Teletraffic analysis and practice was transformed by the discovery of scale invariance properties in packet traffic [1]. The presence of large-scale asymptotic scale invariance, or Long-Range Dependence (LRD), is remarkably universal, and has become indispensable part of traffic modelling, in particular for TCP/IP traffic in the Internet. It appears that this role is destined to continue, as the phenomena has a physical underpinning which is both generic and readily understandable in networking terms, namely the heavy tailed nature of file sizes [2], which, through a well known mechanism [3], results in heavy tailed flows and thereby LRD.

The story of scale invariance in traffic however is not without controversy, both during the initial period of discovery, immediately afterward as the implications for modelling and network performance were being worked out, and through the subsequent debate over the possible existence and origin of multifractal behaviour. In this short paper, which accompanies the talk given in the Stochastic Networks session of this ISI meeting, I will give an historical overview of the subject, explain the current view, and point to the key statistical questions.

II. THE REVOLUTION

In the early 1990’s, packet switched networks were beginning to become widespread, and the Internet was growing rapidly. Engineers from telecommunications carriers were in a position to collect traces of packet traffic, and began to notice some curious behaviour. Some of the first publications detailing these observations were those of Hellstern, Wirth, Yan, Hoeftin [4], where infinite moments where found in IDSN traffic, Leyland and Wilson [5], who observed bursts occurring on ‘all’ time scales, and Erramilli and Willinger [6], who realised that Ethernet data exhibited fractal-like properties. At this point, it was clear that something unexpected was happening, but it was not clear what. The indications of infinite moments were radical for the field, and it was not understood what they meant.

Shortly afterward however, it was realised that the ‘bursts on all time scales’ phenomenon was a form of scale invariance, well known to statisticians, physicists and others. In their award winning and very widely cited paper [1], Leyland, Taqqu, Willinger, and Wilson put the idea of self-similar traffic on the map, building on their earlier work on Ethernet data. This was soon reinforced by similar observations made in local area networks the world over, and then in CCSN/SS7 traffic in 1994, and video traffic (though less clearly) in 1995. Another very influential paper was that of Paxson and Floyd [7] which put to rest the idea that traditional models, such as the Poisson process, could adequately explain modern packet traffic.

At the time it was difficult to obtain data, and the initial claims of ‘fractal traffic’ were met with some scepticism. There were those that claimed that the empirical evidence was in fact an artifact of poor estimation techniques, which were not robust to non-stationarities in the data.

Two developments helped quiet the skeptics view, leading to the situation today, where it is universally accepted that a wide variety of time series extracted from traffic data, including the most important examples, namely packet and byte counts (per discrete time bin), display long-range dependence (LRD). Note that the stronger claim, of self-similarity in the strict (statistical) sense, was only really claimed for Ethernet traffic, or for data aggregated over relatively large time scales such as 1 second and above.

The first development was the discovery of a physically meaningful cause, an underlying property which could be independently tested for, which had a genuine meaning in networking terms, and which could be linked predictively to the observed LRD in a well defined way. This property was heavy tails (with a tail index corresponding to infinite variance) of file sizes. The seminar paper here was that of Cunha, Bestavros and Crovella [8] in 1995, where World Wide Web documents were shown to exhibit this feature. The body of literature on renewal reward processes, notably the work of Taqqu et. al from the 1970s, established the links between this and LRD, which have recently been extended, partly through the stimulation provided by the traffic discoveries.

The second was a greater effort to test for the robustness of the empirical findings. As a greater variety of estimators were used, a common picture emerged: the empirical phenomenon was found to be widespread, and the fractal interpretation was not spurious.

The second was a greater effort to test for the robustness of the empirical findings. As a greater variety of estimators were used, a common picture emerged: the empirical phenomenon was found to be widespread, and the fractal interpretation was not spurious.
Wavelet based estimators in particular ([9], [10]), as they have particular advantages for the study of scaling processes, helped settle the question. The wavelet approach will be discussed in more detail below.

III. THE AFTERSHOCK

Despite the acceptance of a mechanism explaining LRD being almost a decade old, there is even now no one model, or indeed even a group of models, which are accepted as definitive for the description of traffic. There are many which can be put forward which are faithful to the accepted mechanism, but which are different in many other ways both from traffic and each other. Techniques to distinguish between alternatives based on networking realities have been slow to emerge.

In terms of scaling behaviour, the discovery of evidence for multifractal scaling, a much richer form of scaling behaviour associated with non-uniform local variability, raised hopes that another ‘traffic invariant’ had been found which could lead to a complete, robust model of aggregate Wide Area Network (WAN) traffic over all time scales. There is now a literature which accepts the existence of multifractal traffic, exploring alternative multifractal models [11], traffic generators [12], and related performance studies [13]. More broadly, it had become in the late 1990’s somewhat accepted that traffic has multifractal characteristics, despite the fact that physical mechanisms, and network meaning, had never been established in the way it was for LRD.

