

Criticisms of modelling packet traffic using long-range dependence

Richard G. Clegg, Raul Landa, Miguel Rio
Dept of Electrical & Electronic Engineering
University College London, UK
{rclegg,r.landa,m.rio}@ee.ucl.ac.uk

ABSTRACT

This paper criticises the notion that long-range dependence is an important contributor to the queuing behaviour of real Internet traffic. The idea is questioned in two different ways. Firstly, a class of models used to simulate Internet traffic is shown to have important theoretical flaws. It is shown that this behaviour is inconsistent with the behaviour of real traffic traces. Secondly, the notion that long-range correlations significantly affects the queuing performance of traffic is investigated by destroying those correlations in real traffic traces (by reordering). It is shown that the longer ranges of correlations are not important except in one case with an extremely high load.

1. INTRODUCTION

Since the seminal paper of Leland et al [6] it has been considered important that a statistical model of Internet traffic captures the phenomenon of Long-Range Dependence (LRD). In particular it has often been suggested that a model of Internet traffic must capture the Hurst parameter $H \in (1/2, 1)$ of real traffic. LRD is characterised by the unsummability of the autocorrelation function (ACF). It is often stated that this is an important characteristic for the queuing performance of the traffic.

A related topic is that of heavy-tailed distributions. A commonly suggested origin for the LRD in Internet traffic is the heavy-tailed distribution of traffic *on* periods.

DEFINITION 1. *A random variable X is heavy-tailed if, for all $\varepsilon > 0$ it satisfies*

$$\mathbb{P}[X > x] e^{\varepsilon x} \rightarrow \infty \quad \text{as } x \rightarrow \infty. \quad (1)$$

A specific functional form is usually assumed (and will be throughout this paper)

$$\mathbb{P}[X > x] \sim Cx^{-\beta}, \quad (2)$$

where $C > 0$ is a constant and $2 > \beta > 0$. The symbol \sim means asymptotically equal to. If $\beta < 1$ then $E[X]$ is infinite and therefore most models use $\beta \in (1, 2)$.

Suggested models for Internet traffic which generate LRD include fractional Gaussian noise and the related

fractional Brownian motion (fGn, fBm) [7], chaotic maps [4], wavelets [9, 8] and Markov modulated processes [1, 3]. Some of these models output a “traffic level” which represents the mean arrival rate in some notional time period but others are packet based models, that is they produce a model of packets and inter-arrival times. It is the latter class of models (including [4, 1, 3] which are covered by theorem 1 in this paper.

This paper criticises the notion that the long-range correlations in traffic are important to queuing in two ways. In section 2 it is shown that a class of models used to simulate traffic with LRD arising from heavy tails gives an infinite result when queued in infinite buffers. It is demonstrated in section 4 that this is at odds with the behaviour of real traffic. In section 5 real traffic traces are analysed again and reordered to break up correlations beyond a certain level. It is shown that this reordering does not affect the queueing behaviour of the traffic beyond a certain time-scale except when unrealistically high loads are used. The behaviour of the long-range dependent models (and in particular a certain class based on heavy-tails) is theoretically undesirable and fundamentally different to that of real traffic.

2. THEORETICAL RESULTS

Let $\{A_t : t \geq 0\}$ be an arrival process to a queue drained by a deterministic server which serves at a constant rate assumed w.l.o.g. to be one. The mean arrival rate λ is given by $\lambda = \lim_{T \rightarrow \infty} \int_0^T A(t) dt / T$ and it is assumed throughout that A_t is such that this limit exists and $\lambda \in (0, 1)$. Since the server rate is one then λ is equal to the utilisation ρ (the ratio of the rate at which work enters to the maximum rate at which it can be served). Let $\{Q_t : t \geq 0\}$ be the queue process where it is assumed that $Q_0 = 0$. Assume that the queue evolves according to

$$\frac{dQ_t}{dt} = \begin{cases} A_t - 1 & Q_t > 0 \\ \max(0, A_t - 1) & Q_t = 0. \end{cases}$$

