Bulletin of the Maritime Institute

Gdańsk, 1999, 26, 2

JERZY KONORSKI

Department of Marine Electronics
Maritime Institute, Gdańsk, Poland

DETECTION AND QUANTIFICATION OF SELF - SIMILARITY IN DATA TRAFFIC FOR PREDICTION OF PERFORMANCE OF MARINE DATA FILE TRANSFER

Abstract

The time-sensitivity of large marine data files over a communication network necessitates accurate simulative prediction of the file transfer performance. Careful data traffic modelling is required to fit the actual traffic characteristics for subsequent generation of synthetic traffic traces and feeding them into a simulation model. Classical models have recently proved inadequate due to the discovery of self-similarity (fractal behaviour) in data traffic. This paper attempts to systematise the mathematical background of self-similarity and the ways it manifests itself in stochastic processes modelling data traffic. Relevance of self-similarity to traffic description and measurements is discussed. Results of a research effort at the Department of Marine Electronics of the Maritime Institute in Gdańsk are described, which focus on the development of a software tool for detection and quantification of self-similarity in observed or synthetically generated data traffic.

1. Motivation

Within the HIROMB (*High Resolution Operational Model of the Baltic Sea*) project [5], cyclic transfer of large sets of marine data between Swedish Meteorological and Hydrological Institute and the Maritime Institute in Gdańsk takes place using the FTP services of the TCP/IP Internet. The files involved are on order of a few Mbytes in size and contain frequently updated data, with an update cycle on order of hours. It is critical for the HIROMB activity that they reach their destination in time to be processed before the next update. In reality, transfer delays are often close to or beyond the resulting deadlines, some transfers never completing successfully. Given unpredictable but generally heavy background traffic along the route in question, and the fact that FTP does not offer any QoS (quality of service) guarantee, this raises the issue of meeting quasi-real-time requirements in a best-effort communication environment. Accurate prediction of the file transfer performance calls for a simulation project incorporating detailed knowledge about the encountered traffic conditions and aimed at

- evaluation of the performance capability in the present connection configuration,
- recommending optimum file sizes, compression and formatting, as well as optimum transfer modes e.g., parallel transfers,
- projection of performance in view of expected connection reconfigurations and traffic increase,
- comparison of various options as to the future choice of the network service provider e.g., TCP/IP versus Frame Relay, ISDN etc.

A number of ready-made software tools for performance evaluation are currently available, among them CACI's COMNET III$^{®}$ simulator package whose advantages include a user-friendly graphical simulation model creator and the possibility of feeding traces of actually recorded traffic into the created simulation models. The following phasing of the project thus naturally emerges:

1. measurement and parameter estimation of actual network data traffic,
2. traffic modelling to fit the actual traffic characteristics for subsequent generation of synthetic traffic traces, and
3. feeding the generated traffic traces into a simulation model reflecting the assumed route of the file transfer.

This work is primarily concerned with phase 1 and motivated by the fact, extensively reported in recent literature, that classical traffic models (based on Poisson, Markovian, renewal, ARMA processes etc.) have proved inadequate for today's data traffic description. Their failure is mainly due to the occurrence of the so-called self-similarity or fractal behaviour in data traffic, a statistical phenomenon long since discovered in some hydrological or economic processes, but only recently recognised in a telecommunication context. Understanding self-similar data traffic and statistical estimation of its parameters is thus a prerequisite for realistic traffic modelling and subsequent simulation. In the sequel, based on an extensive survey of current research world-wide, an attempt is made to systematise the mathematical background of self-similarity and the ways it manifests itself in stochastic processes modelling data traffic. Next, relevance of self-similarity to traffic description and measurements is discussed. Finally, some results of a research effort at the Department of Marine Electronics of the Maritime Institute in Gdańsk are described, which focus on the development of a software tool for detection and quantification of self-similarity in observed or synthetically generated data traffic.

2. Mathematical background

In this section, the notions of "self-similarity," "fractal" and "long-range dependence" in the context of data traffic in communication networks will be briefly explained in terms of the theory of stochastic processes, with an emphasis upon peculiar scaling behaviour when moving from small to larger time scales. Since fractals are geometrical objects that seem to appeal to the human mind through intuitions they defy rather than agree with, the stochastically counterintuitive behaviour of self-similar data traffic will also be pointed out in a subsequent section for better understanding.

Data traffic observed (actually or conceptually) at some measurement location can be thought of as a succession of arrivals of data units, hereafter called PDUs for *Protocol Data Units*, each carrying a number of bytes. We shall only be interested in the byte volume of the traffic, meaning that PDUs will not be classified by contents, protocol type, source/

destination addresses etc. A suitable mathematical model is that of a realisation (sample path) of a stochastic process, with at least three settings possible, to be chosen from as computational convenience dictates. On the left in Fig. 1, only the PDU arrival epochs are marked thus constituting a one-dimensional point process; weights can possibly be attached to points to record the byte volume. In the middle, a discrete setting is displayed where the time axis is divided into slots of equal size and a time series (X_t, $t = 0, 1, 2,...$) is defined so that the random variable X_t represents the total number of bytes in the t^{th} slot. A continuous setting shown on the right is sometimes easier to handle analytically, with the process (X_t, $t \in \mathbf{R}^+$) representing *momentary traffic rate* variation in time, a limiting case of the former time series as the slot size tends to 0. The X_t‘s can also be viewed as increments of the *accumulated traffic process* defined as $Y_t = \sum_{i=0}^{t} X_i$ or $Y_t = \int_0^t X_u du$ for the discrete and continuous setting, respectively.

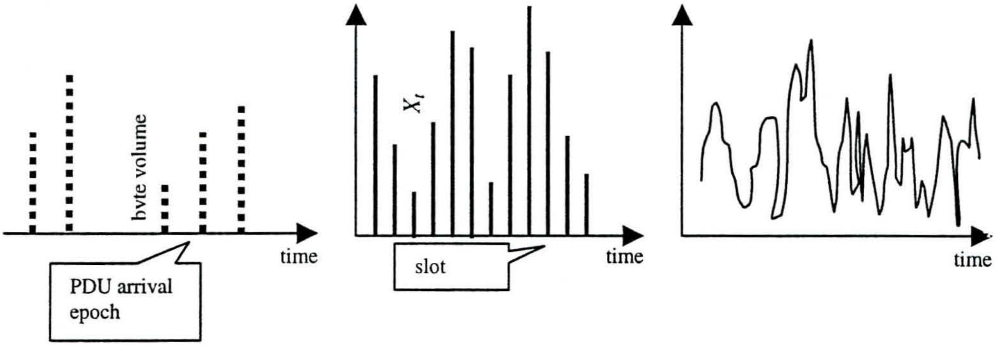


Fig. 1. Point process, discrete and continuous settings of data traffic

Although data traffic is in general non-stationary over a large time scale (e.g., due to diurnal cycles of users' activity, network reconfiguration or breakdowns), it will be further considered on at most hour-long time scales and assumed to be *covariance stationary* i.e., the marginal probability distribution of X_t is assumed to be time-invariant and the autocovariance only a function of the time lag:

$$Cov\{X_t, X_{t+u}\} = E\{\overset{\circ}{X}_t \overset{\circ}{X}_{t+u}\} = \rho(u) \qquad (1)$$

($\overset{\circ}{X}$ denotes the centred version of a random variable X). Moving between different time scales is equivalent to the operation of *traffic aggregation* defined as

$$Y_t^{(m)} = Y_{mt}, \qquad (2)$$

where m, a positive real number or a positive integer depending on the setting, is called the *level of aggregation*. Indeed, as m gets larger, so does the underlying time scale since the increments of Y_t are calculated over larger time intervals (i.e., larger blocks of data):

$$X_t^{(m)} = \sum_{i=m(t-1)+1}^{mt} X_i \quad \text{or} \quad X_t^{(m)} = \int_{mt}^{m(t+\Delta t)} X_u du\,. \qquad (3)$$