The first papers on the topic appeared in 1997. Riedi and Véhel [14], using increment based estimation applied to discrete packet and byte counts of WAN traffic (with bin size $\delta = 150$ms), and packet inter-arrivals, concluding strongly in favour of multifractal scaling. Shortly after Feldmann, Gilbert, Willinger and Kurtz [15] made similar conclusions, although significantly their (wavelet based) analysis focussed instead on scales beginning at $\delta = 10$ms and ending at around 1 second. Beyond this ‘small scale regime’ multifractality was not indicated, so that over ‘large scales’ a monofractal, LRD model was adopted. In contrast to these new WAN findings, Ethernet data was reexamined and no evidence for multifractality found.

In a follow up paper by Feldmann et al, an attempt was made to link the observed behaviour to network characteristics such as the capacity of bottleneck links, and round trip times. The results were not definitive, but questions where raised about whether a true mechanism for multifractal scaling was really present. Meanwhile, in a series of papers, including [16], an even richer scaling class, infinitely divisible cascades (IDC), which includes multifractals as a special case, were used to argue that the two scaling regimes, rather than being entirely separate, could be seen as a variation of the same underlying phenomenon. For the first time, confidence intervals for the multiscale analysis were calculated and used, and it was noted that the case for multifractality over monofractality was marginal, even over the small scales. Thus, the initial enthusiasm and confidence that a new kind of scaling had been found became increasing muted.

More recently, several authors, including [17], pointed out that in many ways packet arrivals looked Poisson at small scales, which seems incompatible with multifractality, and Hohn, Veitch and Abry [18] argued not only that multifractality was not present, but that the empirical evidence had been misleading all along. Finally, the same authors took a fresh look at the available evidence in [19]. Applying a more systematic use of confidence intervals, and using insights into the weaknesses of the available tools gained from [18], they concluded that no case for multifractal models can be convincingly made for the traffic analysed thus far. Essentially, the empirical evidence was due to what could be called ‘artifacts’ and did not represent true scaling at all.

IV. THE ESTIMATORS

To explore the statistical issues further we need to be more precise, and to do so we adopt a wavelet viewpoint, first introduced to traffic analysis in [9], due to its advantageous statistical and computational properties. We use software freely available at [20] to perform the statistical analysis both at second order and at higher order, and are guided by the methodology outlined in [21]. Further details of the use of these tools in the networking context can also be found in the review article [22]. A more mathematical introduction to multifractal processes can be found in [23].

Long-Range Dependence is a form of asymptotic scale invariance in the limit of large scale (low frequency). If $X(t)$ is a continuous time stationary process with power spectral density $\Gamma_X(\nu)$, LRD can be defined as a power law divergence of the spectrum at the origin:

$$\Gamma_X(\nu) \sim c\nu^{-\alpha}, \quad |\nu| \to 0, \quad \alpha \in (0, 1).$$

(1)

To detect this phenomena using wavelets, first define the discrete wavelet transform coefficients as

$$d_X(j, k) = \int_{-\infty}^{\infty} X(t)\psi_{j,k}(t) dt,$$

(2)

where the member $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)$ of the basis function family is generated from the mother wavelet $\psi(t)$ by dilation by a scale factor $a = 2^j$, and translation by $2^k$. At fixed octave $j$, the sequence $\{d_X(j, \cdot)\}$ corresponds to an analysis of $X(t)$ at scale $2^j$. It can be shown that the variance of this process satisfies:

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

(3)
moments, which satisfies $q$ its slope is an estimate of the corresponding spectrum, but much better suited to the study of fractal processes. In the case of LRD it reads

$$E|dx(j, \cdot)|^2 \sim c_2 C(\alpha) 2^{\alpha n}, \quad j \to +\infty,$$

where $C(\alpha) = \int |\nu|^{-\alpha} |\Psi(\nu)|^2 d\nu$ is close to a constant. To estimate the wavelet spectrum from data, the time averages

$$S_2(j) = \frac{1}{n_j} \sum_k |dx(j, k)|^2,$$

(5)

where $n_j$ is the number of $dx(j, k)$ available at octave $j$ (scale $a = 2^j$), perform very well, because of the short range dependence in the wavelet domain. Indeed, the mother wavelet is characterised by an integer $N \geq 1$, known as the number of vanishing moments, which satisfies $\int t^k \psi(t) dt = 0$ for all $k = 1, 2, \ldots, N - 1$, and $\int t^N \psi(t) dt \neq 0$. It plays a central role in the wavelet based analysis of long memory processes, since the wavelet coefficients $\{dx(j, k), k \in \mathbb{Z}\}$ at a given scale $2^j$ are short range dependent provided $N > \alpha - 1$. This has been proven in various contexts, see [21] for a review. To enable approximate but analytic analysis of the performance of the scaling exponent estimation procedures, we idealised this whitening of the $dx(j, k)$ to exact independence. Numerical simulations such as those reported in [9], [24] show that this is a useful approximation.

