The Multiscale Nature of Network Traffic: Discovery, Analysis, and Modelling

Patrice Abry, Richard Baraniuk, Patrick Flandrin, Rudolf Riedi, Darryl Veitch

To appear in IEEE Signal Processing Magazine, April 2002.

Abstract

The complexity and richness of telecommunications traffic is such that one may despair to find any regularity or explanatory principles. Nonetheless, the discovery of scaling behavior in tele-traffic has provided hope that parsimonious models can be found. The statistics of scaling behavior present many challenges, especially in non-stationary environments. In this paper, we overview the state of the art in this area, focusing on the capabilities of the wavelet transform as a key tool for unravelling the mysteries of traffic statistics and dynamics.

Keywords

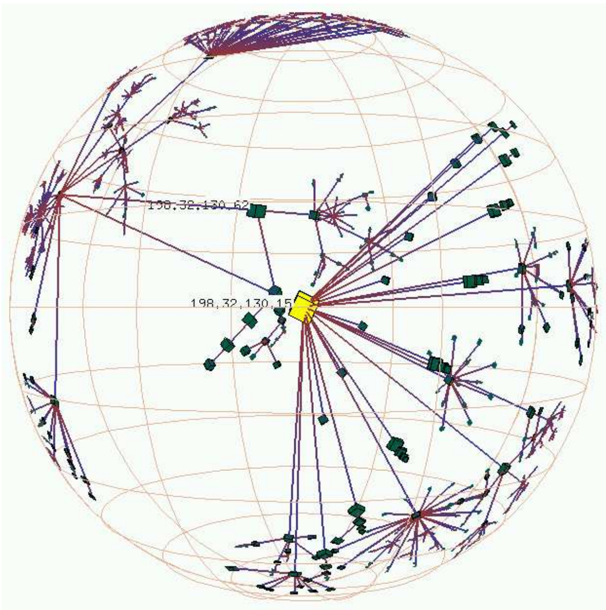
Computer network traffic, Tele-traffic, Wavelets, Scaling, Self-similarity, Long-Range Dependence, Fractals, Multifractals, Cascade Processes, Multiplicative Processes, Infinitely Divisible Cascades, Fractional Brownian Motion.

I. Traffic and Scaling

By the term *telecommunications traffic* or *tele-traffic* we mean the flow of information, or data, in telecommunications networks of all kinds. From its origins as an analog signal carrying encoded voice over a dedicated wire or "circuit", traffic now covers information of all kinds, including voice, video, text, telemetry, and real-time versions of each, including distributed gaming. Instead of the dedicated circuits of traditional telephone networks, *packet switching* technology is now used to carry traffic of all types in a uniform format (to a first approximation): as a stream of packets, each containing a header with networking information and a payload of bytes of "data".

Although created by man and machine, the complexity of teletraffic is such that in many ways it requires treatment as a natural phenomenon. It can be likened to a turbulent, pulsating river flowing along a highly convoluted landscape, but where streams may flow in all directions in defiance of gravity. The landscape is the network. It consists of a deep hierarchy of systems with complexity at many levels. Of these, the "geographical" complexity or connectivity of network links and nodes, illustrated in Box 1, is of central importance. Other key aspects include the size or *bandwidth* of links (the volume of the river beds), and at the lowest level, a wide variety of physical transport mechanisms (copper, optic fibre, etc.) exist with their own reliability and connectivity characteristics. Although each atomic component

Box 1: Tele-Traffic: A Turbulent River over a Rugged Landscape



The geographic and topological complexity of the Internet "infoways" has reached a point that it is now a significant challenge to provide even rough maps of the major tributaries. The *Skitter* program, a CAIDA (Cooperative Association for Internet Data Analysis *http://www.caida.org/*) project, attempts to provide maps such as the one shown here, tracing connectivity of hosts throughout the Internet by sending messages out to diverse destinations and counting the number of links traversed to reach them. Each line represents a logical link between nodes, passing from red on the outbound side to blue on the inbound. The data visible here is only a small part of a large dataset of around 29,000 destinations.
(Figure reproduced with the kind permission of CAIDA. copyright 2001 CAIDA/UC Regents. Mapnet Author: Bradley Huffaker, CAIDA. The three dimensional rendering is provided by the *hypviewer tool*.)

P. Abry and P. Flandrin are with the ENS Lyon, France, E-mail: {pabry,flandrin}@ens-lyon.fr.
R. Baraniuk and R. Riedi are with the ECE Dept., Rice University, E-mail: {richb,riedi}@rice.edu.
D. Veitch is with EMUlab, University of Melbourne, Victoria, Australia. E-mail: d.veitch@ee.mu.oz.au.

is well-understood, the whole is so complex that it must be measured and its emergent properties "discovered". Comprehensive simulation is difficult.

A key concept in networking is the existence of network *protocols*, and their encapsulation. Let us explain with an example: The *Internet protocol* (IP) is used to allow the transport of packets over heterogeneous networks. The protocol understands and knows how to process information such as addressing details contained in the *header* of IP packets. However, by itself IP is only a forwarding mechanism without any guarantee of successful delivery. At the next higher level, the *transfer control protocol* (TCP) provides such a guarantee by establishing a virtual *connection* between two end points and monitoring the safe arrival of IP packets, and managing the retransmission of any lost packets. On a still higher level, web-page transfers occur via the *Hypertext transport protocol* (HTTP), which uses TCP for reliable transfer.

The resulting *encapsulation* "HTTP over TCP over IP", therefore means that HTTP oversees the transfer of text and images etc, while the actual data files are handed over to TCP for reliable transfer. TCP chops the data into datagrams (packets) which are handed to IP for proper routing through the network. This organization offers hierarchal structuring of network functionality and traffic but also adds complexity: each level has its own dynamics and mechanisms, as well as time scales.

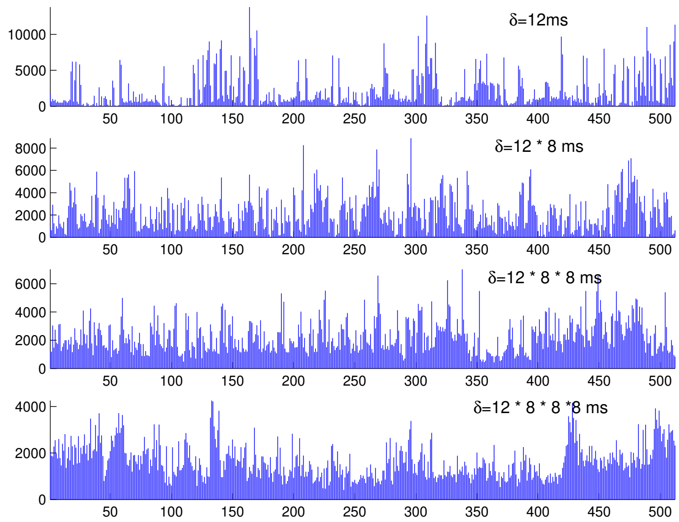
Over this landscape flows the teletraffic, which has even more levels of complexity than the underlying network. Three general categories can be distinguished.

Geographic complexity plays a major role. Although one can think of the Internet as consisting of a "core" of very high bandwidth links and very fast switches, with traffic sources at the network "edge", the distances from the edge to the core vary greatly, and the topology is highly convoluted. Access bandwidths vary widely, from slow modems to gigabit Ethernet local area networks, and mobile access creates traffic which changes its spatial characteristics. Sources are inhomogeneously distributed, for example concentrations are found in locations such as universities and major corporations. Furthermore traffic streams are split and recombined in switches in possibly very heterogeneous ways, and what is at one level a superposition of sources can be seen at another level, closer to the core, as a single, more complex kind of "source".

Offered Traffic complexity relates to the multilayered nature of traffic demands. Users, generating web browsing sessions for example, come and go in random patterns and remain for widely varying periods of time, during which their activity levels (number of pages downloaded) may vary both qualitatively and quantitatively. The users' applications will themselves employ a variety of protocols that generate different traffic patterns, and finally, the underlying objects themselves, text, audio, images, video, have widely differing properties.

Box 2: Temporal Burstiness in Traffic

Here, we present an analysis of a standard trace of Ethernet traffic, "pAug" from [14]. An entry $Y(k)$ of this time series represents the number of bytes observed on the Ethernet at Bellcore during the k-th time slot of duration $\delta = 12$ms of the measurement. Denote by $Y^{(m)}$ the *aggregated series of level* m; for example $Y^{(3)}(1) = (Y(1) + Y(2) + Y(3))/3$ represents then the average traffic observed in time slots of duration 3δ. Through this averaging operator, scale invariance can be illustrated in a simple but powerful way . From top to bottom, the first 512 points of four series are plotted: $Y(k) = Y^{(1)}(k)$, $Y^{(8)}(k)$, $Y^{(64)}(k)$, and $Y^{(512)}(k)$, with δ varying from $\delta = 12$ms to $\delta = 12 * 8 * 8 * 8$ms, or 6.1s.



The decrease in variability with increased smoothing is very slow, consistent with

$$\mathrm{Var}[Y^{(m)}] = O(m^{-\beta}), \quad \beta \approx 0.4 \in (0,1)$$

the so called "slowly decaying variance" of long memory processes.
A wavelet analysis of this series appears in Figure 8, middle plot.

Temporal complexity is omnipresent. All of the above aspects of traffic are time varying, and take place over a very wide range of time-scales, from microseconds for protocols acting on packets at the local area network level, through daily and weekly cycles, up to the evolution of the phenomena themselves over months and years.

The huge range of time-scales in traffic and the equally impressive range of bandwidths, from a kilobytes up to terabytes per second over large optical backbone links, offers enormous scope for scale dependent behavior in traffic. But is this scope actually “exploited” in real traffic? Is traffic in fact regular on most time scales, with variability easily reducible to, say, a diurnal cycle plus some added variance arising from the nature of the most popular data-type/protocol combination? Since the early nineties, when detailed measurements of packet traffic were made and seriously analyzed for the first time [21], [15], [14], we know that the answer is an emphatic “No”. Far from being smooth and dominated by a single identifiable factor, packet traffic exhibits *scale invariance* features, with no clear dominant component.

For instance, long memory is a scale invariance phenomenon that can be seen in the time series $Y(t)$ describing the data transfer rate over a link at time t. Other examples of time series with long memory are the number of active TCP connections in successive time intervals, or the successive interarrival times of IP packets shown in Figure 1.