The association of a class of stochastic processes with fractals stems from scaling considerations. A fractal is ordinarily described as a geometrical object any part of which somehow resembles the whole upon scaling [17]. Such a form of *topological self-similarity* is formally expressed as there existing some exact affine transformation which maps any part F' of a fractal F directly onto F. [4] notes that *stochastic self-similarity* is a weaker version than its topological counterpart in that the existence of an exact transformation is not required. However, stochastic self-similarity can be quantified based on the theory of H-stable probability distributions. Namely, dropping the subscript t for clarity and assuming for the moment that $X^{(m)}$ is centred, either the distribution of $X^{(m)}$ is H-stable or it is attracted by some H-stable distribution, meaning that,respectively,

$$m^{-H}X^{(m)} \overset{D}{=} X \text{ for all } m > 0 \text{ or } m^{-H}X^{(m)} \xrightarrow{D} X^* \text{ as } m \to \infty \tag{4}$$

where $H \geq 0.5$, X^* is a limit random variable with a H-stable distribution and the D marks equality and asymptotic equality in the distributive sense. Informally, (4) expresses a peculiar form of scaling and thus stochastic self-similarity, and standard definitions [2,9] accordingly distinguish between *exact* and *asymptotic self-similarity*, H being called the *Hurst parameter* for historical reasons[1]. Although (4) implies that all distributions are asymptotically self-similar, we note that not all values of H lead to the departure from classical traffic models e.g., most known distributions are attracted to the normal distribution with $H = 0.5$ by virtue of the Central Limit Theorem. Self-similar behaviour becomes interesting for $H > 0.5$ i.e., when the involved distributions seem not to adhere to the Central Limit Theorem. A stochastic process X_t is said to be *self-similar* (or *fractal*) if the equality in (4) extends to all its finite-dimensional probability distributions and *second-order self-similar* (*fractal*) if it is confined to two-dimensional distributions only.

Many traffic statistics exhibit scaling as in (4); especially studying variances is fruitful since it gives insight into the "internal structure" of self-similarity [15]. Taking the continuous setting we have from (1)-(3) the variance of the aggregated accumulated traffic:

$$Var\{Y_t^{(m)}\} = 2mt\int_0^{mt}(1-\frac{u}{mt})\rho(u)du \tag{5}$$

and we see that its asymptotic form (as $m \to 0$) is decided by the integrability of $\rho(\cdot)$. Namely, if $\int_0^\infty \rho(u)du < \infty$ (meaning non-correlated traffic rates that are sufficiently far apart in time) then, regardless of $\rho(\cdot)$,

$$Var\{m^{-1/2}Y_t^{(m)}\} \sim t \tag{6}$$

($\sim$ stands for asymptotic equality). On the other hand, take $\rho(u) = \phi(u)\cdot u^r$, where $\phi(\cdot)$ is an asymptotically constant function and $r \in (-1,0)$, meaning strongly correlated traffic rates that are considerably far apart in time. Then

$$Var\{m^{-(r+2)/2}Y_t^{(m)}\} \sim Var\{Y_t\} \sim t^{r+2}, \tag{7}$$

which states self-similarity with the Hurst parameter $H = (r+2)/2 \in (0.5, 1)$. Consequently, asymptotic self-similarity with $H > 0.5$ and asymptotic power-law decay of $\rho(\cdot)$ are closely connected.

[1] After H.E. Hurst who, almost half a century ago, discovered self-similarity phenomena in certain natural processes.

3. Self-similarity manifestations and related intuitions

Self-similarity with Hurst parameter $H > 0.5$ (also termed the *Joseph effect* after [10]) manifests itself in a number of ways, both stochastically and geometrically (in particular, visually). Careful analysis of these manifestations leads to the formulation of various criteria of self-similarity, equivalent to one another in that they are indicative of the same phenomenon, although it is argued [21] that they differ subtly in mathematical content.

3.1 Visual manifestations

The most obvious for the naked eye is the so-called pictorial proof of self-similarity [9] which consists in plotting the realisations (sample paths) of $X_t^{(m)}$ for a range of levels of aggregation, m. To verify that traffic self-similarity is not unique to the measurements presented in the literature [1,9,12,20], an ad-hoc measurement experiment on an Ethernet LAN segment was carried out; generic plots obtained therefrom are depicted in the left column of Fig. 2. The original byte volume readouts were taken at 10 ms intervals (10 ms slot size) and subsequent aggregation produced 100 ms, 1 s, 10 s and 100 s slot sizes corresponding to $m = 10^1$, 10^2, 10^3 and 10^4. The sample path at the bottom left almost captures individual PDU arrivals and predictably exhibits high burstiness typical of multi-user data traffic. It is nevertheless visibly smoother than a Poisson process on the same time scale appearing at the bottom right, produced from a synthetic trace of a software traffic generator. As m increases, moving toward larger time scales, the burstiness does not vanish in the actual traffic and shows even on the uppermost minutes-long time scale, whereas it does vanish in the Poisson traffic (right column). Thus apart from the time scale, the upper plot appears indistinguishable from the ones beneath, a fact contradicting the intuition based on the law of large numbers. Suggestive ways of expressing this fact include statements like "data traffic is bursty over several time scales" [12], "there is no natural burst size" [4], "traffic 'spikes' ride on longer-term 'ripples' which in turn ride on still longer-term 'swells'" [14] or "the presence of cycles of all frequencies and orders of magnitude displays features suggestive of non-stationarity" [8].

3.2 Stochastic manifestations

Stochastically, the most striking in self-similar traffic with $H > 0.5$ is its non-degenerate correlation structure despite increasing the level of aggregation, a feature referred to as *long-range dependence* and shown in Sec. 2 to entail power-law decay of the autocovariance function. Specifically, denoting by $\rho^{(m)}(\cdot)$ the autocovariance function of $X_t^{(m)}$, we observe for an asymptotically self-similar process that

$$\rho^{(m)}(u) \to \phi(u) \cdot u^{2H-2} \tag{8}$$

as $m \to \infty$.[2] Long-range dependence is reflected by the spectral density of X_t (the Fourier transform of $\rho(\cdot)$) being divergent at the origin, the asymptotic expression near the origin again having a power-law form

$$S(\omega) \sim (1/\omega)^{2H-1}. \tag{9}$$

[2] Recalling the form of $\rho(\cdot)$ assumed earlier, we conclude that for sufficiently large m, $\rho^{(m)}(u) \sim \rho(u)$ as $u \to \infty$, which again justifies the term "asymptotic second-order self-similarity". For an exactly second-order self-similar process, $\rho^{(m)}(u) \equiv [(u+1)^{2H} - 2u^{2H} + |u-1|^{2H}]/2$. Such a process is hard to come across in practice, but often serves as a reference model e.g., the *Fractional Gaussian Noise* [2,5,15].