Let $E[Q(s, t)] = \int_s^t E[Q_u] du / (t - s)$ where $t > s$ and

$E[\cdot]$ denotes expectation. Let $E[Q] = \lim_{t \rightarrow \infty} E[Q(0, t)]$ and note that this limit is not guaranteed to exist (and may tend to infinity). The mean arrival rate at time t is $\lambda_t = E[A_t]$ and the overall mean arrival rate $\lambda = \lim_{T \rightarrow \infty} \int_0^T \lambda_t / T dt$. If $\lambda > 1$ then $\rho > 1$ and the queue must eventually grow to infinity regardless of the details of the arrival process.

THEOREM 1. *Let $\{A_t : t \in \mathbb{R}_+\}$ be an ergodic, weakly stationary arrival process which can only take values $a > 1$ (on) and 0 (off). This is drained by a queue which drains at a fixed rate, w.l.o.g. assumed to be one. Let A_t be such that the arrival rate λ (and hence the utilisation ρ) is in $(0, 1)$. Let $\{X_n : n \in \mathbb{N}\}$ be the length of the n th on period and assume these are i.i.d. with a heavy-tailed distribution $\mathbb{P}[X_n > x] \sim x^{-\alpha}$ for $\alpha \in (1, 2)$ then as $t \rightarrow \infty$ the expected queue length (and hence the expected waiting time) is infinite.*

This theorem can be restated as: if an infinite buffer queuing model is driven by a single on/off source with i.i.d. heavy-tailed on periods then the expected queue length is either zero (if $a \leq 1$) or infinite. This remains true even if the utilisation $\rho \in (0, 1)$ defined as the proportion of the time the server is busy, is much less than one (the queue is empty for an arbitrarily high proportion of the time). It may seem paradoxical that a queue which is empty arbitrarily often can have an infinite expected length. However, this has parallels with the classical Pollaczek–Khinchine formula for an M/G/1 queue [5] where a server with an infinite variance in the service time has an infinite expected queue length even if the mean service time is arbitrarily small and the queue empty an arbitrarily large proportion of the time.

Note that for $\alpha < 1$ the mean length of an on period will not converge and such processes will not, in general, be useful for a queuing system. However, for $1 < \alpha < 2$ the mean length of an on period will be finite and such a process could be used to produce a time series with a known Hurst parameter.

PROOF. First consider a single on period followed by a single off period of such length that the entire queue has drained by the end of the off period. Consider the time period (t_1, t_2) where $Q_{t_1} = Q_{t_2} = 0$, consisting of an on period $(t_1, t_1 + X)$ (where $aX < (t_2 - t_1)$) and an off period $(t_1 + X, t_2)$. Within the period (t_1, t_2) the queue peaks at time $t_1 + X$ when $Q_{t_1+X} = (a-1)X$ and drains completely by time $t_1 + aX$ after which the queue is zero until t_2 . It can be readily seen that, since the queuing process is triangular in shape (rising at rate $a-1$ during the on period and falling at rate 1 during the off period), then $\int_{t_1}^{t_2} E[Q_u] du = (a-1)aX^2/2$ and $E[Q(t_1, t_2)] = (a-1)aX^2/2(t_2 - t_1)$.

Now consider some time period (t_1, t_2) again where $Q_{t_1} = Q_{t_2} = 0$. Let this period contain exactly two on

periods of lengths X_1 and X_2 where $a(X_1 + X_2) < (t_2 - t_1)$. It is clear that $\int_{t_1}^{t_2} E[Q_u] du \geq (a-1)a(X_1 + X_2)/2$ with equality occurring only when the queue empties completely between the two on periods. This argument can be trivially extended to n on periods of lengths X_1, X_2, \dots, X_n all occurring within (t_1, t_2) with $Q_{t_1} = Q_{t_2} = 0$. The mean queue size is minimised if the on periods are such that the generated queues do not overlap.