The philosophy of scale invariance or “scaling” can be expressed as the lack of any special characteristic time or space scale describing fluctuations in $Y(t)$. Instead one needs to describe the steady progression *across* scales. In the case of traffic such a progression has been found empirically and has lead to long memory models and more generally to models with fractal features, as we will explore.

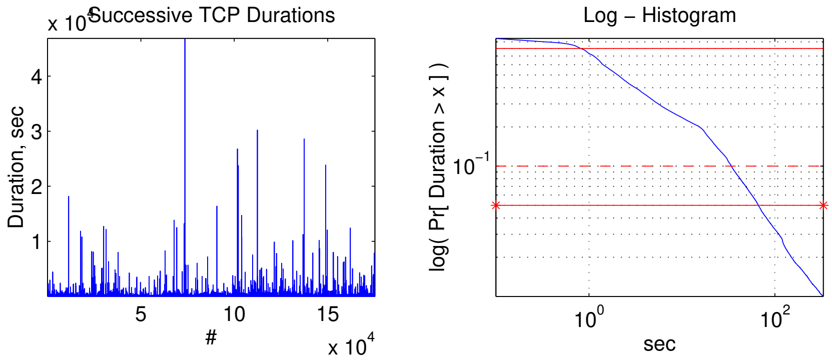
The scale invariant features of traffic can also be thought of as giving precise meaning to the important but potentially vague notion of traffic *burstiness*, which means, roughly, a lack of smoothness. In very general terms, burstiness is important because from the field of performance analysis of networks, and in particular that of switches via queueing theory, we know that increased burstiness results in lower levels of resource utilization for a fixed quality of service, and therefore to higher costs. At the engineering level, service quality refers to metrics such as available bandwidth, data transfer delay, and packet loss. The impact of scale invariance extends to network management issues such as call admission control, congestion control, as well as policies for fairness and pricing.

It is important to distinguish between two canonical meanings of the term burstiness, which have their counterparts in models and analysis. Again let us take “traffic” to be the data *rate* $Y(t)$, nominally in bytes per second, over a link at time t. One kind of burstiness arises from dependencies over long time periods, which can be made precise in terms of the correlation function of $Y(t)$ (assuming stationarity and that second order statistics exist). As shown in Box 2, such *temporal burstiness* was explored when scaling was first found in packet traffic. More precisely, the well known Long-Range Dependent (LRD) property of traffic is a phenomenon defined in terms of temporal correlation, whose network origins are now thought to be quite well understood in terms of the paradigm of heavy tails of file sizes of requested objects, which causes sources to transmit over extended periods [36].

A second kind of burstiness describes variability, the size of fluctuations in value or amplitude, and therefore concerns small scales. It refers therefore to the marginal distribution of $Y(t)$, as characterized for example by the ratio of standard deviation to mean if this exists, as the local singular behavior of multifractal models (described in the next section), or alternatively as a heavy tail parameter

Box 3: Amplitude Burstiness in Traffic.

Consider a particular time series derived from Internet data, the *durations* (in seconds) of successive TCP connections $\mathrm{dur}(k)$, $k = 1, 2 \cdots 175223$, for connections beginning during a 6.4 hour long subset of a much larger trace. The subset was selected for apparent stationarity across a range of criteria.
The left plot shows the time series. Gaussian models can provide in some cases reasonable approximations to traffic traces, but certainly not here. Indeed, the *sample* standard deviation to mean ratio is $\approx$ 12, which given the natural constraint of positivity for the series, is decidedly non-normal!



The marginal of the series is examined in the right plot, in a log-log plot of the sample complementary probability distribution function $P(\mathrm{dur}) > x$. The roughly straight line strongly suggests a heavy power-law like tail, with an index which is close to the boundary of infinite variance. The horizontal lines highlight, from top to bottom, the 20%, 90%, and 95% quantiles respectively.

of the distribution of the instantaneous traffic load in the case of infinite variance models. Box 3 illustrates this latter case for the time series of successive TCP connection durations, derived from measurements taken over a 2Mbps access link, made available at the University of Waikato [22]. Even when an apparently stationary subset is selected, the variation in value or amplitude is very significant, and highly non-Gaussian. Marginals of other time series do not always yield such extreme power-law tails; however Weibullian or log-normal behavior is more common than Gaussian, unless the data has already been highly aggregated or if scales above a few seconds are examined.

The two types of burstiness just described are quite different. However, often it is convenient to work not with a stationary series like $Y(t)$, but with its integrated or "counting process" equivalent $N(t)$, which counts the amount of traffic arriving in $[0, t]$. It is then important to bear in mind that the statistics of $N(t)$ are a function both of the temporal and the amplitude burstiness of the rate process $Y(t)$.

The next step in this introduction to scaling in traffic is to draw attention to the fact that, although at large scales (seconds and beyond) astonishingly clear, simple and relatively well understood scaling laws are found, the same cannot be said at small scales. This is true for example of the *inter-arrival* time series shown in Figure 1, a discrete series giving the successive intervals (in milliseconds) between the arrival of new TCP connections. When examined with the naked eye this series may be accused of having long memory, with a marginal slightly deviating from Gaussianity. In reality, in addition to long memory, it contains much non-trivial scaling structure at small scales (see Figure 8) which is suggestive of a rich underlying dynamics of TCP connection creation. Investigation of such dynamics is beyond the scope of this review, however knowledge of its scaling properties, as examined in section 3 (see [32] for more details), lays a foundation for an informed investigation.

The fact is that much work remains to be done to achieve a clear understanding of traffic scaling over small scales, which is characterized by far higher variability, more complex and less definitive scaling laws, and the necessity of dealing with non-Gaussian data and hence statistics beyond second order. The high variability on small scales is shown in Figures 2 and 3 for a publicly available trace collected at the Lawrence Berkeley Laboratory. The time series of the number of TCP packets arriving per time interval has very irregular local structure, as seen in the blowups in the lower plots. While large scale behavior such as long memory matters for many network design and management issues, understanding small scale behavior is particularly important for flow control, performance and efficiency. In terms of network performance, variability is (almost) always an undesirable feature of traffic data. Therefore, a key motivation for investigating such scaling is to help identify generating mechanisms leading to an understanding of their root causes in networking terms. If for example it were known that a certain feature of the TCP protocol was responsible for generating the observed complex scaling behavior at small scales, then we would be in a position to perhaps eliminate or moderate it via modifications to the protocol. Alternatively, if a property of certain traffic source types was the culprit, then we could predict if the scaling would persist in the future or fade away as the nature of telecommunications services evolve.

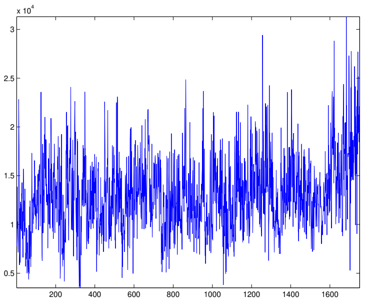


Fig. 1. A series of inter-arrival times of TCP connections, showing highly detailed local structure as well as long memory.

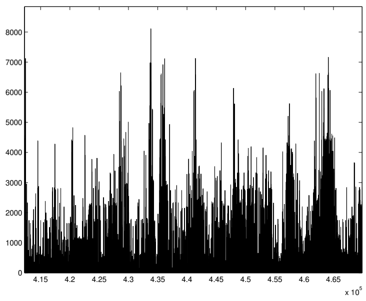


Fig. 2. A snap shot (seconds $415'000 - 470'000$) of the LBL trace of packet arrival per time depicting erratically varying regularity.

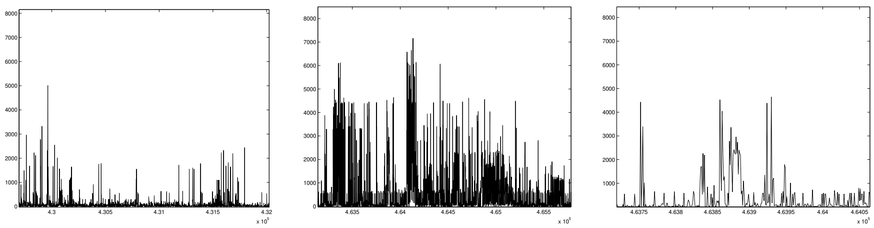


Fig. 3. Zooms: $429'500 - 432'000$, $463'000 - 466'000$, and again: $463'700 - 464'100$.

Demonstrating the existence of long memory as well as the interwoven coexistence of smooth and bursty periods at all times.

To conclude this introduction to scaling in telecommunications, we point out that in many series derived from traffic data, in particular TCP/IP traffic and including the data in Figure 1, (see Figure 8, right most plot and Box 12), a recurring feature is the existence of a characteristic scale at around 1 second, which separates the now classic "mono-scaling" at large scales indicative of long memory, from the more complex, but none-the-less scaling behavior, at small scales. *Multifractal* models are one possible approach for the latter domain, whereas *infinitely divisible cascades* offer the possibility of integrating both regimes in a single description. In the following two sections we will describe these models and the associated traffic phenomena in detail, together with wavelet based statistical methods which enable them to be effectively explored.

II. Scaling Models

The notion of *scaling* is defined loosely, as a negative property of a time series: the absence of characteristic scales. Its main consequence is that *the whole and its parts* cannot be statistically distinguished from each other. The absence of such scales means that new signal processing tools are needed both for analysis and modelling, whilst standard techniques built on characteristic times (for example simple Markov models) must be abandoned. This section provides an introductory review of various models used to give flesh to the spirit of scaling.

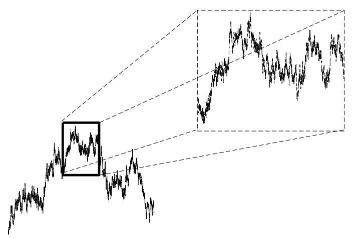


Fig. 4. **Statistical Self-Similarity.** A dilated portion of the sample path of a self-similar process cannot be (statistically) distinguished from the whole.

Self-Similarity. The purest formal framework for scaling is undoubtedly that of *exactly self-similar* processes. Self-similarity (see Figure 4 for an illustration, Box 4 for a technical definition and, e.g., [33] for further information) means that the sample paths of the process $X(t)$ and those of a rescaled version $c^H X(t/c)$, obtained by simultaneously dilating the time axis by a factor $c > 0$, and the amplitude axis by a factor c^H, cannot be statistically distinguished from each other. H is called the self-similarity or Hurst parameter. Equivalently, it implies that an affine dilated subset of one sample path cannot be distinguished from its whole. It is therefore not possible to identify a reference scale of time, and thus there is no such reference scale. Exact statistical self-similarity thereby fulfils the intuition of scaling in a simple and precise way.