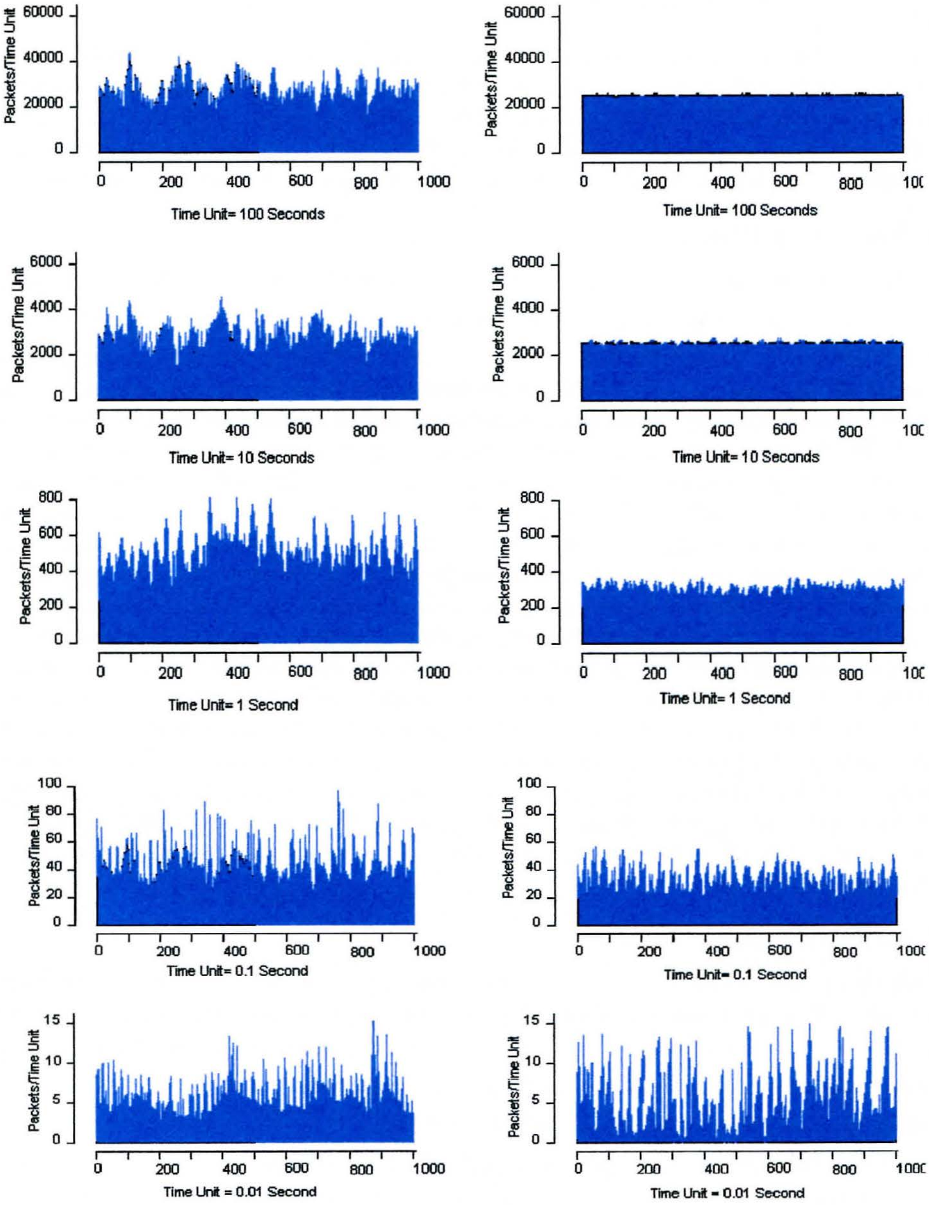


Fig. 2. Pictorial proof of traffic self-similarity (source: ad-hoc Ethernet LAN measurement)

This is termed the *1/f noise effect*. Another name self-similarity sometimes goes by is *slowly-decaying variance*, an effect already demonstrated in Sec. 2, but rewritten more convincingly in the form

$$Var\{\, Y_t^{(m)} \,/\, m \,\} \sim m^{2H-2} \tag{10}$$

as $m \to \infty$; the variance thus decays more slowly than $1/m$. All the effects just described remain in sharp contrast with the intuition based on classical traffic models and observations (e.g., of telephone and early data network traffic) which exhibit

- short-range dependence, reflected by an integrable autocovariance function tending to that of a pure white noise as $m \to \infty$ i.e., $\rho^{(m)}(u) \to 0$ for all $u \neq 0$,
- finite spectral density at the origin, and
- $Var\{ Y_t^{(m)} / m \}$ decaying as $1/m$.

Yet another stochastic manifestation of self-similarity is produced by fitting the so-called *ON-OFF traffic model* to the point process representing individual PDU arrivals over time. Because data traffic is bursty in nature, PDU arrivals tend to cluster into *PDU trains* (or *ON-periods*), separated by idle *intertrain* intervals (or *OFF-periods*), both of variable length. It was noted long ago [10] that self-similarity and long-range dependence are tightly related to the probability distributions of the ON- and OFF-period lengths exhibiting power-law decay (or, to use a more descriptive term, having *heavy tails*). An example is the Pareto distribution:

$$Pr[\text{ON/OFF length} < x] = 1 - (c/x)^b \tag{11}$$

with $b, c \geq 0$ and $x \geq c$; b and c are the *shape* and *location parameters*, respectively. If $b \leq 2$ then the Pareto distribution has infinite variance and if $b \leq 1$, infinite mean as well. The fact that observed data traffic fits the Pareto ON-OFF model with $b \leq 2$ gives rise to the term *infinite variance syndrome*, also called the *Noah effect*, after [10] (see also [19,21]). It might appear at first glance that, since the delimitation of PDU train and intertrain intervals in a bursty point process is to some extent arbitrary (it results from superimposing the ON--OFF model rather than from the process' internal mechanisms), the probability distributions of the ON- and OFF-period lengths should critically depend on how a PDU train is defined. E.g., one might decide that a minimum of Δ s without PDU arrivals constitutes an intertrain interval, PDU trains being thus determined unequivocally. The Noah effect, however, turns out largely insensitive to the threshold Δ due to a unique property of the Pareto distribution, called *invariance under truncation from below* [14]. It states that conditioning the Pareto distribution $Pr[I < x]$ on $I \geq y$ leaves the distribution unchanged except for y becoming the new location parameter. As a consequence, it may be shown that $E\{\text{OFF length}\}$ is independent of Δ [21] whereas

$$E\{\text{ON length}\} = o(\Delta) \tag{12}$$

for $b < 2$,with $o(\Delta)$ defined by $o(\Delta)/\Delta \to 0$ as $\Delta \to 0$ [14]. Note that common intuition based on finite-variance distributions dictates that $E\{\text{ON length}\}$ should be proportional to Δ, which is true for $b \geq 2$. In fact, the PDU train lengths do grow with Δ, but only very slowly, while the intertrain lengths do not at all, causing the traffic to retain its apparent burstiness over many time scales. The connection between Pareto distributions and self--similarity is further confirmed by a key theorem [19] stating that a superposition of sufficiently many processes fitting the Pareto ON-OFF model with possibly distinct shape parameters b_1 and b_2 for the ON- and OFF-periods converges to an exactly self-similar process with

$$H = (3 - \min\{b_1, b_2\})/2. \tag{13}$$

Table 1 below summarises the stochastic manifestations of asymptotic self-similarity with Hurst parameter H.

Table 1. Stochastic manifestations of self-similarity

Manifestation	Statistical measure involved	Relation to the Hurst parameter
Joseph effect	Probability distribution	$m^{-H} X^{(m)} \xrightarrow{D} X*$
Long-range dependence	Autocovariance function	$\rho(u) \sim u^{2H-2}$
$1/f$ noise effect	Spectral density	$S(\omega) \sim (1/\omega)^{2H-1}$
Slowly-decaying variance	Variance of accumulated traffic	$Var\{ Y_t^{(m)} / m \} \sim m^{2H-2}$
Noah effect (infinite variance syndrome)	PDU train/intertrain length distribution	Pareto with shape parameters b_1, b_2 s.t. $H = (3 - \min\{b_1, b_2\})/2$

3.3 Geometrical manifestations

Sample paths of momentary traffic rate or collections of PDU arrival epochs on the time axis can be mapped onto geometrical objects whose fractal properties can be investigated based on topological notions. Leaving sample paths out for the moment (cf. Higuchi's method in Sec. 6), we shall touch on two measures related to the correlation structure in the point process setting: the *coincidence rate* [17] and *correlation dimension* [3]. Both attempt to establish how densely the point process fills the space it is embedded in (i.e., the time axis). The coincidence rate, given by

$$g(u) = \lim_{\delta \to 0} \frac{Pr[\text{PDU arrivals in } (0, \delta) \text{ and } (u, u+\delta)]}{\delta^2} \tag{14}$$

is akin to the autocovariance function and is known to scale as $u^{\alpha-1}$ for self-similar processes, $\alpha \in (0, 1)$ being called the *fractal exponent* of the process. The correlation dimension is derived from a set of PDU arrival epochs by calculating the proportion of pairs of epochs that are u seconds apart or less. For self-similar processes it is known to scale as u^{α} over a range of time scales, where α is the correlation dimension. For more analysis of geometrical measures see [6].

4. Relevance of self-similarity to understanding data traffic

Understanding the nature of data traffic is important both from the network designer's perspective, for proper dimensioning of the network resources the traffic is to be handled by, and from the user's perspective (essential in the present work), for proper prediction of performance of a planned activity on the network e.g., a series of time-sensitive large file transfers. To paraphrase [12], understanding data traffic has two possible levels:

- Concrete causal (or microscopic), whereby one thinks of individual events like file transfers, interactive sessions etc. and builds a traffic model by aggregation of the event models, and
- Abstract statistical (or macroscopic), whereby one seeks a stochastic traffic model that fits some relevant aggregate statistics without worrying about its consistence with the individual event models.