A plot of the logarithm of the estimates $S_2(j)$ against $j$ we call the Logscale Diagram (LD):

$$LD: \quad \log_2 S_2(j) \text{ vs } \log_2 a = j. \quad (6)$$

In these diagrams, straight lines constitute empirical evidence for the presence of scaling. For example, a straight line observed in the range of the largest scales with slope $\alpha \in (0, 1)$ both reveals long memory and measures its exponent $\alpha$.

The above definition of scaling is second order based. If $X$ were Gaussian then this would be sufficient, but this is far from the case for TCP/IP traffic over small timescales. One can generalise the 2nd order definition and study $q$th order quantities $E|dx(j, \cdot)|^q$, for arbitrary $q \in \mathbb{R}$, by using the estimates

$$S_q(j) = \frac{1}{n_j} \sum_k |dx(j, k)|^q.$$

(7)

Just as the wavelet spectrum serves as a statistically effective summary of second order statistics regardless of whether scaling is at issue or not, a plot of $S_q(j)$ against $j$, the $q$th order Logscale Diagram:

$$q-LD: \quad \log_2 S_q(j) \text{ vs } j. \quad (8)$$

is a useful way of examining the raw $q$th order content of data, independently of any multifractal question.

For $q$ fixed, a behaviour $E|dx(j, \cdot)|^q = c 2^j \alpha_n$ over some scale range is seen as a straight line in the $q$-LD, and a measurement of its slope is an estimate of the corresponding $q$-specific scaling exponent $\alpha_q$. If, for each $q$, a straight line is found in the $q$-LD over the same range of scales, then the scaling exponents $\{\alpha_q\}$ are the manifestation of a single underlying scaling phenomenon which we refer to as multiscaling.

We now explain the connection to multifractals. For many multifractal processes, the collection of exponents $\{\alpha_q\}$ are related to the so called multifractal spectrum, which captures the essential details of the multifractal scaling, and can be used to estimate it [23]. We do not attempt to estimate the multifractal spectrum itself, as this would introduce even more estimation difficulties. Instead we adopt the simpler operational approach [21] of testing for linearity of the function $\zeta(q)$. This is because for simple cases such as the exactly self-similar (H-SS) processes with Hurst exponent $H$, for the corresponding increments processes (which are stationary) $\zeta(q) = qH = \alpha_q + q/2$ is a simple linear function, and one speaks of monofractality, whereas for true multifractal processes this is not the case. The same is true for LRD (stationary) processes for which at large scales $\zeta(q) = q(\alpha_2 - 1)/2$. Deviations from linearity can therefore be taken as evidence for the more complex multifractal (MF) behaviour, where a single scaling exponent is insufficient.

V. A TELLING RESULT

Figure 1 is adapted from [19] (the paper mentioned above that concluded against multifractality) It captures one of the main points of that paper. On the bottom row is the wavelet based multifractal analysis of backbone Internet data for three different $q$ values. The top row shows the same analysis for a fitted Poisson cluster model introduced in [18]. The results are remarkably similar (not just in form but even down to actual values, including those of the fitted slopes and therefore of the scaling exponents), despite the fact that the model was fitted used first and second order moments, and tail behaviour, and not higher order statistics.

Although the close agreement between the data and the model is satisfying, the key point here is something quite different. The cluster model is not multifractal, and yet, it reproduced, empirically, a non-trivial multiscaling signature (at least to the same extent as the data). In fact, as we will explain in the talk in more detail, the cluster model can give rise to this apparent or pseudo scaling through a transition effect between the large and small scale regimes.

Informally, what is at work here is low power of the estimation procedure, pointing to the need of a formal hypothesis test for the presence of multifractality. In the talk these issues will be discussed in more depth. The development of more powerful tools is one of the key developments required to settle questions of scaling behaviour which may be present after all in traffic today (for example at large scales for certain traffic components), or into the future and the dynamics of networks and traffic protocols evolve.
Fig. 1. \( q \)-th order Logscale Diagrams for the point process of packet arrivals comparing traffic data (bottom row) from a fitted model (top). In each case \( q = \{0.5, 2, 4\} \) from left to right. In each plot there is evidence of twin scaling regimes: at coarse scales, \([j^*, j] = [1, 6]\), and fine scales, \([j^*, j] = [-7, 1]\).

REFERENCES