Self-similar processes are, by definition, non stationary, as can be seen from equation (2). However the most important subclass, namely *self-similar processes with stationary increments* (H-sssi processes), are non-stationary in a very homogeneous way. They can be thought as the integral of some stationary process. Fractional Brownian motion is the unique Gaussian self-similar process with stationary increments, and is the most widely used process to model scaling properties in empirical times series. For example it has been used to model the data shown in Box 2, more specifically to model the variability of the number of Ethernet bytes in the interval $[0, t]$. Practically, self-similarity is usually tested for and analyzed through its increments and the relation (6).

Box 4: Self Similar Processes with Stationary Increments.

A process $X(t)$ is said to be self-similar, with self similarity parameter $H > 0$, if

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

where $\stackrel{d}{=}$ means equality for all finite dimensional distributions. A major consequence of this definition is that the moments of X, provided they exist, behave as power-laws of time:

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

For applications, one usually restricts the class of *self-similar processes* to that of *self-similar processes with stationary increments* (or H-sssi processes). A process X is said to have stationary increments $Y(\delta, t)$ if

$$\{Y(\delta, t) := Y_\delta(t) := X(t+\delta) - X(t),\ t \in \mathcal{R}\} \stackrel{d}{=} \{X(\delta) - X(0)\}, \forall \delta, \tag{3}$$

or, in other words, if none of the finite dimensional laws of $Y(\delta, t)$ depend on t.
For a H-sssi process X, the self-similarity parameter necessarily falls in $0 < H < 1$ and the covariance function, when it exists, takes a specific, unique, and constrained form: $\mathbb{E}X(t)X(s) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right), \sigma^2 = \mathbb{E}|X(1)|^2$. Moreover, it can be shown that the autocovariance function of the increment process Y_δ reads:

$$\mathbb{E}Y_\delta(t)Y_\delta(t+s) = \frac{\sigma^2}{2}\left(|s+\delta|^{2H} + |s-\delta|^{2H} - 2|s|^{2H}\right). \tag{4}$$

The self similarity of the process X is transferred to its increments insofar as:

$$Y(\delta, t) \stackrel{d}{=} c^H Y(\delta/c, t/c), \tag{5}$$
$$\mathbb{E}|Y(\delta, t)|^2 = \mathbb{E}|X(t+\delta) - X(t)|^2 = \sigma^2 |\delta|^{2H}. \tag{6}$$

Limitations of Self-Similarity. Self-similar processes with stationary increments, and more specifically fractional Brownian motions, are very attractive models to describe scaling because they are mathematically well-defined and well-documented. In addition, their great advantage lies in being simple and parsimonious: each of their properties is defined and controlled by the same parameter, H. Their main drawback however, lies in them being . . . simple. It is indeed unlikely that the wide variety of scaling encountered in data can be modelled by a process with a single parameter. The model is overly rigid in several respects. First, definition 1 is valid for all positive real c, which means that the scaling exists for all scales or dilation factors ranging from 0 to ∞. Equivalently, one can say, looking at equation (5), that the scaling relation holds whatever the value of the scaling factor. In actual real world data, scaling can naturally exist only within a finite range of scales and will typically only be approximative.

Moreover, one may find evidence for scaling only in the *asymptotic* regions, i.e., only within the very large (or the very small) scales. Second, self-similarity implies (see equation (2) that scaling holds for each moment order q (provided it exists), with scaling exponent qH. In empirical data, moments of different orders may have scaling exponents that are not controlled by a single parameter, and some moments may simply not exhibit scaling at all. Even worse, the empirical moments might be misleading when the theoretical moments of the true distribution do not exist at all, as is the case with stable laws. In the case of traffic data, most often scaling models with a single parameter are appropriate at large scales, but at small scales more parameters are required. In rarer cases, definitive evidence for scaling is lacking altogether. Infinite moments can play a role for quantities such as TCP connection durations, but in term of scaling models, those most commonly used are of the finite (positive) moment type.

The remainder of this section details more flexible models that enable such deviations from exact self-similarity. We first explore those that concentrate on scaling in second order statistics, that is, involving autocovariance functions and spectra or power spectral densities. Processes whose spectra obey a power-law within a given (and sufficiently wide) range of frequencies (scales) are often referred to as $1/f$ processes:

Box 5 : Long-Range Dependence

Let $\{X(t), t \in \mathcal{R}\}$ denote a second-order stationary stochastic process, and r_X and Γ_X its covariance function and spectral density. We will say that the process $\{X(t), t \in \mathcal{R}\}$ is *Long-Range Dependent* (LRD) if either

$$r_X(\delta) \sim c_1|\delta|^{\gamma-1}, \ \delta \to +\infty, \ \gamma \in (0,1) \tag{7}$$

or

$$\Gamma_X(\nu) \sim c_2|\nu|^{-\gamma}, \ \nu \to 0, \ \gamma \in (0,1)\,, \tag{8}$$

with $c_2 = 2(2\pi)^{-\gamma}\Gamma(\gamma)\sin((1-\gamma)\pi/2)c_1$. In most practical situations, r_X is regularly varying or even asymptotically monotone, in which case these relations are in fact equivalent.

With this definition, the autocovariance function decreases so slowly, the past is so weighty, that its sum diverges, i.e., for any $A > 0$,

$$\int_A^\infty r_X(\delta)d\delta = \infty.$$

$$\Gamma_X(\nu) = C_0|\nu|^{-\gamma}, \ \nu_m \leq |\nu| \leq \nu_M.$$

The two special cases where the scale range is semi-infinite, either at small frequencies, $\nu_m \to 0$ (equivalently, large scales) or at large frequencies, $\nu_M \to \infty$ (small scales), define two interesting models, namely those of *Long-Range Dependent* processes (see Box 5) and *monofractal processes* (see Box 6).

Long-Range Dependence. Long-range dependence (LRD) or long memory [5] is a model for scaling observed in the limit of the largest scales, and is defined in terms of second-order statistics (see Box 5). LRD is usually equated with an asymptotic power law decrease of the autocovariance function, that should be compared to the exponential one encountered in more classical models (like ARMA processes). An exponential behavior implies, by definition, a characteristic time while a power law, in contrast, is naturally scale invariant.

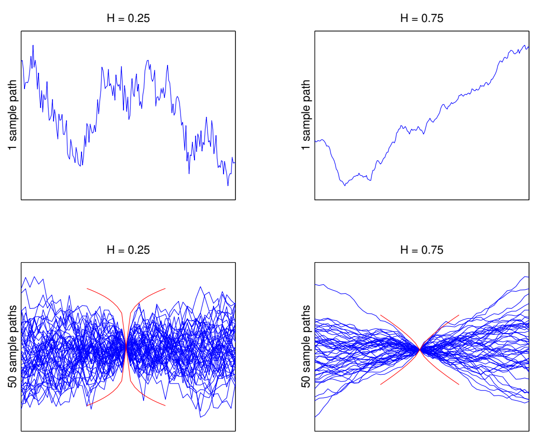


Fig. 5. **Hurst and Hölder in fractional Brownian motion.** The larger the Hurst exponent H, the smoother the sample path (top row). The Hölder characterization of roughness can be visualized by binding together a number of realizations at some arbitrary point, and by superimposing (in red) the right-hand side of eq.(10), with $h = H$ and $K = 3\sigma$ (bottom row).

All processes with exact self-similarity exhibit LRD. Indeed, let X be a H-sssi process with finite variance. Then it follows from equation (4) that, asymptotically, the covariance function of its increments Y_δ reads

$$r_{Y_\delta}(s) := \mathbb{E}Y_\delta(t+s)Y_\delta(t) \sim \sigma^2 H(2H-1)\, s^{2(H-1)}, \ s \gg \delta.$$

which shows that, for $1/2 < H < 1$, the increments are long-range dependent processes with $\gamma = 2H - 1$.

Long range dependence is often theoretically and practically studied through the technique of aggregation. As explained and illustrated in Box 2, aggregation consists of studying windowed average versions of the data as a function of the window width T. The covariance functions of the aggregated LRD processes converge to the form given in equation (4) for the fractional Gaussian noise (the increment process of fBm), which is itself invariant under aggregation. This explains its canonical role in analyzing long-range dependence in empirical time series. The variance of the aggregated LRD process also behaves as a power-law of the aggregation length with an exponent controlled by γ (Box 2). This property provides the basis for simple time domain estimators for the exponent (see, e.g., [34]). For traffic data, LRD models have been the most widely used. For example both the Ethernet data of Box 2 and the TCP data of Figure 1 exhibit strong LRD.

Fractal Processes. Local Hölder regularity (see Box 6) describes the regularity of sample paths of stochastic processes by means of a local comparison against a power-law function, and is therefore closely related to scaling in the limit of small scales [9]. The exponent of this power-law, $h(t)$, is called the (local) Hölder exponent and depends typically on both time and the sample path of X. Processes for which the Hölder exponent $h(t)$ is the same for all t, such as fractional Brownian motion, exhibit constant regularity along their sample paths; they are often referred to as *monofractal* processes. The Hölder exponent $h(t)$ provides a measure of local path-regularity or roughness which generalizes the notion of differentiability: sample paths exhibit more and more variability as h is decreased from 1 to 0. This is clearly seen for fractional Brownian motion in the top row of Figure 5.

While a rigorous proof is hard, it is easy to convince oneself of the monofractal character of fractional Brownian motion exploiting its H-sssi property combined with the centered nature of the Gaussian marginals. Indeed, from equation (6) the autocovariance of the increments Y_δ of a second order H-sssi process X behaves as

$$\mathbb{E}|Y(\delta,t)|^2 = \mathbb{E}|X(t+\delta)-X(t)|^2 = \sigma^2|\delta|^{2H}.$$

which is independent of t. In Box 6 we find in equation (9) an asymptotically equivalent property for some stationary processes with a certain autocorrelation function. Let us add the assumption that our process X is Gaussian, i.e., restrict X to fractional Brownian motion. Since the Gaussian distribution is well centered, meaning that most samples are within a few standard deviations from the mean, the net result is that the oscillations of X over intervals of length δ are roughly of the size $\delta^H \cdot \sqrt{\mathbb{E}[X^2(1)]}$. Indeed, it can be shown that for any $h < H$ (and for no $h > H$) almost all sample paths satisfy (10) at each t_0. Thus, the variability (oscillations) of fBm are of equal strength everywhere, confirming its monofractal character which it is entirely controlled by H. Another heuristic argument uses self-similarity to re-scale time and space through $X(ct) = c^H X(t)$ (see (1)) with the same ratio between time and space at "all" times. Similar as for long-range dependence, also local Hölder regularity is often studied through the increments of the process, according to relation (9).