Because of the multitude of individual events that would have to be kept track of, and because the stochastic paradigm has been so entrenched within the data traffic research community, macroscopic understanding tends to prevail. It is somewhat ironical, then, that the initial resistance toward self-similar traffic models used microscopic arguments. [21] quotes two questions that stood in the way, concerning the lack of a "physical" explanation for the observed traffic self-similarity and its unclear impact on queuing processes within the network. Both questions, it may be conjectured, stemmed from the long-lived tradition of measuring data traffic either on very small time scales (to track individual PDUs) or very large ones (to determine average resource utilisation e.g., to fix tariffs), where the medium-term correlation structure remained invisible. Over time, the opposition subsided faced with hard empirical evidence including surprisingly poor performance at relatively low traffic rates and highly irregular timing of PDU losses due to resource overflow, both phenomena now attributed to traffic self-similarity. As argued in [4], any realistic traffic model of today should incorporate self-similarity or long-range dependence or heavy-tailedness of PDU train/intertrain distribution, or all the three.

Conflicting views exist as to whether and why self-similar traffic models should be given attention, cf. [7,8,12,16]. However, three aspects of self-similarity seem to safeguard its future role: accuracy of statistical inference, parsimonious traffic modelling and queue distribution tail behaviour.

4.1 Accuracy of statistical inference

Not taking proper account of self-similarity phenomena may prove disastrous in classical significance tests and confidence interval assessment. Cases are known where sampling errors obtained for conventional traffic models are wrong by a factor that tends to infinity as the sample size increases [8]. A similar point can be made using standard estimation arguments. Consider the mean traffic rate $\lambda = E\{X_t\}$ whose straightforward estimator is $\hat{\lambda} = E\{Y_t\}/T$ (T is the observation period). It is unbiased and, as seen from Table 1,

$$Var\{\hat{\lambda}\} = C{\cdot}T^{2H-2}. \tag{15}$$

For accuracy, one needs to keep the coefficient of variation of $\hat{\lambda}$ below a threshold ε, which in turn requires

$$T > \sqrt[1-H]{\frac{\sqrt{C}}{\lambda\varepsilon}}\ . \tag{16}$$

Unfortunately, the right-hand side of (16) grows very quickly as $H \to 1$; e.g., for values typical of a LAN segment traffic ($C = 4{\cdot}10^{11}$ bit·s, $\lambda = 2{\cdot}10^6$ bit/s) and a moderate accuracy $\varepsilon = 10\%$, one requires $T > 5$ minutes if $H = 0.8$, but $T > 27$ hours at $H = 0.9$. Such large observation periods are certain to encompass diurnal non-stationarities thus discrediting any statistical inference geared to covariance stationary processes.

Proximity of the Hurst parameter to 1 also proves dangerous to the estimator of the momentary traffic rate variance. Based on the Fractional Gaussian Noise model and assuming that N samples of X_t are taken over the observation period, it follows that the estimator is biased this time, the bias vanishing slowly with N (in fact, as N^{2H-2} [15]). Fig. 3 illustrates that the number of samples required to keep the bias under any given level grows dramatically with H; for large H, at least one realisation of X_t is likely to occur for which the variance estimator is meaningless even for large N.

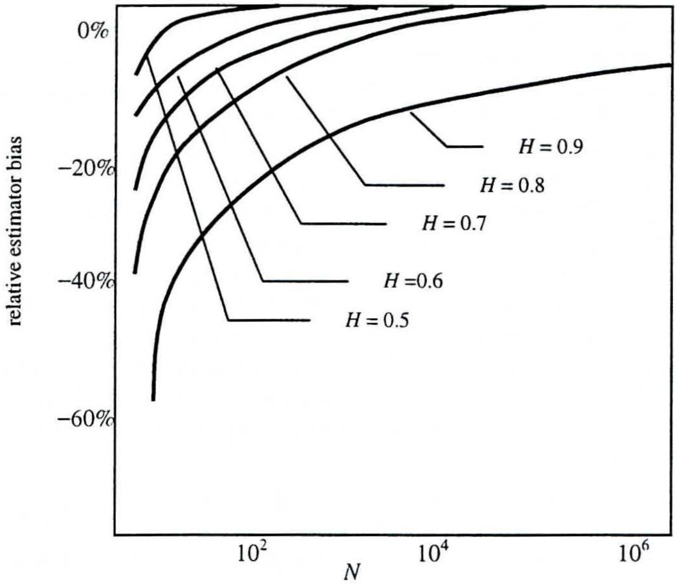


Fig. 3. Estimator bias of the momentary traffic rate variance [15]

4.2 Parsimonious traffic modelling

Generation of synthetic traces of traffic based on specific self-similar traffic models plays a role in performance-oriented computer network simulation. Markovian traffic models were extensively used for telephone and early data network traffic and were successively enriched as WWW, multimedia and interactive traffic was beginning to dominate (cf. ARMA, MMPP and Markov-modulated ON-OFF source models [15]). Their increased sophistication, meant to capture the behaviour of individual traffic components, was not unlike approximating a hyperbolic function by a sum of exponential ones [8]: it involves prohibitively many parameters whose "physical" interpretation becomes impossible. Meanwhile, the traffic descriptor that the models were to fit (consisting of classical measures like the peak-to-mean ratio) was becoming a problem in itself. E.g., [8] cites LAN traffic measurement data where the peak-to-mean ratio varies from 150 in any 5 ms interval to over 700 in any 5 s interval, making it an entirely unrepresentative measure.

Applying the self-similarity apparatus instead reduces the number of correlation-capturing parameters to just one - the Hurst parameter, measuring the degree of asymptotic self-similarity (as H increases from 0.5 to 1, so does the process' asymptotic self-correlation, from a pure noise to almost deterministic). This enables parsimonious modelling whereby the multitude of parameters reflecting individual traffic components is replaced by a few parameters, much in the spirit of macroscopic understanding. A suitable traffic descriptor has been proposed based on the following expression for the autocovariance function of a fractal point process [17]

$$\rho^{(m)}(u) = C \cdot \frac{m^{2H-1}}{\tau^{2H-1} + m^{2H-1}} \cdot u^{2H-2} \tag{17}$$

where C and τ are constants, the latter being called the *fractal onset time* and marking the lower limit of significant power-law decay of the autocovariance function. The traffic is fully specified by the triple (λ, H, τ).

4.3 Queue distribution tail behaviour

While the discovery of self-similarity in data traffic was of great cognitive value and the challenging task of testing for self-similarity opened up an interesting field of statistical research, a truly practical question regards the impact upon the expected network performance. PDU delays and losses, which subsequently cause the flow and congestion control mechanisms to throttle the user-perceived throughput, are incurred in a series of queues the traffic encounters at the access control and switching elements on its way from source to destination. A generic FIFO queuing system fed by long-range dependent traffic is known to behave differently from one with non-correlated (e.g., Poissonian) input traffic in that very long queues and resulting PDU losses seem abnormally frequent. One way of quantifying this fact is to plot the complementary probability distribution function of the queue length i.e., the probability that the queue exceeds a given threshold as a function of the threshold. A typical curve obtained by feeding a trace of actual long-range dependent traffic is shown in Fig. 4. Compared with the non-correlated input case (the dashed curve) it exhibits an uncomfortably heavy tail – according to some analyses, governed by the Weibullian distribution [12]. To prevent suspicions that this heavy-tailedness might have arisen due to a specific marginal distribution at the input, the trace was subsequently shuffled (to remove the long-range dependence while retaining the marginal distribution) and again fed into the queue, producing the dotted curve that almost coincides with the dashed one. Another uncomfortable fact is that, since a FIFO queue behaves like a low-pass filter [11], there is little chance for the traffic to lose its long-range dependence as it passes successive queues. On the other hand, some authors argue [12] or even demonstrate via simulation [7] that the flow and congestion control mechanisms within the network may act to suppress the traffic self-similarity to some extent.