Consider the process A'_t which has the same mean arrival rate and is the process A_t reordered in time according to the following rules:

- on periods occur in the same order and have the same length as A_t with the first on period starting at $t = 0$,
- an on period of length X_i is followed by an off period of length exactly $X_i(a/\lambda - 1)$.

This off period is long enough that the queue has always completely drained before the end of the off period (since $\lambda < 1$). It can easily be shown that such a reordering is possible since the on periods are of exactly the same length in the same order and the off periods have the same mean length.

Let Q'_t be the queue process for A'_t (assuming the same server process). Clearly $E[Q'] \leq E[Q]$ since the queues due to A'_t never overlap (with equality occurring only when the queues never overlap in A_t either).

It can be shown that

$$\begin{aligned} E[Q'] &= \lim_{N \rightarrow \infty} \frac{\sum_{i=1}^N \int_{t_i}^{t_{i+1}} Q'_t dt}{t_{N+1}} \\ &= \lim_{N \rightarrow \infty} \frac{a(a-1) \sum_{i=1}^N X_i^2}{2 \sum_{j=1}^N X_j} \end{aligned}$$

Taking expectations a second time gives

$$\begin{aligned} E[E[Q']] &= E[Q'] = \lim_{N \rightarrow \infty} \frac{a(a-1) \sum_{i=1}^N E[X_i^2]}{2 \sum_{j=1}^N E[X_j]} \\ &= \frac{a(a-1)E[X^2]}{2E[X]}, \end{aligned}$$

where the last equality follows since the X_i are i.i.d.

$E[Q']$ is a lower bound for $E[Q]$. $E[Q]$ does not converge if $E[X^2]$ does not converge. If $1 < \alpha < 2$ then $E[X]$ is finite but $E[X^2]$ is not and the expected queue is infinite. The result follows. \square

A similar result holds for discrete time on-off arrival processes $\{A_n : n \in \mathbb{N}\}$ with heavy-tailed on periods. However, the result is not true, for example, if the queue is driven by two or more heavy-tailed sources each of which has an arrival rate less than one but together having an arrival rate over one. Processes such as fractional Gaussian noise exhibit LRD but have a finite expected

queue length in an infinite buffer. In fact the theorem has an obvious corollary.

It should be noted that heavy-tailed arrival processes of the type from Theorem 1 which give rise to a finite value of $E[Q]$ are possible but paradoxically these have an arrival rate $\lambda = 0$. For example, let $q > 0$ and let every *on* period of length X_i (at rate a) be followed by an *off* period of length $\max((a-1)X_i, (a-1)aX_i^2/2q - X_i)$. The queue drains completely in every *off* period and the mean queue size for that *on* period and *off* period is at most q (with equality attained when $((a-1)aX_i^2/2q - X_i) \geq (a-1)X_i$) and therefore $E[Q] \leq q$. However, in the heavy-tailed case, $E[X_i^2]$ does not converge and hence neither does the expected length of an *off* period. This implies a proportion of time in the *off* period tending to one and a mean arrival rate λ of zero.

It might still be argued that real traffic traces have this property but the outcome of infinite queue size is not seen in real life because they are fed to a finite sized array. This possibility will be investigated in section 4.

3. SIMULATION FRAMEWORK FOR THIS PAPER

CAIDA data: This data set is taken from a trace approximately an hour long. It is referred to on the CAIDA website ¹ as

20030424-000000-0-anon.pcap.gz and was captured on the 24th April 2003. It was captured on an OC48 link with a rate of 2.45 Gb/s. The first 550,000 packets are used in the analysis here. This trace has a relatively low Hurst parameter $H = 0.6$ (see [2]).

Bellcore data: This is a much studied data set and, while certainly not representative of modern traffic, it is included as one of the original traces from [6]. The data here is taken from an August 1989 measurement referred to as BC-pAug89.TL. The data was collected on an Ethernet link². The first 1,000,000 packets are studied here. This trace has a relatively high Hurst parameter $H = 0.8$ (see [2]).