Moving beyond monofractality, one could think of allowing the exponent h in relation (9) to be a function of time:

$$\mathbb{E}|X(t+\delta)-X(t)|^2 \sim C(t)\,|\delta|^{2h(t)}.$$

Such a process could describe data which have locally fractal properties which evolve slowly and fairly smoothly over time. If $0 < h(t) < 1$ is a deterministic function with enough regularity, the process X is said to be *multifractional* or, when Gaussian, *locally self-similar*. This means that locally around time t, $X(t)$ is very much like a fBm with parameter $H = h(t)$ (see [25] for details). Such a multifract*ion*al model clearly no longer has stationary increments, since their distributions depend by definition on the deterministically changing $h(t)$. Also, such a model is not multifractal in the true sense: although locally fractal with a varying exponent $h(t)$, it suffers from two deficiencies. First, the local irregularity $h(t)$ at a given time t is "deterministic", meaning that it is the same for almost all realizations, whereas it is random for truly multifractal processes. Second, $h(t)$ varies very slowly or "smoothly" while true multifractal processes exhibit a full range of different values $h(t)$ in *any* time interval, however small. For these two reasons, multifractional models really aim at describing a form of non-stationarity. Network traffic, however, can exhibit rich, true *multifractal* behavior (see Figures 1 and 2).

Box 6: Local Hölder Regularity

Let $\{X(t), t \in \mathcal{R}\}$ denote a second-order stationary stochastic process, whose autocovariance function has the cusp-like behavior $\mathbb{E}X(t)X(t+\delta) \sim (\sigma^2/2C)(1 - C|\delta|^{2h})$ (with $h > 0$) when $\delta \to 0$. This implies that small step increments of X satisfy:

$$\mathbb{E}|X(t+\delta)-X(t)|^2 \sim \sigma^2|\delta|^{2h}, \delta \to 0. \tag{9}$$

This relation gives an information on the regularity of X since the condition $h > 0$ guarantees mean-square continuity, whereas differentiability can only be achieved if $h > 1$. In other words, within the range $0 < h < 1$, sample paths of X are *everywhere continuous and nowhere differentiable.*

The description of such "wild" trajectories can be made more precise by referring to *Hölder exponents.* A signal $X(t)$ is said to be of Hölder regularity $h \geq 0$ in t_0 if one can find a local polynomial $P_{t_0}(t)$ of degree $n = \lfloor h \rfloor$ and a constant $K > 0$ such that $|X(t) - P_{t_0}(t)| \leq K\,|t-t_0|^h$. In the case where $0 \leq h < 1$, the regular part of $X(t)$ reduces to $P_{t_0}(t) = X(t_0)$, leading to the simpler relation, based on increments only:

$$|X(t_0+\delta) - X(t_0)| \leq K\,|\delta|^h, \tag{10}$$

and the largest such value of h is the Hölder exponent.

Hölder regularity is also closely connected to the algebraic behavior (9) of the increments variance, and even in the case of non-stationary processes, provided they have stationary increments. Stochastic processes that present a local Hölder regularity that is constant along their sample paths are often referred to as *monofractal* processes. More sophisticated situations can be encountered, where the Hölder exponent is no longer unique, but can vary from point to point. This is especially the case in *multifractal* situations (see Box 7).

Multifractals. When the regularity $h(t)$ is itself a highly irregular function of t, possibly even a

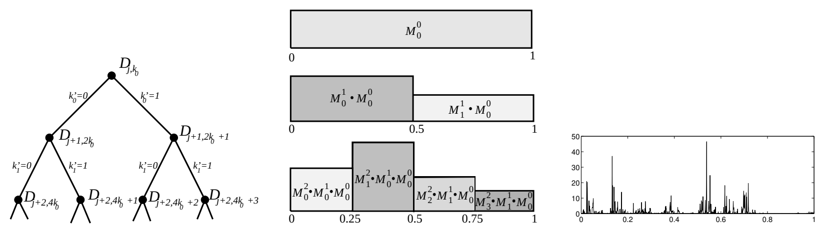


Fig. 6. **Binomial Cascade.** Dyadic tree based construction – first three intermediate stages (values of the measure on coarsest intervals) – a sample path.

random process rather than a constant or a fixed deterministic function, the process X is said to be *multifractal*. In such situations, the fluctuations in regularity along paths are no longer described in terms of a function $h(t)$ but through the so-called *multifractal spectrum* $D(h)$ (see Box 7 and [9], [29]). Tele-traffic time series, for example those in Figures 1, 2, in fact often have local Hölder exponents $h(t)$ which change erratically with location t. Such behavior is loosely termed *multifractal*. A model class which is rich enough to capture multifractal properties is that of *multiplicative cascades*. One of the most celebrated examples is that of the *Binomial cascade* X, defined here for convenience on $[0,1]$ through:

$$X((2k+1)/2^{n+1}) - X(2k/2^{n+1}) \stackrel{d}{=} M_{2k}^{n+1} \cdot (X((k+1)/2^n) - X(k/2^n)) \stackrel{d}{=} \prod_{i=1}^{n+1} M_{k_i}^i \cdot (X(1) - X(0)). \qquad (11)$$

Here the $M_{k_i}^i$ are independent positive random variables called the *multipliers* such that "siblings" add up to one: $M_{2k}^{n+1} + M_{2k+1}^{n+1} = 1$. Thus, (11) "repartitions" the increments of X iteratively. Setting $X(0) = 0$ and $X(1) = 1$) (for convenience) defines the process on $[0,1]$. This is a particular incarnation of a general approach to the generation of multifractal processes, namely the iteration of a multiplicative procedure. Note that all increments are positive and that the aspect ratios, given by the $M_{k_i,i}$, depend explicitly on the location where the re-scaling is done. This is in stark contrast to the scaling of fractional Brownian motion and the relation (5) for self-similarity, and is the most immediate reason for the multifractal structure of cascades. An illustration of this construction procedure as well as an example of resulting sample path is shown on Figure (6). Comparing by eye with the network time series of Figures (1, 2), a clear visual agreement is evident. A disadvantage of binomial cascades is that they are not even second order stationary. Stationary multifractal models are only just appearing in the literature [20].

One of the major consequences of multifractality in processes lies in the fact that quantities usually called *partition functions* present power law behaviors in the limit of small scales:

Box 7: Multifractals

Let $\{X(t), t \in \mathcal{R}\}$ denote a stochastic process. The *local Hölder exponent* $h(s)$ of the process at time s is a random variable defined pathwise as the largest $h > 0$ such that $|X(t) - P_s(t)| \leq K\,|t-s|^h$. Here, $P_s(t)$ is the local polynomial of degree $n = \lfloor h \rfloor$ as in Box 6. If the Taylor polynomial of degree n exists, then this polynomial is necessarily that Taylor polynomial; but in general the path of X might not have n derivatives.
In the case where the local polynomial P_s is constant then $h(s)$ is the largest h such that

$$|X(s+\delta) - X(s)| \leq K\,|\delta|^h. \qquad (12)$$

holds. Note that $h(s)$ may very well be larger than 1, as is the case with all cascades. A simple argument yields [31] the more useful dual statement: *if* the largest h satisfying (12) is non-integer, *then* the local polynomial P_s is necessarily constant and $h(s)$ can be computed using (12).
Fig 5 demonstrates the simple scaling structure of fractional Brownian motion; for almost every path and at any time instance one finds the same local scaling exponent: $h(t) = H$. In real world data such as network traffic the local scaling $h(t)$ changes erratically and randomly in time. The multifractal spectrum D of a process X provides a mean to capture this complexity; it is defined path-wise and is, thus, random. Denoting the *Hausdorff dimension* of a set E by $\dim(E)$ the spectrum is

$$D(a) := \dim(\{t \in \mathcal{R} \,:\, h(t) = a\}) \qquad (13)$$

The multifractal spectrum of cascades and self-similar processes is the same for almost all paths. In particular, for fBm it consists of only one point: $D(H) = 1$, while it has an inverted-"U" shape for multiplicative cascades.
While estimating D from traces is very hard, there exist almost sure upper bounds which are easier to estimate (see Box 11). For an overview see [31].

$$S_\delta(q) = \sum_{k=1}^{1/\delta} |Y((k+1)\delta, \delta))|^q = \sum_{k=1}^{1/\delta} |X((k+1)\delta) - X(k\delta)|^q \simeq c_q |\delta|^{\zeta(q)-1}, \;\; |\delta| \to 0. \qquad (14)$$

For instance, for the binomial cascade above, assuming that all multipliers in (11) are identically distributed, (14) holds (and also (19) below), at least for lags $\delta = 1/2^n$ and with $\zeta(q) = -\log \mathbb{E} M^q$. For processes with stationary increments, the time averages $S_\delta(q)/\delta$ can be seen as estimators for the statistical averages $\mathbb{E}|X_\delta(t)|^q$. Therefore, relation (14) above is highly reminiscent of the fundamental equation (2) implied by self-similarity. A major difference, however, lies in the fact that the exponent $\zeta(q)$ need not a priori follow the linear behavior qH of self-similarity. In other words, to describe cascades using one single exponent or parameter is impossible and an entire collection of exponents is needed. The measurement of the $\zeta(q)$ exponents offer, through a Legendre transform, a useful way to estimate the multifractal spectrum (see Box 11 and [31]).

Beyond power laws. Multifractal scaling offers an extension to self-similarity insofar as the scaling of moments is no longer governed by one single exponent H but by a collection of exponents. However, it maintains a key feature: moments behave as power laws of the scales. When analyzing actual data, it may very well be observed that this is not the case, see e.g., [39]. To account for those situations, the infinitely divisible cascade (IDC) model provides an extra degree of freedom.