5. Qualitative self-similarity assessment

Self-similarity in data traffic would not have been discovered without a surge of massive traffic measurement projects first undertaken at the beginning of the present decade. It was only upon analysing very long traces of observed traffic that the inadequacy of such classical measures as the peak-to-mean ratio, short-term variance [1], indices of dispersion for counts and intervals [11] etc. for capturing long-range dependence became apparent. The main findings so far include:

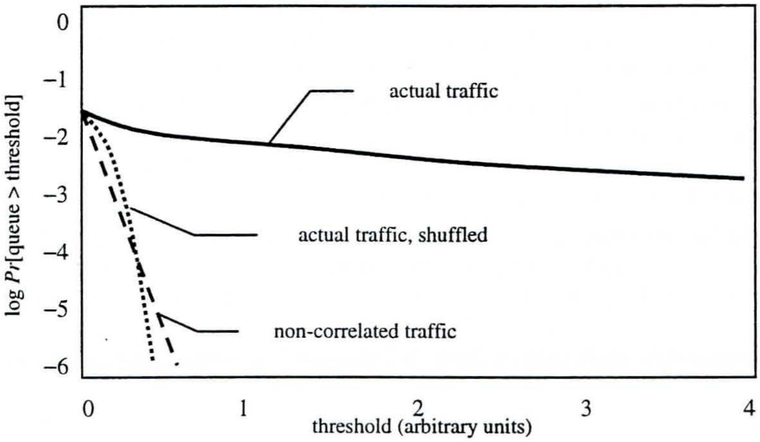


Fig. 4. Effect of self-similarity upon queuing performance [1]

- PDUs related to TELNET activity on the Internet form very different arrival processes depending on whether they carry session requests or intra-session data; the former are Poisson whereas the latter exhibit the Noah effect [14] (similar conclusions pertain to FTP traffic). Some traces of multi-protocol (all-TCP) traffic fit the Fractional Gaussian Noise model.
- Both internal and external Ethernet LAN traffic is self-similar with the Hurst parameter in the range (0.7, 0.95) as computed from ca. 30-minute traces. Most surprisingly, H tends to increase with the mean traffic rate i.e., as more traffic components superpose [9], contrary to the common intuition that traffic superposition has a "smoothing" effect.
- WWW traffic measurements over time scales of 1 s or larger yield H significantly greater than 0.5 (mostly between 0.75 and 0.85); self-similarity is more pronounced on backbone links shared by many source-destination pairs and seems to be related to the distribution of transferred file sizes being heavy-tailed [2].
- Several-minute measurements on a 155 Mbit/s ATM-based WAN reveal $H \approx 0.7$ as well as certain robustness of self-similarity despite attempts to remove it by traffic shaping [11].
- Self-similarity and the infinite variance syndrome have also been detected in most types of networks other than the Internet e.g., ISDN and CCSN (using the SS7 signalling system) [3].

While the methodology of traffic measurement is relatively simple except that below--millisecond timing accuracy is necessary (an output trace simply records the arrival epochs and byte volumes of successive PDUs), the detection and quantification of self-similarity is challenging. The Joseph, Noah etc. effects are but idealisations (not unlike the Markov process) never to be fully validated in finite data [21] – even an imperfect validation should involve very large data sets. Rigorous point estimation may, at a high computational cost, yield conclusions hard to benefit from (e.g., "the shape parameter of the PDU train distribution is 2.0 ± 0.05"). Consequently, besides point estimation, some data-intensive methods have been developed that rely more on visual than quantitative assessment. They include sample path assessment, textured plots and log-moment generating functions.

5.1 Sample path assessment

By plotting the sample paths of a self-similar process over many time scales, just like in Fig. 2, and using analogous plots of Markovian traffic for reference, one may attempt to visually assess the degree of self-similarity at which the Markovian and self-similar traffic begin to look the same. E.g., in [13], traffic self-similarity on a backbone link was found to be induced by the transferred file size distribution being Pareto with the shape parameter $b \le 2$; visual assessment helped to determine $b = 1.95$ as the limit beyond which the two types of traffic were indistinguishable.

5.2 Textured plots

Recall that the delimitation of PDU train and intertrain intervals in the ON-OFF model of a point process is largely a matter of convention. To visually assess the appropriateness of the Pareto ON-OFF model (i.e., to verify that ON- and OFF periods can indeed be distinguished), one constructs a *textured plot* [19,21], which turns the one-dimensional time axis into a dotted strip, each PDU arrival epoch being marked by a dot and displaced vertically by a random amount within the strip. If the time scale is large enough and the Pareto ON-OFF model does fit the observed traffic, one gets a characteristic striped texture sketched in Fig. 5. As noted in [19], this type of texture is much less likely to appear for Markovian (exponential ON-OFF) traffic.

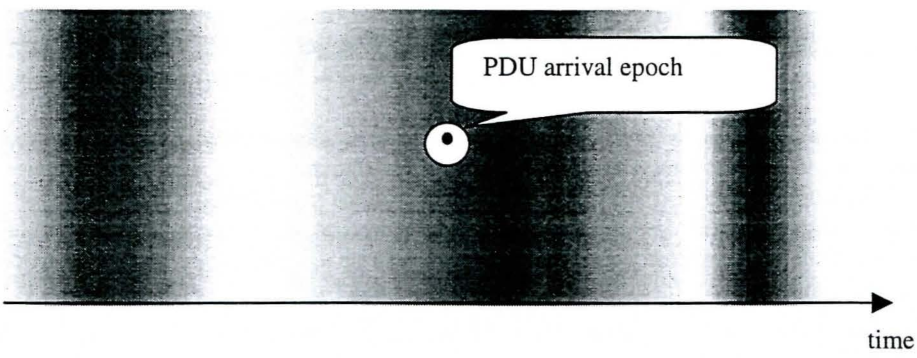


Fig. 5. A typical texture plot for Pareto ON-OFF traffic

5.3 Log-moment generating functions

The detection of self-similarity or long-range dependence in the observed traffic using correlation or spectral measures alone gives little insight into the performance impact it would have when fed into a FIFO queue: queuing theory has always been more oriented toward the workload input rate than its spectral characteristic. A combination of spectral and workload characterisation of data traffic has been proposed [20] using the discrete setting and the *log-moment generating function*:

$$G^{(m)}(\omega) = \frac{\log E\{\exp(\omega X_t^{(m)})\}}{\omega m} \tag{18}$$

As before, the level of aggregation m determines the time scale used (e.g., $m = 1, 2, 4, 8, \ldots$), whereas the variable ω (e.g., $\omega = \ldots, 10^{-1}, 10^{0}, 10^{1}, \ldots$) acts as a weighting of measurement between that of the mean and maximum traffic rate. Indeed, $G^{(m)}(\omega) \to E\{\ X_t^{(m)}\ \}/m$ as $\omega \to 0$ and $G^{(m)}(\omega) \to E\{\sup X_t^{(m)}\ \}/m$ as $\omega \to \infty$. Despite its quantifying potential, (18) only offers qualitative detection of self-similarity – by assessment of shapes. Namely, the shape of the two-dimensional surface $G^{(m)}(\omega)$ over the (m, ω)-plane, in particular that part of it above the constant plane corresponding to the queue's service rate, has been shown to differ qualitatively for self-similar and Markovian traffic.

6. Point estimation of self-similarity-related parameters

Point estimates of the Hurst parameter can be obtained either directly, using one of the methods described below, or indirectly, via the assessment of shape parameters b of heavy tails in superposing traffic components. The latter ordinarily follows by plotting the complementary probability distribution function of the observed PDU train/intertrain lengths on a log-log scale and looking for asymptotically linear behaviour [2,21]. A least squares fit then produces a straight line whose slope is equal to $-b$. To verify how much of the data variability is explained by the Pareto power-law decay, calculation of the relative variance of error (the R^2 statistic) is recommended. Repeated analysis of independent traffic traces permits to compute the confidence intervals for b, if necessary.

Alternatively, the *Hill estimator* of b can be computed from the set of observed PDU train/intertrain lengths $\{I_1, \ldots, I_N\}$ as

$$\hat{b}(k) = \left[\frac{\sum_{i=0}^{k-1} \log(I_{(N-i)} / I_{(n-k)})}{k} \right]^{-1} \tag{19}$$

and plotted against k (the *Hill plot* [21]), where $I_{(j)}$ is the j^{th} order statistic in the set $\{I_1, ..., I_N\}$ and $k \in (1, N)$ is the number of the lowest-order statistics taken. For a heavy-tailed PDU train/intertrain distribution, the Hill plot initially varies considerably with k to level off for large k, whereas in the case of an exponential tail it continues to decrease with k. Confidence intervals can be computed using independent traces or the asymptotic normality property of the Hill estimator.