The simulations used in this paper are all based upon an extremely simple queuing model. Packets arrive in a FIFO buffer which never drops packets. The buffer has a given bandwidth b (in bytes/second) – which can be adjusted to give a specific level of utilisation. While the absolute level of the queue changes, the results presented here are not very sensitive to this parameter. A packet of length l bytes takes l/b seconds to depart the queue. If Q_t is the queue length in packets at time t and the simulation runs until time T then the mean queue \bar{Q} is evaluated as $\bar{Q} = \int_0^T Q_t/T dt$ (this integral can be evaluated exactly since Q_t is a constant between

¹See <http://www.caida.org/data/passive/> for more information about this trace.

²See <http://ita.ee.lbl.gov/html/contrib/BC.html> for more information about this traffic

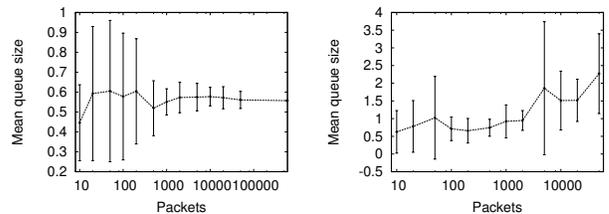


Figure 1: Mean queue size versus number of packets for the CAIDA data.

arrivals and departures from the queue).

This simple model has been used by the first author to assess several the queuing performance of several models which attempt to replicate the statistical nature of Internet traffic. This work is reported in [2]. None of the long-range dependence based models replicated the queuing performance of the real traffic traces they were tuned to match. Such a simple model is, of course, open to much criticism. It does not account for TCP feedback mechanisms. However, if the question being addressed is about an “open loop” model of traffic and whether it replicates the characteristics of real traffic then we should expect both to have the same behaviour at a queue.

4. SIMULATION RESULTS ON REAL VERSUS HEAVY-TAILED TRAFFIC

The first simulations here consider the theoretical results presents in section 2. The first results show how the infinite expected queue size in the model reveals itself in simulation results. Obviously any experiment with a model of the form in Theorem 1 will produce a finite value for \bar{Q} the mean queue size. However, this finite value will increase as the model is run for longer and longer (up to the numerical accuracy of the model). The value of \bar{Q} generated will (in theory at least – in practice the finite accuracy of computers limits this) increase as the runtime increases. The question may be asked if this is true of real data.

The experiment performed here is to take different sized samples of the real data and to queue those samples with the model from the previous section. The behaviour of the mean queue length versus sample size is investigated. Experiments on LRD are notorious for their high (often theoretically infinite) variability. Here ten replications are performed for each size are used and the mean plotted. In addition error bars of the size of the standard deviation (one standard deviation above and below the mean) are added (standard confidence interval techniques are not applicable in the case of LRD data).

Figure 1 (left) shows the results for the CAIDA data.

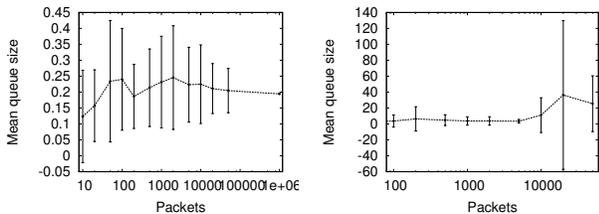


Figure 2: Mean queue size versus number of packets for the Bellcore data.

The x axis shows the number of packets in the sample and the y axis the mean queue size (or rather the mean of the ten means for the ten experiments with that sample size). Since the queue is assumed to start empty then very small sample sizes would naturally have a smaller expected queue. However, beyond this, figure 1 shows no clear influence of the sample size on the expected queue. As the samples get larger the standard deviation bounds from the ten experiments gets smaller. It should be noted that the final (rightmost) point is the whole data set and no error bars can be included.