The concept of infinitely divisible cascades (IDC) was first introduced by B. Castaing in [6] and rephrased in the wavelet framework in [4]. Box 8 briefly recalls its definition, consequences and relations to other models. The central and defining quantity of an IDC is the *propagator* or kernel $G_{\delta,\delta'}$. Infinite divisibility generalizes the concept of self-similarity; it simply says that the marginal distributions at different scales are related to each other through a simple convolution with the propagator G; thus, G completely captures and controls the multiscale statistics. Leaving details to Box 8, let us be explicit in the case of self-similarity where the propagator takes a particular simple form due to (1): $G_{\delta,\delta'}$ is a Dirac function. In more precise terms, the distribution at scale δ' is obtained by convolving the distribution at scale δ with $G_{\delta,\delta'}(\ln\alpha) = \delta(\ln\alpha - H\ln(\delta/\delta'))$. Since the Laplace transform reads as $\tilde{G}_{\delta,\delta'}(q) = \exp\{qH\ln(\delta/\delta')\}$ we may interpret $G_{\delta,\delta'}$ as the $\ln(\delta/\delta')$-fold self-convolution of an elementary propagator G_0 which describes a "unit change of scale". For comparison, we note

Box 8: Infinitely Divisible Cascades

Self-similarity implies that the probability density function (pdf) p_δ of the increments X_δ at scale δ, is a dilated version of the pdf of those at a larger scale δ': $p_\delta(x) = (1/\alpha_0)\, p_{\delta'}(x/\alpha_0)$ where the dilation factor is unique : $\alpha_0 = (\delta/\delta')^H$. In the cascade model, the key ingredient is that there is no longer a unique factor but a collection of dilation factors α ; consequently p_δ will result from a weighted sum of dilated incarnations of $p_{\delta'}$:

$$p_\delta(x) = \int G_{\delta,\delta'}(\ln\alpha)\, \frac{1}{\alpha}\, p_{\delta'}\left(\frac{x}{\alpha}\right) d\ln\alpha.$$

The function $G_{\delta,\delta'}$ is called the kernel or the ***propagator*** of the cascade. A change of variable shows that the definition above relates the pdfs $\underline{p}_\delta$ and $\underline{p}_{\delta'}$ of the log-increments $\ln|X_\delta|$ at different scales through a convolution with the propagator :

$$\begin{aligned} \underline{p}_\delta(\ln|x|) &= \int G_{\delta,\delta'}(\ln\alpha)\, \underline{p}_{\delta'}(\ln|x| - \ln\alpha)\, d\ln\alpha \\ &= (G_{\delta,\delta'} * \underline{p}_{\delta'})(\ln\alpha). \end{aligned} \tag{15}$$

Infinite divisibility implies by definition that no scale between δ and δ' plays any specific role, i.e, if scale δ'' lies between scales δ and δ' then $G_{\delta,\delta'} = G_{\delta,\delta''} * G_{\delta'',\delta'}$. This convolutive property implies that propagators can be written in terms of an elementary function G_0 convolved with itself a number of times, where that number depends on δ and δ'

$$G_{\delta,\delta'}(\ln\alpha) = [G_0(\ln\alpha)]^{*\left(n(\delta)-n(\delta')\right)}.$$

Here, G^{*n} denotes n fold convolution of G with itself.
Using the Laplace transform $\tilde{G}_{\delta,\delta'}(q)$ of $G_{\delta,\delta'}$, this can be rewritten as $\tilde{G}_{\delta,\delta'}(q) = \exp\{H(q)(n(\delta) - n(\delta'))\}$, with $H(q) = \ln\tilde{G}_0(q)$. This yields (compare with eq. (20)): the following relations, fundamental for the analysis [39]:

$$\ln\mathbb{E}|X_\delta|^q = H(q)n(\delta) + K_q \tag{16}$$

$$\ln\mathbb{E}|X_\delta|^q = \frac{H(q)}{H(p)}\ln\mathbb{E}|X_\delta|^p + \kappa_{q,p}. \tag{17}$$

A possible interpretation of this relation is that the function G_0 defines the elementary step of the cascade whereas the quantity $n(\delta) - n(\delta')$ quantifies the number of times this elementary step is to be applied to proceed from scales δ to δ'. The derivative of n with respect to δ describes in some sense the speed of the cascade at scale δ. When the function n takes the specific form $n(\delta) = \ln\delta$, the infinitely divisible cascade is said to be *scale invariant* and reduces to multifractal scaling. The exponents $\zeta(q)$ associated to the multifractal spectrum are then related to the Laplace transform of the propagator through $\xi(q) = H(q)$ (see Box 8). As detailed in the text, self-similarity is also included as an even more special case. For further details on infinitely divisible cascade, see [39].

$$\begin{aligned} \textit{Self-Similarity} \quad \mathbb{E}|X_\delta(t)|^q &= c_q|\delta|^{qH} = c_q\exp(qH\ln\delta) & (18)\\ \textit{Multifractal Scaling} \quad \mathbb{E}|X_\delta(t)|^q &= c_q|\delta|^{\zeta(q)} = c_q\exp(\zeta(q)\ln\delta) & (19)\\ \textit{Infinitely Divisible Cascade} \quad \mathbb{E}|X_\delta(t)|^q &= c_q\exp(H(q)n(\delta)) & (20) \end{aligned}$$

where the function $n(\delta)$ is not necessarily $\ln\delta$, just as the function $H(q)$ is not a priori qH.

III. Wavelets for Analysis and Inference

We saw from the previous section that diverse signatures of scaling can be observed both with respect to *time* (regularity of sample paths, slow decay of correlation functions,...), or to *frequency/scale* (power-law spectrum, aggregation, zooming, small scale increments,...). This suggests that to identify and characterize scaling an approach which *combines* time and frequency/scale, and which formalizes properly the idea of a simultaneous analysis at a continuum of scales, should be taken. In this respect, *wavelet analysis* appears as the most natural framework.

By definition, wavelet analysis (see Box 9 for basics and [18] for a comprehensive survey) acts as a mathematical microscope which allows one to zoom in on fine structures of a signal or, alternatively, to reveal large scale structures by zooming out. Therefore, when a signal or a process obeys some form of scale invariance, some self-reproducing property under dilation, wavelets are naturally able to reveal it by a corresponding self-reproducing property across scales. Moreover, the time-dependence of the wavelet transform allows for a time-localization of scaling features.

In its discrete version operating on dyadic scales, the wavelet transform (WT) is a rigorous and invertible way of performing a *multiresolution* analysis, a splitting of a signal into a low-pass *approximation* and a high-pass *detail*, at any level of resolution. Iterating the procedure, one arrives at a representation which consists of a low-resolution approximation, and a collection of details of higher and higher resolution. From the perspective of more classical methods used for scaling data, iterating low-pass approximations, at coarser and coarser resolutions, is an implicit way of *aggregating* data, whereas evaluating high-pass details, as differences between approximations, is nothing but a refined way of computing *increments* (of order N for a wavelet

Box 9: A Wavelet Primer

In contrast to the Fourier transform which analyzes signals in terms of oscillating sinusoidal waves $e^{j2\pi ft}$, the wavelet transform conducts a *local Fourier analysis* by projecting the signal $X(t)$ onto locally oscillating waveforms, referred to as "wavelets." A *wavelet* $\psi(t)$ is a bandpass function which oscillates with some central frequency f_0. Scaling (by dilating or compressing) and shifting the wavelet:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k), \tag{21}$$

moves its central frequency to $2^{-j} f_0$, and shifts its time center by $2^j k$.

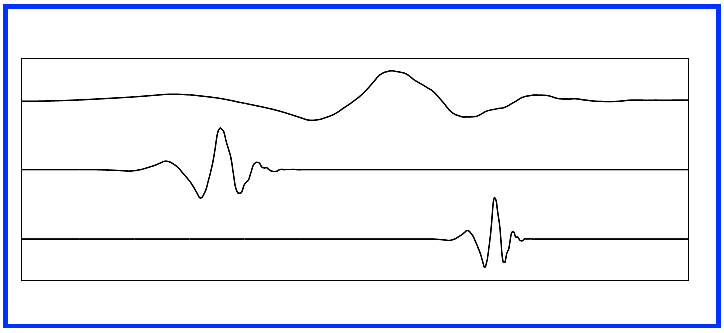


Fig. 7. Wavelets from a length-8 Daubechies filterbank. ¿From top to bottom: $\psi_{0,0}(t)$, $\psi_{1,3}(t)$, $\psi_{3,22}(t)$.

Besides the wavelet $\psi(t)$, a wavelet decomposition makes use of a companion low-pass function $\phi(t)$ (referred to as a *scaling function*) which can be scaled and shifted in the same way. Just as a signal can be built up from a sum of weighted sinusoids, it can be built up from a sum of weighted scaling functions and wavelets

$$X(t) = \sum_k c_X(j_0,k)\phi_{j_0,k} + \sum_{j\le j_0}\sum_k d_X(j,k)\psi_{j,k}(t). \tag{22}$$

The $c_X(j_0,k)$ are called the *scaling coefficients*; and the $d_X(j,k)$ the *wavelet coefficients*. The first term reconstructs a coarse-resolution approximation to $X(t)$. The second term adds in detail information at finer and finer scales (higher and higher frequencies) as $j \to -\infty$. By careful design, the wavelet and scaling functions can be constructed to be orthogonal, meaning we can compute the wavelet and scaling coefficients as simple inner products:

$$c_X(j,k) = \langle X, \phi_{j,k}\rangle, \quad d_X(j,k) = \langle X, \psi_{j,k}\rangle. \tag{23}$$

As an extension to the band-pass requirement (i.e., ψ has zero mean), a further property of any wavelet is its *number of vanishing moments*, i.e., the largest number $N \ge 1$ such that

$$\int t^k\, \psi(t)\, dt = 0;\;\; k = 0, 1, \ldots N-1. \tag{24}$$

There are large families of orthogonal wavelets and scaling functions. The Daubechies–8 wavelets pictured above (for which $N = 4$) are but one example.

From a practical point of view, the scaling and wavelet coefficients are related by a *filterbank*. To create $c_X(j,k)$, $d_X(j,k)$, we pass $c_X(j+1,k)$ at the next finer scale through both a lowpass and a highpass discrete-time filter and then downsample by skipping every other sample. The filter responses are elegantly related to the continuous-time scaling and wavelet functions. This algorithm is applicable also to discrete-time signals and is extremely efficient ($O(n)$ time to compute all available scales of a n point signal).

with N vanishing moments). Combining these two key elements makes of multiresolution a natural language for scaling processes.

As explained in Section II, self-similarity is the canonical reference model for scaling behavior. Self-similar processes with stationary increments are traditionally analyzed through their increments, however reasons for resorting to wavelets are at least threefold:

1 *Scaling* — Due to its built-in scaling structure, the wavelet transform reproduces any scaling present in the data, with a geometrical progression of all (existing) moments across scales, as:

$$\mathbb{E}|d_X(j,k)|^q = \mathbb{E}|d_X(0,k)|^q \cdot 2^{jq(H+1/2)}. \tag{25}$$

2 *Stationarization* — Due to the bandpass nature of admissible wavelets, sequences of wavelet coefficients can be seen as (filtered) increment processes at different scales: this makes the analysis extensible to non-stationary processes with stationary increments (like H-sssi processes), resulting in stationary sequences at each scale.

