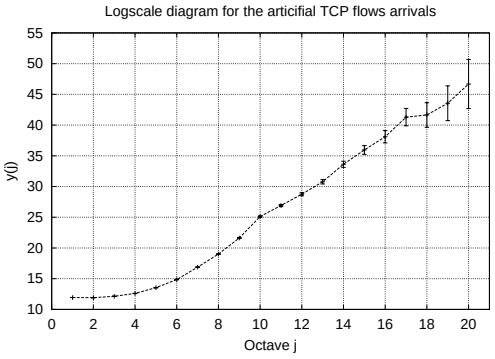

Figure 9: LD for the simulated TCP flow arrivals.

8 Conclusion

In this paper we have proposed a new graphical method for studying non-stationary and scaling processes with the 3D-LD. We have shown throughout this paper that this graphical method is useful for analyzing complex signals whose scaling properties vary with time. In consequence, we argue that this method is useful in every situation where (second-order) signals are expected to present non-stationarity and scaling. Even in the situation of a stationary process, or in the absence of scaling, the 3D-LD is useful to check for stationarity.

We have also studied the process of the TCP flow arrivals based on 2 recent IP traffic traces. We first studied the scaling properties over the whole traces and showed that no strict scaling was present. We also detected an abnormal signature over the timescales of several seconds and less, which we found to be caused by rate limitation. We then looked at the time-dependent scaling properties of the traffic using the 3D-LD. It showed the first type of non-stationarity of the TCP flow arrivals process, namely its dependence on the "time of the day" and the "day of the week" seasonalities. We also showed that a non-stationary Poisson process could be identified for timescales up to a couple of seconds for non-http traffic. In contrast, http traffic contained statistical dependence even at the subsecond timescales. A second type of non-stationarity was found within the timescales from seconds to several minutes, that rendered the scaling at these timescales highly irregular although dependent on peak hours. We identified this scaling as dependence due to the "within session" behavior of the users that generates correlations among TCP flows for timescales between seconds and minutes. Then, looking at the timescales of minutes to hours revealed self-similar (or LRD) scaling, which we attributed to the user sessions variability. TCP flow arrivals therefore exhibit scaling properties inconsistent with pure LRD or self-similarity, but scaling in TCP flows arrivals is time- and timescale-dependent.

Finally, we generated an artificial TCP flow arrivals process based on the different components identified in the traffic traces. We showed that it was possible to get qualitatively close to the characteristics of a real TCP flow arrivals process, but also that generating a complex signal with different scaling properties over different timescales was not trivial due to the bias of self-similar scaling.

Acknowledgements

We acknowledge all the people having contributed to the collection of the used traffic traces, and particularly B. Piret from BT Ignite Belgium for the Yucom trace. We are also grateful to the ITC reviewers for their constructive remarks. This work was partially supported by the European Commission within the IST ATRIUM project.

References

[1] W. Leland and D. Wilson, "High time-resolution measurement and analysis of lan traffic: Implica-

tions for lan interconnection," in *Proc. of* IEEE INFOCOM'91, 1991.

[2] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," IEEE/ACM *Transactions on Networking*, 1994.

[3] K. Park and W. Willinger (editors), *Self-Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, 2000.

[4] A. Feldmann, A. Gilbert, W. Willinger, and T. Kurtz, "The changing nature of network traffic: scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.

[5] V. Paxson, "Empirically-derived analytic models of wide-area tcp connections," IEEE/ACM *Transactions on Networking*, vol. 2, no. 4, pp. 316–336, 1994.

[6] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modeling," IEEE/ACM *Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[7] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic evidence and possible causes," in *Proc. of* ACM SIGMETRICS'96, May 1996, pp. 160–169.

[8] G. Kim K. Park and M. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," in ICNP'96, 1996.

[9] A. Downey, "Evidence for long-tailed distributions in the internet," in *Proc. of* ACM SIGCOMM *Internet Measurement Workshop*, November 2001.

[10] S. Uhlig, "3D-LD: a graphical wavelet-based tool for analyzing non-stationary and scaling processes," Infonet technical report Infonet-2002-04, available at `http://www.infonet.fundp.ac.be/doc/tr/`, April 2002.

[11] J. Beran, *Statistics for Long-Memory Processes*, Monographs on Statistics and Applied Probability, Chapman & Hall, 1994.

[12] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," *In [3]*.

[13] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The multiscale nature of network traffic: Discovery, analysis, and modelling," IEEE *Signal Processing Magazine*, vol. 19, no. 3, pp. 28–46, May 2002.

[14] B. Mandelbrot, *Gaussian Self-affinitiy and Fractals: Globality, The Earth, 1/f Noise, and R/S*, Springer-Verlag, 2001.

[15] I. Daubechies, *Ten Lectures on Wavelets*, Number 61 in CMBS-NSF Series in Applied Mathematics, SIAM, Philadelphia, 1992.

[16] P. Abry D. Veitch, "A statistical test for the time constancy of scaling exponents," IEEE *Transactions on Signal Processing*, vol. 49, no. 10, pp. 2325–2334, 2001.

[17] Cisco, "Netflow services and applications," White paper, available from `http://www.cisco.com/warp/public/732/netflow`, 1999.

[18] S. Romig, M. Fullmer, and R. Luman, "The osu flow-tools package and cisco netflow logs," in *Proc. of* USENIX LISA, December 1995.

[19] K. Keys, D. Moore, R. Roga, E. Lagache, M. Tesch, and K. Claffy, "The architecture of coralreef: an internet traffic monitoring software suite," in *In Proc. of* PAM2001, April 2001.

[20] A. Feldmann, "Characteristics of tcp connection arrivals.," *In [3]*.

[21] J. Cao, W. Cleveland, D. Lin, and D. Sun, "On the nonstationarity of internet traffic," in *Proc. of* ACM SIGMETRICS 2001, June 2001, pp. 102–112.

[22] W. Cleveland, D. Lin, and D. Sun, "Ip packet generation: Statistical models for tcp start times based on connection-rate superposition," in *Proc. of* ACM SIGMETRICS 2000, June 2000, pp. 166–177.

[23] T. Dang and S. Molnar, "On the effects of non-stationarity in long-range dependence tests," Periodica Polytechnica *Ser. El.*, vol. 43, no. 4, pp. 227–250, 1999.

[24] M. Roughan and D. Veitch, "Measuring long-range dependence under changing traffic conditions," in *Proc. of* INFOCOM'99, 1999.

[25] B. Mandelbrot, "Robustness of R/S in measuring noncyclic global statistical dependence," *Water Resources Research*, , no. 5, pp. 967–988, 1969.

[26] S. Uhlig and O. Bonaventure, "Understanding the long-term self-similarity of internet traffic," in *Proc. of* QOFIS'01, September 2001.

[27] S. Ma and C. Ji, "Modeling heterogeneous network traffic in wavelet domain," IEEE/ACM *Transactions on Networking*, vol. 9, no. 5, pp. 634–649, 2001.

[28] V. Paxson, "Fast, approximate synthesis of fractional gaussian noise for generating self-similar network traffic," *Computer Communication Review*, vol. 27, pp. 5–18, 1997.

[29] S. Ledesma and D. Liu, "Synthesis of fractional gaussian noise using linear approximation for generating self-similar network traffic," *Computer Communication Review*, vol. 30, no. 3, pp. 4–17, 2000.