Figure 1 (right) shows the same experiment but performed on a simulated data set with the same mean arrival rate and the same Hurst parameter using the techniques described in [2] (the Wang model from that paper). The simulation does not well reflect the queuing performance of the trace and this is because of theorem 1 which applies here. In contrast with figure 1 the mean queue size increases with the number of packets in the sample. In addition the error bars (representing one standard deviation either side of the mean) stay as large or become larger.

Figure 2 (left) shows the same experiment for the Bellcore data. This data has a higher Hurst parameter and the outcome is less clear. However, remembering that the last point has no error bars (representing the entire data) this figure is not consistent with the idea that the mean queue rises as the length of the sample rises apart from in the early part of the plot. (The rise in the mean and the larger error bars in the center of the plot coincides with a single very large burst of a particular duration).

Figure 2 (right) shows simulated data with the same Hurst parameter and same mean arrival rate as the Bellcore data. As in figure 1 (right) and in accordance with theorem 1 the mean queue size rises with the number of packets simulated (although the connection is certainly not unambiguous). As can be seen, these simulations can be problematic to work with and a researcher looking only at the early part of the graphs could be convinced that they had used sufficiently many packets for the simulation to converge to a good estimation of the mean queue length.

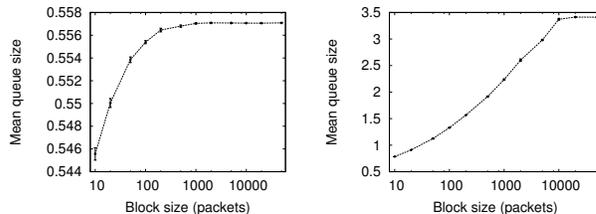


Figure 3: Mean queue size versus blocksize for CAIDA data.

It may be thought that the problem may be connected with the fact that the LRD based methods dramatically overestimated levels of queuing. However, repeating the experiment with lower bandwidth on the real data for both Bellcore and CAIDA traces does not dramatically alter the shape of the graph although obviously the mean queue level increases.

5. SIMULATION RESULTS ON REORDERING OF REAL TRAFFIC TRACES

This section takes a different approach by deliberately destroying correlations in the data to see which scales of correlation are important to the queuing properties. It is often stated that LRD is an extremely important property for queuing in real data. If this is the case, then deliberately truncating the correlation beyond a certain scale should have important effect on the queuing.

The experiment performed in this section is to take a certain blocksize B and to split the data into blocks each containing B packets (and associated delays). The order of these blocks is then randomised so no correlation can persist beyond B packets. The entire trace is then queued and the mean queue length recorded. Again ten replications are performed to assess the repeatability.

Figure 3 (left) shows this experiment on the CAIDA data. Again there are ten repetitions of each blocksize and the graph shows mean of the means and the standard deviation of the means above and below. Note the small scale on the y axis. Even for very small block sizes the variation in the queuing performance is not great. Beyond a block size of 1,000 the correlation seems unimportant to the queuing performance and the resultant mean queue size is the same to several decimal places.

By contrast, in figure 3 (right) for the LRD simulation all scales of correlation theoretically affect queuing performance (within the bounds implied by the fact that only a finite sample of data is used). Here, the relationship between correlation and queuing is clearly shown. Correlations of block sizes up to 10,000 packets are important to the queuing performance of the simulated data.

Figure 4 shows the same experiment on the Bellcore

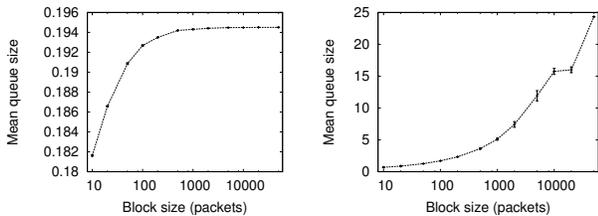


Figure 4: Mean queue size versus blocksize for Bellcore data.