Two points are worth making when estimating the Hurst parameter directly. First, having obtained an H-estimate close to 1 (or whose 95% confidence intervals covers 1) one should check if the apparent self-similarity is genuine and not induced by some hidden deterministic mechanisms [9]. Second, H-estimates obtained from non-overlapping blocks of the same trace may differ significantly and it is important to decide whether this variability is due to randomness or to the traffic non-stationarity (a suitable χ^2–like test has been developed [8]).

The following methods of the Hurst parameter estimation assume that the observed traffic is at least covariance-stationary (cf. [2,11,15]). Except for Whittle's method, all of them lead to power-law plots of specific statistics against the level of aggregation, m, with an exponent dependent on H. Therefore, H follows from a least squares fit to the corresponding log-log plot. Discrete setting is assumed throughout, $X_1, ..., X_N$ being as before the observed samples of the traffic rate. Formulae (2) and (3) apply too.

6.1 Time-domain methods

The samples of the momentary traffic rate can be used in many ways to construct meaningful m-sensitive statistics. To obtain a *variance-time plot*, one estimates Var{ $X_t^{(m)}$ }:

$$V^{(m)} = \frac{1}{N'(m)} \sum_{t=1}^{N'(m)} \left(X_t^{(m)}\right)^2 - \left(\overline{X}^{(m)}\right)^2 \tag{20}$$

for increasing m (as long as $m \ll N$), where $N'(m) = \lfloor N/m \rfloor$ and $\overline{X}^{(m)} = \frac{1}{N'(m)} \sum_{t=1}^{N'(m)} X_t^{(m)}$.

By (10), $V^{(m)} \sim m^{2H-2}$ (a slope of −1 on the log-log plot indicates the absence of self-similarity). More generally, the empirical k^{th} central absolute moment of $X_t^{(m)}$ scales as $m^{k(H-1)}$. A related measure called the *Index of Dispersion for Counts* (IDC), essentially an estimate of *Var*{ $X_t^{(m)}$ }/*E*{ $X_t^{(m)}$ }, scales as m^{2H-1}. Another statistic, proposed in *Higuchi's method*, is the sample path length estimator:

$$L^{(m)} = \frac{N-1}{m^3} \sum_{t=1}^{m} \frac{\sum_{n=1}^{N''(m,t)} \left| Y_{t+nm} - Y_{t+(n-1)m} \right|}{N''(m,t)} \tag{21}$$

known to scale as m^{2-H} (where $N''(m,t) = \lfloor (N-t)/m \rfloor$ and the exponent is called the *fractal dimension* of the process X_t). Yet another interesting method consists in applying a least squares fit to m-long parts of the sample path of Y_t and estimating the minimum mean square error. For a class of self-similar processes, this error scales as m^{2H}. Finally, we mention the *rescaled adjusted range* (*R/S*) *statistic*, the most historic of all (instrumental in the discovery by H.E. Hurst). First, *R/S*(m) is calculated for $m \leq N$ as the diameter of the set $\{\Delta Y_1, ..., \Delta Y_m\}$, where ΔY_t is the deviation of the accumulated traffic process Y_t from its estimated mean, normalised w.r.t. the standard error estimate from $\{X_1, ..., X_t\}$. Next, using a number of non-overlapping sample series $X_1, ..., X_N$ (possibly successive blocks of observed samples), an estimate of $E\{R/S(m)\}$ is computed, known to scale as m^H.

6.2 Periodogram-based methods

By (9), the Hurst parameter can be deduced from the spectral density estimate, or *sample periodogram*, of $X_1, ..., X_N$ plotted against frequency near the origin. In practice, a least squares fit to the periodogram is biased toward higher frequencies (frequencies tend to "gravitate to the right" on a log scale); accordingly, modifications of the periodogram method are proposed e.g., the periodogram is first averaged over equal-size intervals on the log-frequency scale or the frequency range for the least squares fit is cut off close to the origin. The resulting H-estimates are nevertheless criticised for often being inaccurate.

Whittle's estimator is of maximum-likelihood type, where the process X_t presumably matches a family of self-similar model processes parameterised by the values of H e.g., FARIMA(p, d, q) or Fractional Gaussian Noise [8,18], and the resulting H-estimate is the one that gives the best match. Whittle's integral, given by

$$W(H) = \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega, H)} d\omega\,, \tag{22}$$

serves as the match criterion (subject to minimisation), with $S(\cdot)$ denoting the sample periodogram and $S'(\cdot, H)$ the model process' spectral density e.g., $S'(\omega, H) = [2\sin(\omega/2]^{-(2H-1)}/2\pi$ for FARIMA(0, d, 0), where $d = H - 0.5$.

7. Self-similarity detection using TSA

In order to develop a tool for data traffic analysis geared to self-similarity, a research project was undertaken at the Department of Marine Electronics of the Maritime Institute in Gdańsk, with the aim to

- adapt the available UNIX software and conduct measurements to produce traces of actual data traffic on a 10 Mbit/s LAN segment, and
- implement a number of the above methods of the Hurst parameter estimation in a single software package to obtain H-estimates for the measured traffic.

The software package, called TSA for *Time Series Analysis*, was written in Visual C++ as an MDI application for Windows NT. Its main functions include

- import of observed or synthetic (software-generated) traffic traces as text files,
- sampling the imported traces to produce $X_t^{(m)}$ with various levels of aggregation,
- computation of marginal distributions and autocovariance functions,
- computation of variance-time, R/S and IDC plots, periodograms and Whittle's integrals, and creation of the corresponding log-log plots,
- application of least square fits to the log-log plots or parts thereof along with the computation of R^2, and
- shuffling the imported traces by rearranging blocks of data in a random order to remove long-range dependence.

Using the *pcap* library, a separate program was written to record the data traffic at a specified network interface and output a trace file. A few dozens of measurement experiments were conducted, each resulting in an approximately 20-Mbyte file (corresponding to a little less than 90 minutes of traffic), of which a March 30, 1999, file was selected for further analysis. As a reference, a synthetic Poisson traffic trace was produced and verified to yield H-estimates about 0.5.

The marginal distribution of the measured traffic, displayed in Fig. 6, differs noticeably from Poissonian; the implied mean traffic rate is about 1 Mbit/s amounting to 10% of the available bandwidth. Fig. 7 shows the log-log variance-time plot and the range of time scales for which a least aquares fit was applied yielding a slope of –0.186. This corresponds to $H = 0.907$ ($R^2 \approx 0.95$). The limiting time scales were chosen so as to disregard the short-range dependence on the one hand and avoid the inaccuracy due to too few samples of $Y_t^{(m)}$ on the other. A least squares fit applied to the R/S statistic (Fig. 8) yields $H = 0.890$ ($R^2 \approx 0.97$) and when applied to the IDC plot for the range of time scales shown in Fig. 9, $H = 0.921$ ($R^2 \approx 0.99$). Two periodogram-based methods, the periodogram plot (Fig. 10) and Whittle's method (Fig. 11) yield $H = 0.850$ ($R^2 \approx 0.76$) and $H = 0.860$, respectively. In the latter, Whittle's integral is plotted against d using FARIMA(0, d, 0) [18] as the reference self-similar traffic model, yielding a minimum at $d = 0.360$.

In spite of some discrepancies between the obtained H-estimates, the analysed traffic can safely be regarded as self-similar with H somewhere in the range (0.85, 0.92) (note the high degree of self-similarity at a relatively low mean traffic rate). Thus LAN traffic traces, verified to be self-similar by the TSA tool, can serve as the input traffic in computer simulation models geared to the prediction of performance of large file transfer in realistic traffic conditions.