3 *Almost decorrelation* — Whereas direct manipulation of LRD processes is hampered by slowly-decaying correlations, it turns out that [11], [37]

$$\mathbb{E}d_X(j,k)d_X(j,k+m) \sim C(j)|m|^{2H-2N}, \quad |m| \to \infty,$$

N being the number of vanishing moments of the wavelet. Under the mild condition $N \geq H+1/2$, global LRD existing among the increments of H-sssi processes, can thus be turned, at each scale, into short-range dependence.

Another advantage is that, due to the frequency interpretation of wavelets, wavelet analysis can serve as a basis for useful substitutes for spectral analysis. Indeed, it can be shown that for stationary processes X with power spectrum $\Gamma_X(\nu)$, we have

$$\mathbb{E}d_X(j,k)^2 = \int \Gamma_X(\nu)\, 2^j\, |\Psi(2^j\nu)|^2\, d\nu.$$

When in addition X is a long range dependent process, this yields

$$\mathbb{E}|d_X(j,k)|^2 \sim C' 2^{j\alpha},\ j \to +\infty, \tag{28}$$

and it can be shown [2] that the corresponding wavelet coefficients are also short range dependent as soon as $N \geq \alpha/2$.

Wavelet coefficients are also useful to study Hölder regularity. This relies on the fact that if X is Hölder continuous of degree $h(t)$ at t then the wavelet coefficients at t decay as

$$|d_X(j,k)| \leq 2^{j(h(t)+1/2)} \tag{29}$$

as the intervals $[k2^j,(k+1)2^j]$ close in on t ($j \to -\infty$). Under certain conditions, the bound is asymptotically tight [13], [7]. For monofractal processes, that is for processes for which Hölder exponents $h(t)$ remain constant along sample paths, we have the following relation,

$$\mathbb{E}|d_X(j,k)|^2 \sim C'' 2^{j(2h+1)},\ j \to -\infty, \tag{30}$$

to be compared to equations (25) and (28) above.

To summarize, the wavelet transform closely reproduces the scaling properties that exist in data, be it self-similarity, long range dependence, or monofractality, and, at the same time, replaces one single poorly behaved (non-stationary, LRD) time series by a collection of much better behaved sequences (stationary, SRD), amenable to standard statistical tools. Therefore, second order statistical scaling properties can be efficiently estimated from marginalized scalograms, that is squared wavelet coefficients averaged over time, circumventing the difficulties usually attached to scaling processes.

Box 10: Wavelet Analysis of 2nd Order Scaling

Scaling processes (be they LRD, $1/f$-type, mono- or multifractal) share the property of exhibiting power-law spectra in some frequency range, whence the idea of estimating scaling exponents from a spectral estimation. The wavelet transform offers an alternative to classical spectrum analysis [2], based on a power law behavior of the wavelet detail variances across scales

$$\mathbb{E}|d_X(j,k)|^2 \sim C 2^{j\gamma}, \tag{26}$$

reminiscent of equation (25) with $q=2$ for self-similarity, (28) for long range dependence and (30) for monofractality. These are all suggestive of a *linear* relationship $\log_2 \mathbb{E}d_X(j,k)^2 \sim \gamma j + C$ in a log-log plot.
The *stationarization* property together with the *almost decorrelation* property (see points 2 and 3 in text) justify that the variance involved in (26) can be efficiently estimated on the basis of the simple empirical estimate:

$$\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} d_X(j,k)^2, \tag{27}$$

where n_j is the number of coefficients available at octave j. The graph of $\log_2 \mu_j$ against j (together with proper confidence intervals) is referred to as the (second-order) *Logscale Diagram* (LD) [3]. Examples are given in Figure 8. Straight lines in such diagrams can be understood as evidence for the existence of scaling in analyzed data, while the range of scales involved gives information on its precise nature (self-similarity, long memory, ...). Estimation of scaling exponents can be carried out from such graphs via weighted linear-fit techniques (see [3], [38], [1] for details). The possibility of varying the number of vanishing moments of the mother wavelet bring robustness to the analysis procedure against non-stationarities.

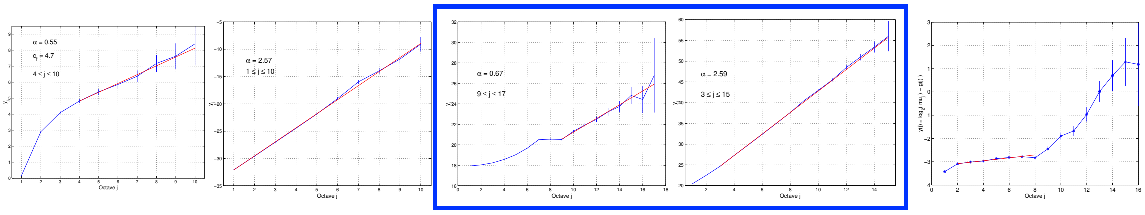
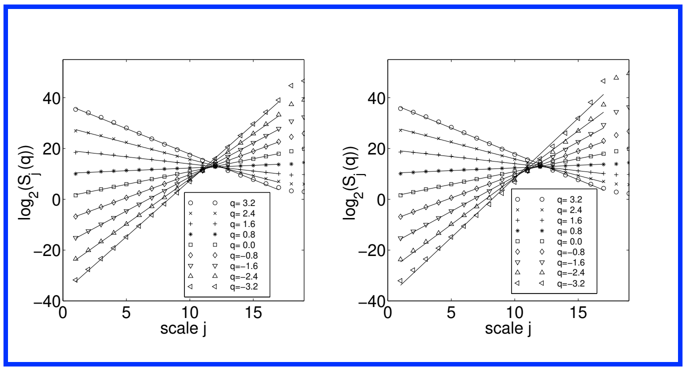


Fig. 8. **Second Order Logscale Diagrams.** For each of five different time series, scaling behavior is identified over the range fitted in red, as described in Box 10. Left two plots: a LRD series with scaling at large scales, and a self-similar process, where the scaling is seen across all scales. The next two plots are from the same "pAug" Ethernet trace as Box 2. Left: discrete time series of IP packet inter-arrival times showing LRD, and Right: the bytes per bin data of Box 2, showing empirical self-similarity. Far right plot: Interarrival time series of TCP connections (see Figure 1), showing an abrupt change point separating two apparently different scaling behaviors, at a characteristic time scale of about 1s. These two scaling regimes can be linked via the Infinitely Divisible Cascade model.

Using this idea, Box 10 details the steps leading to an estimation of the exponent of second order scaling, in a log-log plot known as the Logscale Diagram.

Examples of such second order analysis are given in Figure 8 for two synthesized time series and three series from traffic data, as detailed further in the caption. The plots grouped in the box are two different time series extracted from the same celebrated Ethernet trace [14] discussed in Box 2. Series from this trace provided one of the first clear indications of long range dependence in traffic. The advent of wavelet-based analysis added precision and completeness to the study of the empirical scaling, and to the corresponding measurements of the Hurst parameter [3], [38], as well as estimates of the prefactor C' (equation 28), of importance in applications. Crucially, it also helped settle controversy as to the interpretation of the discovery, by showing that the observed scaling in the time series was not the result of corrupting non-stationarities, but actually corresponded to long range dependencies.

The diversity of behavior in the examples of Figure 8 illustrates an important advantage of a semi-parametric analysis framework, such as the wavelet approach described here. The analysis need not make any a priori assumption about the range of scales over which scaling may exist. The range is rather inferred from the analysis itself, leading to an identification of the scaling type, such as LRD at large scales and/or multifractality at small scales, prior to any estimation phase. Indeed, the rightmost plot shows two different scaling regimes for a series derived from Internet data, which (from a purely second order viewpoint), requires two independent estimations. In contrast, parametric methods can easily give very misleading results if the data is not close to the assumed model class, making them unsuitable for the exploration of real, and complex, data. The comments of the previous paragraph could be expressed as "robustness with respect to model class". Another form of robustness enjoyed by wavelets is their insensitivity to deterministic trends which may be superimposed onto a process of interest, with undesirable consequences. These include invalidating the stationarity property of the LRD process under study, or mimicking LRD correlations when added to a short-range dependent process [1]. Wavelets are a versatile solution to this crucial issue, since they offer the possibility of being blind to polynomial trends. Recall that any admissible wavelet has zero mean. This is equivalent to having a zeroth order vanishing moment, or in other words, to be orthogonal to constants. In fact N vanishing moments implies that the wavelet is blind to polynomials up to order $p \leq N-1$. Trends which are "close" to polynomial can be effectively eliminated in this manner [3], and the advantage of being able to do so without even testing for their presence is an important one when making sense of real data, and in particular when trying the distinguish non-stationarity from scaling behavior. Building on the advantages of the wavelet approach, a statistical test for the constancy of a scaling exponent can be defined [40] which helps resolve this difficult issue.



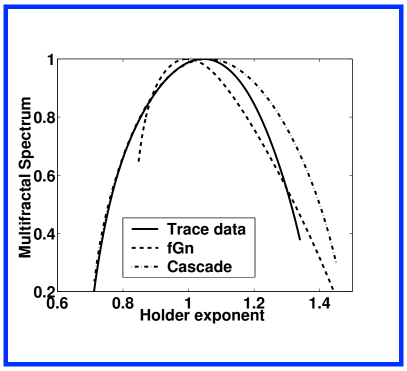


Fig. 9. Left: Superimposed log-log plots, at several values of q, of the partition sum against scale for a time series of bytes per bin of TCP traffic (taken from the LBL-TCP3 trace [24]) (on left), and a matched binomial cascade (right).

Fig. 10. Right: Multifractal spectrum of local Hölder exponents estimated via the Legendre transform.

Finally, the analysis of scaling processes is often faced, and particularly so in the case of tele-traffic, with enormous quantities of data, thereby requiring methods which are efficient from a computational point of view. Because of their multiresolution structure and the related ability to be implemented as a filter bank, wavelet-based methods are associated with fast algorithms, out performing FFT-based competitors with a complexity of only $O(n)$ in computation (compared to $O(n \log(n))$) and $O(\log(n))$ in memory, for n data points. These advantages hold not only at second order, but more generally, including for the more advanced types of analysis we now discuss.