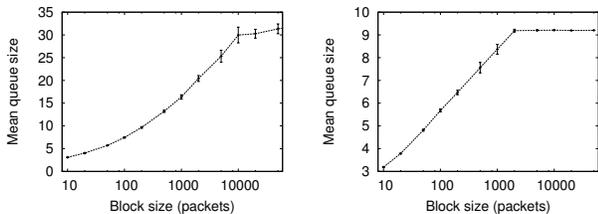


Figure 5: Mean queue size versus blocksize for real data with high load.

data. Again the LRD method used drastically over predicts the level of queuing compared to the real data when the same mean arrival rate and Hurst parameter is used. Again the same result is seen only a small difference in the amount of queuing when the correlations are broken up. The correlations of long time scale are unimportant in the real data but important in the artificial data.

Again the question might be asked would the same conclusion hold true for the real data if the bandwidth used for the experiment were reduced. Figure 5 (left) shows this for the Bellcore data with the bandwidth reduced so the queue occupancy is 5 (right) shows this for the CAIDA data (with an even higher queue occupancy of 0.62). In the CAIDA data there seems to be a very clear transition from important correlation scales (at lengths below 2000 packets) and unimportant correlation scales (above). Indeed this transition is so marked it raises questions about whether something in the data capture process or in the system itself would cause this. In the case of the Bellcore data, more timescales are important but not nearly so many as in the long-range dependent data of figure 4 (right).

The conclusion of this section is clear. The claim that correlations over long scales is important to queuing behaviour is not true of the CAIDA data and arguably true of the Bellcore data only when the system occupancy is extremely high.

6. CONCLUSIONS

This paper criticises long-range dependence as a use-

ful model for packet traffic. Firstly, a theoretical problem with a class of models used to simulate LRD is shown. This class of models predicts either no queuing or an infinite expected queue length when fed into an infinite buffer. At the very least, experimenters should be aware of this problem to ensure that simulations are not affected by it (the answer given by the simulation is a product of the runtime of the simulation rather than a stable reflection of queuing performance). The effect is shown to be different to the queuing performance of real traffic. In the second part of the paper it is shown using simulated queuing on real traffic that for real traffic traces long-range correlations are not important for queuing behaviour except with extremely high traffic. These results should be replicated on more traffic traces to work out how general this conclusion is.

7. REFERENCES

- [1] M. Barenco and D. K. Arrowsmith. The autocorrelation of double intermittency maps and the simulation of computer packet traffic. *Dynamical Systems*, 19(1):61–74, 2004.
- [2] R. G. Clegg. Simulating internet traffic with Markov-modulated processes. In *Proceedings of UK Performance Engineering Workshop*, 2007. Available online at http://www.richardclegg.org/pubs/rgc_ukpew2007.pdf.
- [3] R. G. Clegg and M. M. Dodson. A Markov based method for generating long-range dependence. *Phys. Rev. E*, 72:026118, 2005.
- [4] A. Erramilli, R. P. Singh, and P. Pruthi. Chaotic maps as models of packet traffic. *Proc. 14th Int. Teletraffic Cong., 6-10 June 1994 North-Holland Elsevier Science B.V.*, 1:329–338, 1994.
- [5] A. Y. Khinchin. Mathematical theory of stationary queues. *Mat. Sbornik*, 39:73–84, 1932.
- [6] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. In D. P. Sidhu, editor, *Proc. ACM SIGCOMM*, pages 183–193, San Francisco, California, 1993.
- [7] V. Paxson. Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. *Computer Comm. Rev.*, 27:5–18, 1997.
- [8] R. H. Riedi. Multifractal processes. In P. Doukhan, G. Oppenheim, and M. S. Taqqu, editors, *Theory And Applications Of Long-Range Dependence*, pages 625–716. Birkhauser, 2003.
- [9] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Special Issue On Information Theory*, 45:992–1018, 1999.