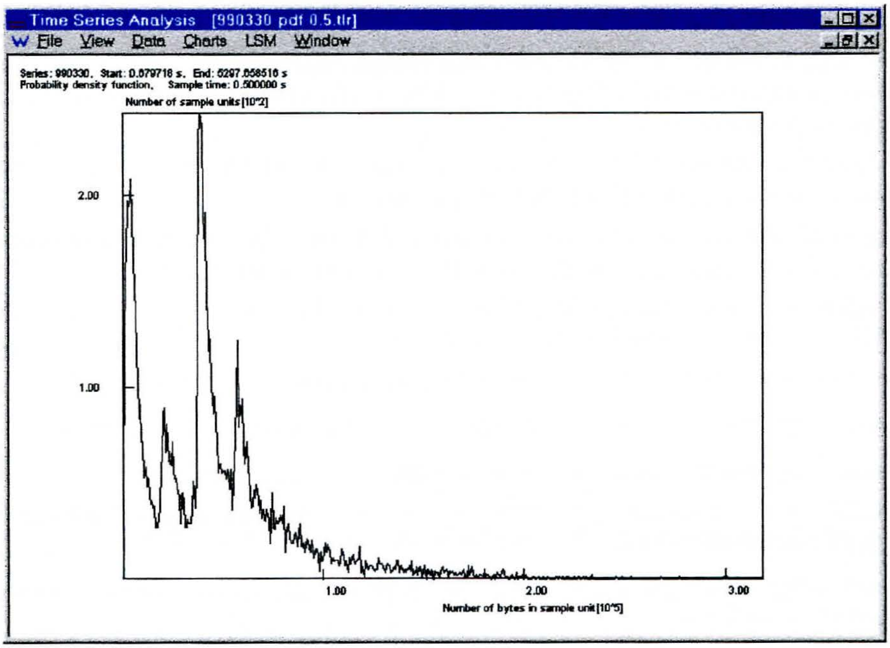
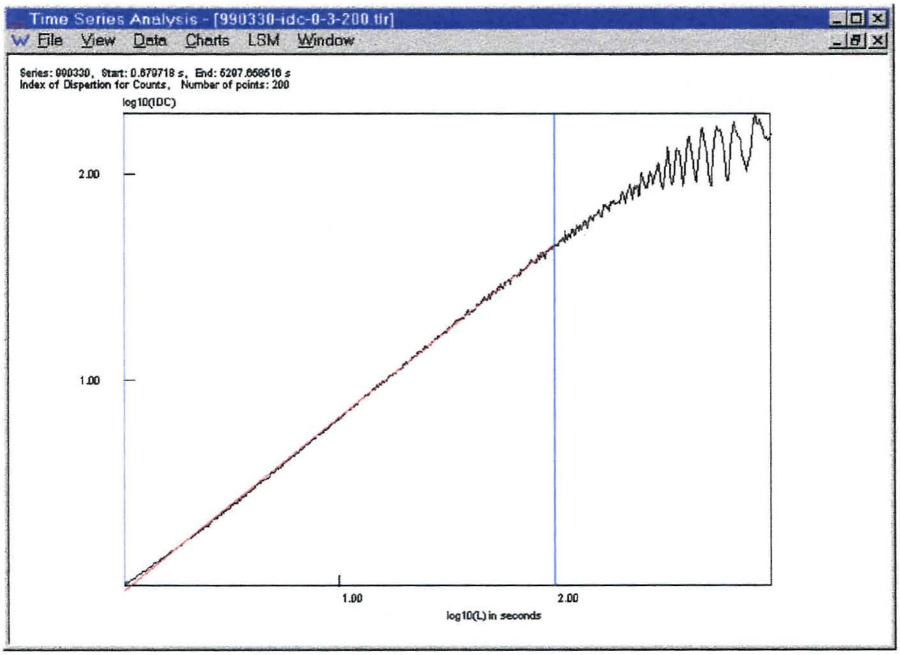
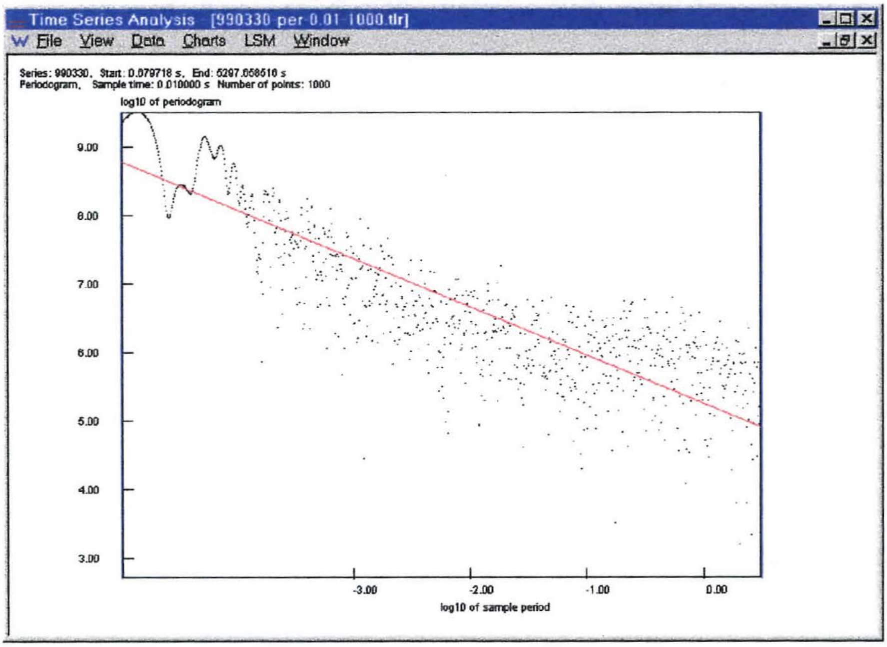


Fig. 6. Marginal distribution of the traffic rate



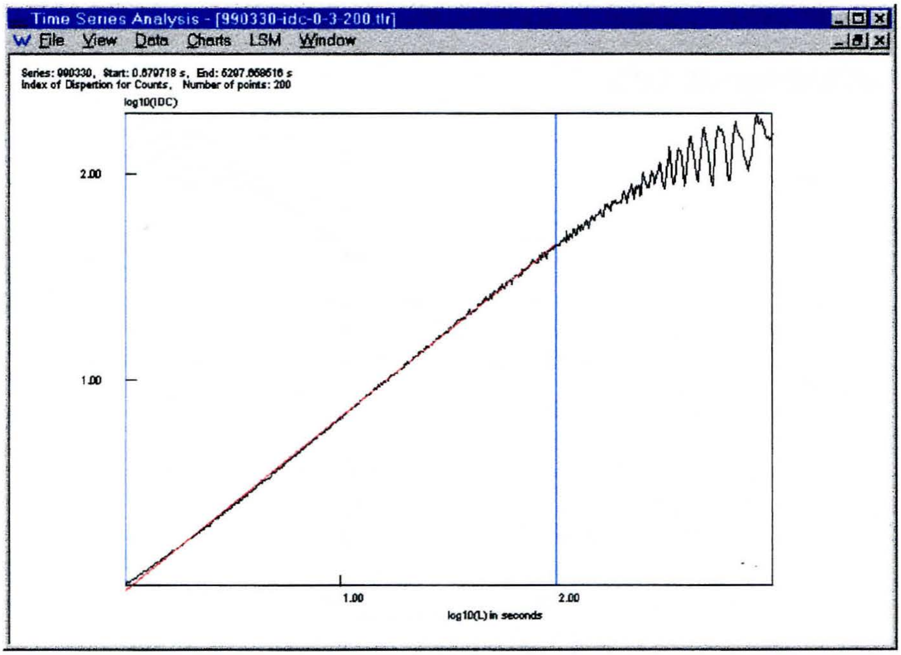
Series: 990330, Start: 0.079718 s, End: 5297.658516 s
Index of Dispertion for Counts, Number of points: 200