Beyond Second Order Analysis As explained in Section 2, scaling may involve statistics beyond second order, which if observed in the limit of small scales, calls for a multifractal interpretation. Multifractal analysis provides a "finger print" of local scaling properties of the paths of a process X through the multifractal spectrum $D(h)$, and the *multifractal formalism* provides a powerful approach to numerically estimating it. Just as for second order scaling analysis, estimates can be based on increments of the process or time series, however, from arguments close to those developed at second order, wavelet coefficients offer themselves as an ideal alternative. Notably, tuning the number of vanishing moments of the mother wavelet allows the analysis of processes with Hölder exponents larger than 1. Box 11 gives a more detailed pictures of this wavelet based multifractal analysis.

Box 11: Wavelet-based Multifractal Formalism

The wavelet based partition function,

$$S_j(q) = \sum_k |2^{-j/2} d_X(j,k)|^q, \tag{31}$$

constitutes the wavelet counterpart of the traditional partition function (equation (14)). It can be bounded from below by summing only over a subset of indices k, say those for which $|2^{-j/2} d_X(j,k)| \sim 2^{jh}$. For the sake of argument we assume that this marks the locations where the Hölder regularity of the path is indeed h (compare (29)). It follows then from box-counting methods, a standard technique in fractal geometry, that the number of such indices grows asymptotically at least as $2^{-jD(h)}$, implying that $S_j(q)$ grows at least as $2^{j(qh-D(h))}$. Since the choice of h was arbitrary, we arrive at the asymptotic bound

$$S_j(q) \geq 2^{\inf_h (j(qh-D(h)))}, \tag{32}$$

which is provably tight in the limit $2^j \to 0$ using a steepest descent argument. Estimating $\zeta(q)$ from the decay of estimates of the moments $S_j(q) \sim 2^{j\zeta(q)}$, we arrive at an asymptotic estimate

$$D(h) \leq D^{**}(h) = \zeta^*(h), \tag{33}$$

where $g^*(x) = \inf_y (xy - g(y))$ denotes the Legendre transform of a function g. Note that applying the transform twice yields the concave hull g^{**} of g. It is notable, that the statistically and numerically robust global estimator ζ provides information on the delicate local properties captured in $D(h)$, which would be almost impossible to access directly.

In practice, $\zeta(q)$ is estimated as the least square slope of a log-log plot of the partition sum against scale, i.e., $\log(S_j(q))$ against $\log 2^j$. Comparing with Box 10, this demonstrates quite explicitly how multifractal analysis goes beyond second order statistics. Figure 9 shows examples. This wavelet based estimator can be further developed using the wavelet maxima method [23], [4] which addresses in particular the invertibility of (29).

Figure 9 depicts log-log plots of $S_j(q)$ against 2^j for a real world trace (the LBL-TCP3 trace of [24]) and a synthetic cascade which has been designed to match the second moments of the series on all dyadic scales. It is notable that also the sample moments of orders $-3.2 \leq q \leq 3.2$ agree closely. Consequently, the functions $\zeta(q)$ and the estimated spectrum $D(h) = \zeta^*(h)$ are very close. This is demonstrated in Figure 10 where the spectrum of an additive tree model is added for comparison. This additive model matches the same second order moments as the cascade, but it is Gaussian in nature with only little variation in its local Hölder exponents and consequently shows a narrow spectrum different from the real trace. This example again shows that in numerous computer network time series, scaling occurring at small scales cannot be described by a single exponent but require an entire family. Current research focusses on its impact on performance evaluation, network design and control [8], [28].

The infinitely divisible cascade model, introduced in Box 8 using increments for simplicity, can also be rephrased in wavelet terms [4], [39] with, again, many advantages similar to those detailed above for the second order case. Box 12 illustrates the analysis, estimation, and verification procedure of this more practical wavelet incarnation. The time series is that of Figure 1, the list of successive inter-arrival times of TCP connections. The study of the nature of such a series gives us direct insight into the statistical genesis of TCP connections in a heterogeneous environment. The series was extracted from exceptionally precise TCP/IP trace made available by the WAND group at the University of Waikato. This archive, the "Auckland II" traces, are taken from both directions of the access link of the University of Auckland to the external Internet [22]. As detailed in Box 12, an infinitely divisible cascade model provides a relevant description of the analyzed time series on a wide range of scales: $2^3 \leq 2^j \leq 2^{14}$. The key observation is that no other scaling model could have been applied over the full range, because of the change in behavior at the change point at around j_*.

Box 12: Extracting an Infinitely Divisible Cascade

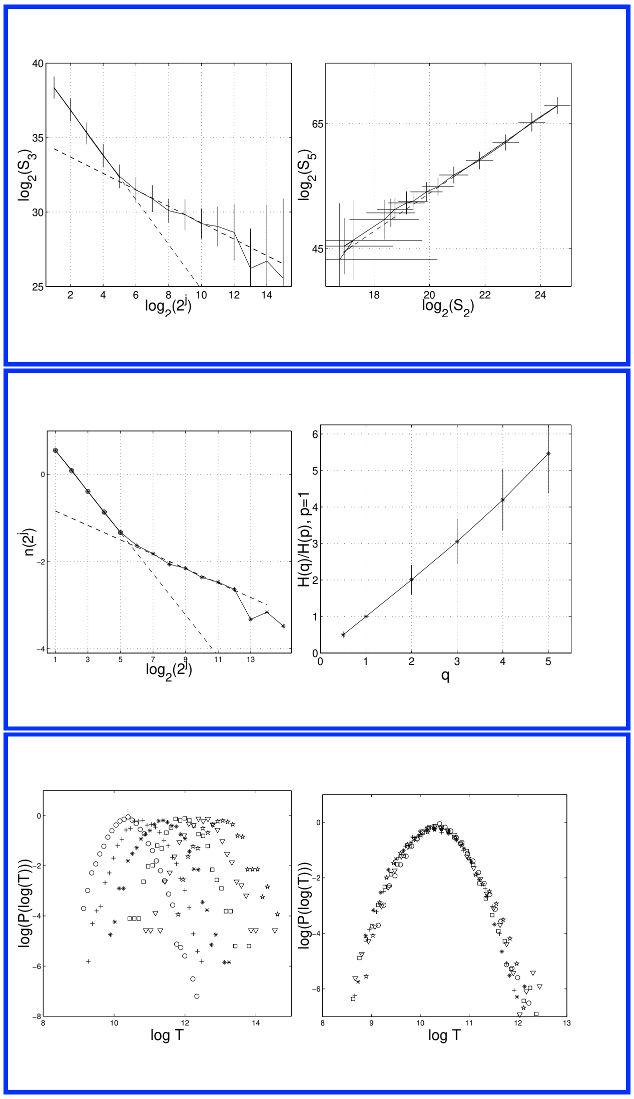


Fig. 11. **Scaling for a TCP/IP connection Inter Arrival time series.**

The IDC model is used here to analyze the TCP connection inter-arrival time series of Figure 1. The top left plot shows that the third order moments of the wavelet coefficients do not behave as power-laws of scale over the full range of scales, disallowing a self-similar or even a multifractal model over this range. The top right plot shows that **relative** power-laws do exist over the full range for the (for example, fifth order) moments, suggesting an Infinitely Divisible Cascade model can apply. Note the confidence intervals in both directions, as estimates are plotted on both axes. The middle plots show respectively the estimates of the functions $n(2^j)$ and $H(q)$, defining the IDC propagator. The bottom left plot shows the estimated probability density functions of the wavelet coefficients at scales 2^6 to 2^{11}. In the bottom right plot, those densities have been numerically "propagated" through the cascade, using the estimated propagator. The collapse of the curves illustrates the meaningfulness of the fitted Infinitely Divisible Cascade model, as well as the accuracy of the estimated propagator.

Here as in many other series extracted from Internet traces, j_* corresponds to a characteristic time of 2.5 to 3.5 seconds, in keeping with findings in [10], and of our own measurements of *round trip times* of TCP/IP connections. Indeed, when examining individual log-log plots such as the top left in Box 12 (or the far right in Figure 8), without the IDC framework one would be forced to conclude that two entirely different scaling models apply, over two different scaling ranges. Using the IDC formalism it is possible to note that the change is restricted to $n(2^j)$, the wavelet counterpart of the $n(\delta)$ function introduced in Box 8, whereas $H(q)$ is typically observed to be close to linear. We can therefore integrate the observations into a single scaling picture over the full range of scales, and interpret the piecewise-log form of $n(2^j)$ as an abrupt change of speed of some underlying multiplicative mechanism, described by $H(q)$, which is itself unchanged. Although "only statistical", such a specific hypothesis leads us to search for causal explanations, in traffic sources, networks themselves and their protocols, that could be capable of generating effects of this type. Using infinitely divisible cascades to model a variety of time series describing different aspects of the same raw Internet data, is a starting point for ongoing modelling work, some early results of which can be found in [39], [32].

IV. Selected applications of multiscale traffic models

A triumph of multiscale analysis techniques in networking has been the discovery of strong scaling phenomena as well as convincing evidence pointing to causes behind it: networking mechanisms, protocols, source characteristics and so on. But the multiscale concept is applicable to network related problems beyond the mere analysis of traffic traces. In this section, we briefly outline some applications that directly leverage the multiscale framework.

Multiscale Queuing Analysis Since the construction of network routers consists largely in combining queues (buffers), queuing analysis plays a crucial role in their design and performance. In the simplest queuing analysis, an aggregate traffic input $X(t)$ is fed into a single-server queue of size B bytes with service rate s bytes/s, and we wish to determine information about $Q(t)$, the queue size in bytes at time t. For example, we might desire the average queue size or the probability that the queue will overflow, the tail queue probability $P(Q > B)$. Queuing analysis in general is extremely difficult, owing to the inherent non-linearities associated with a queue emptying (few packet arrivals) and overflowing (too many packet arrivals).

A distinct advantage of the classical Poisson traffic model for $X(t)$ is the existence of analytic formulae for $P(Q > B)$ [17]. However, the fact that real traffic is not Poisson renders these results of limited utility in real-world situations.

Another, approximate approach is to study only the so-called *critical time scale* that dominates queue overflow.

But as we have seen, real traffic is not typically dominated in a simple way by a single time scale. Real traffic is multiscale, and so we should study the queue size $Q(t)$ at multiple time scales and fuse the results into a single statistic. A multiscale model for $X(t)$ (such as fBm or a binomial cascade) facilitates the investigation of the distribution of $Q(t)$ at multiple scales, incorporating the full multiscale structure. In this framework, the distributions of the wavelet coefficients of the fBm model, or multipliers in the cascade models, are combined into a simple formula that provides a close approximation to the tail queue probability. See [27] for more details.