Fig. 7. Variance-time plot



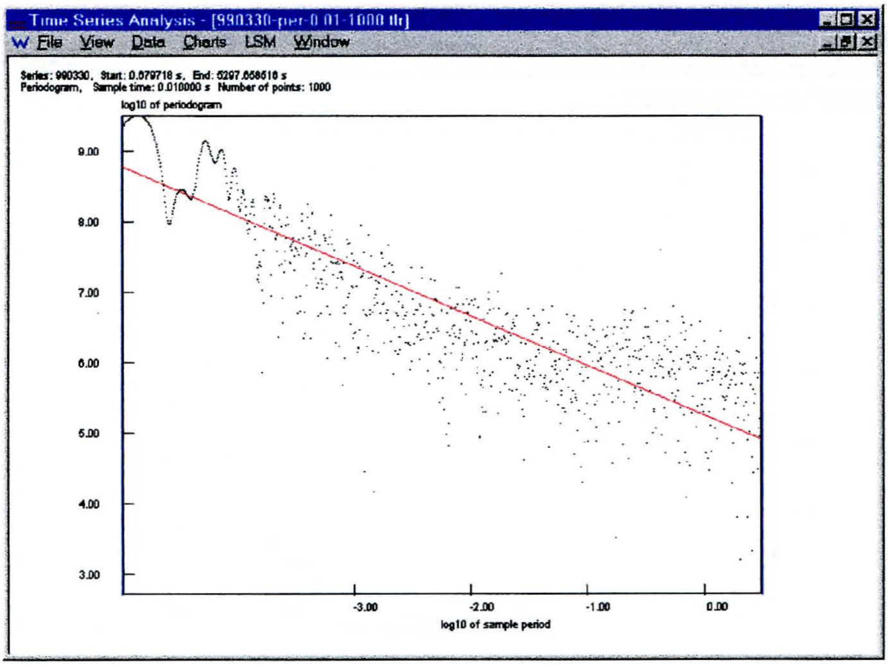
Series: 990330, Start: 0.079718 s, End: 5297.658516 s
Periodogram, Sample time: 0.010000 s Number of points: 1000

Fig. 8. R/S plot



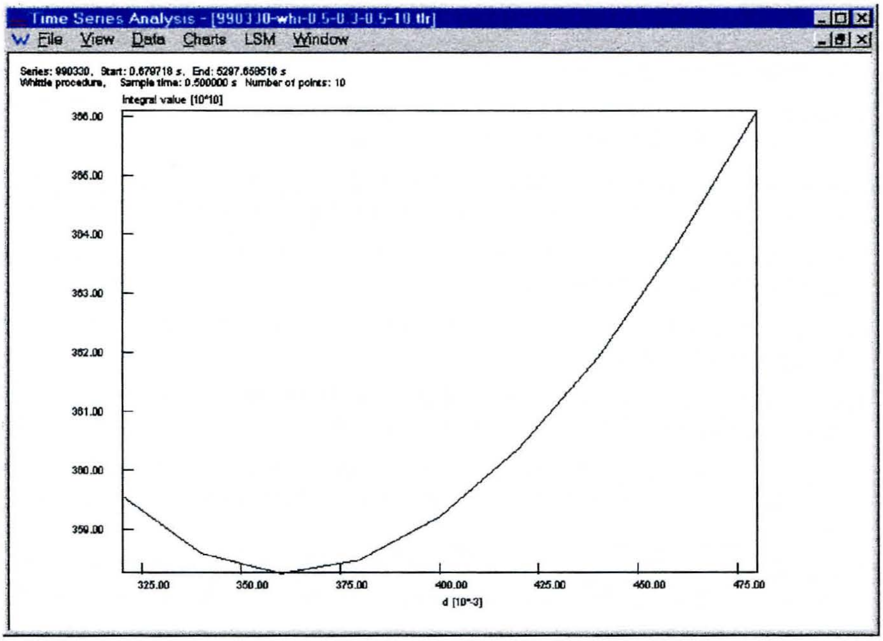
```
Series: 990330,  Start: 0.079718 s,  End: 5297.658516 s
Index of Dispertion for Counts,   Number of points: 200
```

Fig. 9. IDC plot



```
Series: 990330,  Start: 0.079718 s,  End: 5297.658516 s
Periodogram,   Sample time: 0.010000 s   Number of points: 1000
```

Fig. 10. Periodogram plot



Series: 990330, Start: 0.679718 *s*, End: 5297.659516 *s*
Whittle procedure, Sample time: 0.500000 *s* Number of points: 10

Fig. 11. Minimisation of Whittle's integral

References

[1] Addie, R.G. and Zuckerman, M., 1998, *Broadband Traffic Modelling: Simple Solutions to Hard Problems*, IEEE Comm. Mag. 8.

[2] Crovella, M.E. and Bestavros, A., 1997, *Self-similarity in World Wide Web Traffic: Evidence and Possible Causes*, IEEE/ACM Trans. on Networking 5(6).

[3] Erramilli, A., Pruthi, P. and Willinger, W., 1994, *Application of Fractals in Engineering for Realistic Traffic Processes*, Proc. 14th Int. Teletraffic Congress, Sophia-Antipolis, France.

[4] Fiorini, P.M., Crovella, M. and Lipsky, L., 1998, *On the Connection between Power-Tail Distributions and Long-Range Dependencies*, available from crovella@cs.bu.edu

[5] Gajewski, J., and Staśkiewicz, A., 1998, *Validation of Hydrodynamic Models of the Baltic Sea in Polish Waters – HIROMB as an Example*, Bull. Mar. Inst., XXV, (2).

[6] Jędruś, S., 1999, *Modelling of packet traffic intensity in computer networks using multi-fractal measures*, Ph.D. dissertation, Inst. of Theoretical and Applied Informatics, Gliwice. [In Polish].

[7] Karlsson, P. and Arvidsson, A., 1999, *Traffic Modelling of TCP/IP over ATM*, draft for 16th Int. Teletraffic Congress, Edinburgh, UK.

[8] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson D.V., 1993, *On the Self Similar Nature on Ethernet Traffic (Extended Version)*, draft, ftp://ftp.bellcore.com/pub/world/wel/tome.ps.Z.

[9] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson D.V., 1994, *On the Self Similar Nature on Ethernet Traffic (Extended Version)*, IEEE/ACM Trans. on Networking 2(1).

[10] Mandelbrot, B., 1983, *The Fractal Geometry of Nature*, Freeman, New York.

[11] Molnar, S. and Vidacs, A., 1997, *On Modelling and Shaping Self-similar ATM Traffic*, Proc. 15th Int. Teletraffic Congress, Washington D.C.

[12] Norros, I., 1995, *On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks*, IEEE J. Selected Areas In Comm., 13 (6).

[13] Park, K., Kim, G. and Crovella, M., 1997, *On the Effect of Traffic Self-similarity on Network Performance*, Proc. SPIE Int. Conf. on Performance and Control of Network Systems.

[14] Paxson, V. and Floyd, S., 1995, *Wide Area Traffic: The Failure of Poisson Modelling*, IEEE/ACM Trans. on Networking 3(3).

[15] Roberts, J., Mocci, U. and Virtamo, J. (eds.), 1996, *Broadband Network Teletraffic*, Springer-Verlag, Berlin Heidelberg.

[16] Ryu, B., 1997, *Fractal Network Traffic Modelling: Past, Present and Future*, Proc. 35 Allerton Conf. on Comm., Control and Computing, http://www.wins.hrl.com/people/ryu/fsndpMS95.ps.gz

[17] Ryu, B. and Lowen, S.B., 1998, *Point Process Models for Self-Similar Network Traffic with Applications*, [in] Neuts, M. (ed.), *Stochastic Models* 14 (3), http://www.wins.hrl.com/people/ryu/sm98.ps.gz

[18] Taqqu, M.S., Teverowsky, V. and Willinger, W., 1995: *Estimators for Long-Range Dependence: an Empirical Study*, Fractals 3(4), http://math.bu.edu/people/murad/pub/estimators-posted.ps

[19] Taqqu, M.S., Willinger, W. and Sherman, R., 1997, *Proof of a Fundamental Result in Self-Similar Traffic Modelling*, Computer Comm. Review, 27, http://math.bu.edu/people/murad/pub/ccr97-onoff-posted.ps

[20] Taralp, T., Devetsikiotis, M. and Lambadaris, I., 1998, Traffic Characterisation for QoS Provisioning in High-Speed Networks, Proc. Hawai'i Int. Conf. On System Sciences, ftp://www.sce.carleton.ca/pub/bbnlab/doc026.pdf

[21] Willinger, W., Taqqu, M.S., Sherman, R. and Wilson, D.V., 1997, *Self-similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*, IEEE/ACM Trans. on Networking 5(1).