Multiscale Path Probing To understand and predict the performance of end-to-end protocols such as TCP and modern streaming protocols, it is crucial to understand the dynamics of the end-to-end paths through a network. In particular, we could have interest in the delays and losses experienced by packets transmitted end-to-end. Here we focus on delay rather than loss.

Information on packet delay can be obtained either by actively probing the path with packets or by passively monitoring packets as they pass a fixed point. We will focus on an active strategy. The delay a packet will incur is bounded below by the propagation delay from the transmitter to receiver. However, it can be considerably larger if there is significant *cross-traffic* that forces the packet to wait in a buffer before it is serviced. Clearly, modelling the end-to-end packet delay process implicitly involves modelling the cross-traffic, since large delays are caused by large traffic flows along the path.

A typical Internet end-to-end path can easily pass through fifteen or more queues, which complicates analysis and modelling considerably. Fortunately, in certain cases, an end-to-end path can be replaced by single "bottleneck" queue that is driven both by the probe traffic and an "effective cross-traffic" stream that models the contributions of all competing traffic along the path. Our fundamental observation for this bottleneck queue model is as follows: the delay spread at the receiver between two probe packets transmitted closely spaced in time corresponds directly to the amount of cross-traffic along the path.

Inherent in any probing scheme is an uncertainty principle, or "accuracy/sparsity tradeoff." The volume of cross-traffic entering the bottleneck queue between the two probes can be computed essentially exactly from the delay spread of the two packets at the receiver *provided* the queue does not empty in between. Unfortunately, this emptying will certainly occur unless the probes are spaced very closely. Even worse, long probing trains of closely spaced packets will overwhelm the very network we are trying to model. If the probes are spaced far apart, then the queue can empty in between, which results in uncertainty in the cross-traffic measurement.

Again, help is on the way with a multiscale model. Modelling the cross-traffic as a multiscale process (fBm or binomial cascade for example), we can transmit a stream of packets that probes simultaneously at several time scales. For example, by spacing the packets exponentially (two packets with small spacing T followed by a packet every 2^kT, $k = 1, 2, \ldots$, we probe the bottleneck queue at a multitude of dyadic scales.

This so-called "chirp packet train" balances the accuracy/sparcity tradeoff by being highly accurate initially and highly sparse at the end [28]. Packet chirps allows us to estimate the cross-traffic volume (or equivalently delay distribution) at any dyadic scale of interest. The algorithm works quite well in simulation studies; currently it is under more exhaustive testing on real networks.

V. Conclusions

In this paper, we have seen that the complexity and richness of tele-traffic is well matched by the multiscale analysis and modelling frameworks of self-similarity, long-range dependence, fractals, multifractals, and infinitely divisible cascades. These frameworks not only allow us to confirm and formalise the presence of multiscale behavior in traffic, but also point to possible causes of multiscale structure in the physical networking infrastructure. The choice of framework, from a simple fBm to a more complicated multifractal or cascade, clearly depends on the application and the data at hand. But whatever the framework, the multiscale wavelet transform provides a parsimonious and efficient domain for processing.

Finally, we note that the tools overviewed here have found a home in numerous other areas of science and engineering, including turbulence and percolation, among many others[1].

Acknowledgements This work was supported by grants from USA DARPA, DOE, and NSF, the French CNRS, and by Ericcson.

[1] Matlab routines implementing the analysis/estimation procedures described throughout this text are available at the following URLs: `www.emulab.ee.mu.oz.au/~darryl` and `www.dsp.rice.edu/`

References

[1] P. Abry, P. Flandrin, M.S. Taqqu and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. Chapter 2, pp. 39–88, in [26].

[2] P. Abry, P. Gonçalvès and P. Flandrin. Wavelets, spectrum estimation and $1/f$ processes. in A. Antoniadis and G. Oppenheim, eds., *Wavelets and Statistics, Lecture Notes in Statistics*, **103**, pp. 15–30, Springer-Verlag, New York, 1995.

[3] P. Abry and D. Veitch. "Wavelet analysis of long-range dependent traffic", *IEEE Trans. on Info. Theory*, 44(1), pp. 2–15, 1998.

[4] A. Arnéodo, J.F. Muzy and S.G. Roux. "Experimental analyis of self-similar random cascade processes: application to fully developed turbulence", *J. Phys. II France*, vol. 7, pp. 363–370, 1997.

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

[6] B. Castaing. The temperature of turbulent flows. *J. Phys. II France*, 6:105–114, 1996.

[7] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, New York, 1992.

[8] A. Erramilli, O. Narayan, A. Neidhardt and I. Saniee, "Performance Impacts of Multi-Scaling in Wide Area TCP/IP Traffic", Proceedings of IEEE Infocom'2000, March 2000, Tel Aviv, Israel.

[9] K. Falconer. *Fractal Geometry—Mathematical Foundations and Applications.* Wiley, New York, 1990.

[10] A. Feldmann, A.C. Gilbert, P. Huang, W. Willinger, Dynamics of IP traffic, A study of the role of variability and the impact of control, Proceedings of *ACM/Sigcomm'99*, Cambridge MA, August 1999, vol. 29(4), pp. 301–313.

[11] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion", *IEEE Trans. on Info. Theory*, vol. 38, pp. 910–917, 1992.

[12] A.C. Gilbert, "Multiscale analysis and data networks", *Appl. Comp. Harm. Anal.*, 10(3), pp. 185–202, 2001.

[13] S. Jaffard. Local behavior of Riemann's function. *Contemporary Mathematics*, 189:287–307, 1995.

[14] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the self-similar nature of Ethernet traffic", *Computer Communications Review*, vol. 23, pp. 183–193, 1993.

[15] W.E. Leland and D.V. Wilson, "High time-resolution measurement and analysis of lan traffic: Implications for lan interconnection", in *Proceedings of IEEE Infocom'91*, Bal Harbour, FL, 1991, pp. 1360–1366.

[16] J. Lévy-Véhel and R. Riedi, Fractional Brownian motion and data traffic modeling: the other end of the spectrum, in J. Lévy-Véhel, E. Lutton and C. Tricot, eds., *Fractals in Engineering*, pp. 185–203, Springer, London, 1997.

[17] D. V. Lindley. The theory of queues with a single server. Proc. Cambridge Phil. Soc. vol. 48, pp. 277–289, 1952.

[18] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Boston, 1997.

[19] B.B. Mandelbrot and J.W. Van Ness, "Fractional Brownian motions, fractional noises and applications", *SIAM Rev.*, vol. 10, pp. 422–437, 1968.

[20] P. Mannersolo, I Norros and R. Riedi, "Multifractal products of stochastic processes: construction and some basic properties", COST (European Cooperation in the field of Scientific and Technical Research) 257, technical report COST257TD(99)31, 1999. Submitted to *Applied Probability*, August 2001.

[21] K. Meier-Hellstern, P.E. Wirth, Y.-L. Yan, and D.A. Hoeflin, "Traffic models for ISDN data users: office automation application", in *Proceedings of 13th ITC, pp. 167–172, Copenhagen*, 1991.

[22] Jörg Micheel, Ian Graham, N. Brownlee, *The Auckland data set: an access link observed.* to appear, Proceedings *14th ITC Specialist Seminar*, Barcelona, April 2000.

[23] J.F. Muzy, E. Bacry and A. Arnéodo, "The multifractal formalism revisited with wavelets", *Int. F. Bifur. Chaos*, 4(2), pp. 245–301, 1994.

[24] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, 1995.

[25] R.F. Peltier and J. Lévy-Véhel, Multifractional Brownian motion: definition and preliminary results, INRIA Research Report No. 2645, 1995.

[26] *Self-Similar Network Traffic and Performance Evaluation.* K. Park and W. Willinger, eds. Wiley, 2000.

[27] V. Ribeiro, R. Riedi, M. S. Crouse, R. Baraniuk. Multiscale queuing analysis of long-range-dependent network traffic. Proc. IEEE Infocomm, Mar. 2000 (available at dsp.rice.edu).

[28] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, R. Baraniuk. Multifractal cross-traffic estimation. Proc. ITC Specialist Seminar on IP Traffic Measurement, Modeling, and Management, Sept. 2000. pp. 15/1-10 (available at www.dsp.rice.edu).

[29] R. Riedi, "An improved multifractal formalism and self-similar measures", *J. Math. Anal. Appl.*, vol. 189, pp. 462–490, 1995.

[30] R. Riedi, M.S. Crouse, V.J. Ribeiro, R.G. Baraniuk, "A Multifractal Wavelet Model with Application to Network Traffic", *IEEE Trans. on Info. Theory*, Special Issue, 45(3):992–1018, April, 1999.

[31] R. Riedi, "Multifractal Processes", in: "Long range dependence : theory and applications", eds. Doukhan, Oppenheim and Taqqu, to appear 2001.

[32] S. Roux and D. Veitch and P. Abry and L. Huang and P. Flandrin and J. Micheel, "Statistical Scaling Analysis of TCP/IP Data", Proceedings of *ICASSP 2001, Special session, Network Inference and Traffic Modeling*, Salt Lake City, Utah, May, 2001.

[33] G. Samorodnitsky and M.S. Taqqu, *Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 1994.

[34] M.S. Taqqu, V. Teverovsky and W. Willinger, "Estimators for long-range dependence: an empirical study", *Fractals*, 3(4), pp. 785–798, 1995.

[35] M.S. Taqqu, V. Teverovsky and W. Willinger, "Is network traffic self-similar or multifractal?", *Fractals*, vol. 5, pp. 63–74, 1997.

[36] M.S. Taqqu, W. Willinger and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling", *Comput. Commun. Rev.*, vol. 26, pp. 5–23, 1997.

[37] A.H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion", *IEEE Trans. on Info. Theory*, vol. 38, pp. 904–909, 1992.

[38] D. Veitch and P. Abry, "A wavelet-based joint estimator for the parameters of long-range dependence", *IEEE Trans. on Info. Theory*, 45(3), pp. 878–897, 1999.

[39] D. Veitch, P. Abry, P. Flandrin and P. Chainais, Infinitely Divisible Cascade Analysis of Network Traffic Data. Proceedings of *ICASSP 2000*, Istanbul, June 2000.

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

Matlab routines implementing the analysis/estimation procedures described throughout this text are available at the following URLs: `www.emulab.ee.mu.oz.au/~darryl` and `www.ece.rice.edu/`.