A Study of the Daily Variation in the Self-Similarity of Real Data Traffic

Matthew Roughan and Darryl Veitch [a] *

[a]Software Engineering Research Centre, RMIT University,
Level 3, 110 Victoria St, Carlton, Vic 3053, Australia
E-mail: {matt,darryl}@serc.rmit.edu.au.

In the last few years the discovery of the self-similar nature of packet traffic has highlighted the need for the estimation of parameters quantifying scaling phenomenon, such as the Hurst parameter. An important practical question concerns the variation of the Hurst parameter with time. The on-line real-time version of the wavelet based estimator of Abry and Veitch, which allows data of unlimited length to be analysed, was used to collect almost 6 months of Ethernet data. The question of the diurnal variation of the Hurst parameter is investigated in detail, and its correlation with network load.

1. Introduction

In the last few years the discovery of the *self-similar* or *scaling* nature of many kinds of packet traffic [7,9] has inspired a small revolution in the way study of high-speed traffic. No single model is accepted as definitive, but the *Hurst* parameter H, which describes the degree of self-similarity, holds a central place in the description of such traffic. Its accurate measurement is therefore of considerable importance for the provision of quality of service as well as for capacity planning.

In current models of traffic, H is a constant describing the scaling nature of traffic which is deemed to be stationary. Naturally in real data this assumption holds only approximately, or perhaps not at all. For example diurnal variation in load is a recognised feature of traffic in most contexts, is this also true for H? Some central questions are:

- What are the empirical variations in H estimates observed over timescales including minutes, hours, days and weeks?
- Over what time-scales can H be considered to be constant? (in other words to what extent does it make sense to speak of H 'varying', given that it is a parameter describing stationary data?)
- What systematic variations in H are present, for example is there a consistent diurnal pattern?
- What structural causes can be found explaining the variations of H, for example in what way is it connected to load, if at all?

In this paper we present a preliminary investigation into the variation of H in Local Area Network (LAN) traffic. A central aim is to gain experience in, and develop methods

*The authors gratefully acknowledge the support of Ericsson Australia.

for dealing with the time variation of H. This is necessary if fractal models are ever to bridge the gap between useful but over idealized theoretical tools, and workable practical descriptions of real traffic. A knowledge of the time variation of H is also of interest in its own right, as LAN traffic is an important component of the traffic in wide area networks.

In order to address such questions, two essential problems need to be resolved. The first is that of reliable and practical H estimation. Many estimation methods suffer from poor statistical performance, and/or high computational complexity. Recent work based on wavelets however has provided a semi-parametric estimator for H, referred to here as the Abry-Veitch (AV) estimator, which gives unbiased estimates with low variance together with significant computational advantages, notably a run time complexity of only $O(n)$. In the present context, however, it is essential that the estimator behave robustly in realistic non-stationary environments. The AV estimator has important avantages in this area also, for example it can distinguish between true scaling behaviour and certain kinds of non-stationary which lead most estimators to erroneously conclude that fractal behaviour exists when in fact it does not. Properties of the estimator are summarized below in Section 2, and can be found in [14] (see also [2,3]). Details and discussion of the robustness properties and related stationarity issues can be found in [3,14,11,13].

The second problem is that of the difficulties posed by the need for data collection and analysis over extended periods. In [12] an on-line, real-time implementation of the estimator on inexpensive hardware was described. It was applied successfully to the real-time measurement of 10 Mbps Ethernet traffic and has since been extended to 155 Mbps Asynchronous Transfer Mode (ATM) traffic. The on-line implementation allows measurements to be made over essentially unlimited time periods, withour extensive memory requirements. Using it, we have no difficulty in performing continuous monitoring over a period of months. Another aim of the paper is to illustrate the potential of such an ability, and how the long term measurements it provides can be profitably used.

We focus on the diurnal or daily variation in traffic parameters. Section 4 shows results based on nearly 6 months of Ethernet data. The results of the analysis are strongly suggestive of a number of features that are likely to be applicable to more than just our LAN, as they can be plausibly explained in terms of human usage of the network. Briefly the important features are:

- A weak diurnal cycle in load linked with use of the network by humans, and with backup loads. The natural variations during the day may be of larger magnitude than diurnal variations.
- Differentiated diurnal cycle in load and H value on weekends (as opposed to weekdays), again due to the different usage of the people on the system.
- A weak diurnal variation in the nature of the data as a function of scale, and not merely a change in the Hurst parameter.
- A dependence in the qualitative type of scaling observed on the nature (human or machine generated) of the traffic on the network.
- Variation in the scaling behaviour on time scales of the order of 1–4 hours which may be of larger magnitude than diurnal variations.

A conclusion of practical importance based on the above is that continuous measurements of the parameters of long-range dependence are required for real time network needs such as call admission control or applications which uses rate adaptation, in addition to the network load measurements. This is because, although diurnal variations exist, local changes are also very significant and may dominate. Our results suggests that in the case of Ethernet traffic that such measurements should be based around time scales from 1 to 4 hours: any smaller and the data sets are not sufficient to obtain an accurate estimate, any larger and the parameters may change substantially over the measurement interval.

2. The Abry-Veitch (AV) estimator

The basic theoretical framework of this paper is as follows. The time varying *rate* $x(t)$ of traffic is the key data, and we model it as a stationary stochastic process. Basic features of this process are its mean $\mu_x = E[x]$, variance $\sigma_x^2 = E[(x-\mu_x)^2]$, and correlation function $\gamma_x(k) = E[(x(t+k)-\mu_x)(x(t)-\mu_x)]$. The self-similar properties of traffic manifest themselves in a particular form of $\gamma_x(k)$, namely a decrease with lag k so slow that the sum of all correlations downstream from any given time instant is always appreciable. The past therefore exerts a long term influence on the future, exaggerating the impact of traffic variability and rendering statistical estimation problematic. This phenomenon is known as *Long Range Dependence* (LRD), and is commonly defined by $\gamma_x(k) \sim c_\gamma |k|^{-(1-\alpha)}$, $\alpha \in (0,1)$, or equivalently as the power-law divergence at the origin of its power spectrum: $f_x(\nu) \sim c_f|\nu|^{-\alpha}$, $|\nu| \to 0$. The Hurst parameter describes the (in practice, asymptotic) self-similarity of the cumulative traffic process $\int_0^t x(s)ds$ while the LRD parameter α describes the rate process $x(t)$. It is nonetheless common practice to speak of H in relation to LRD via the relation $H = (1+\alpha)/2$, and we follow this convention here.

In [14] a semi-parametric joint estimator of (α, c_f) is described based on the *Discrete Wavelet Transform* (DWT) [5]. Wavelet transforms in general can be understood to be a more flexible form of Fourier transform, where $x(t)$ is transformed into a time-scale wavelet domain rather than a frequency domain. Thereby allowinging simultaneous observation of a time series over different scales a, whilst retaining the time dimension of the original data. No information is lost if we sample the continuous wavelet coefficients at a sparse set of points in the time-scale plane known as the *dyadic grid*, defined by $(a,t) = (2^j, 2^j k)$, $j,k \in \mathbb{N}$, leading to the DWT with discrete coefficients $d_x(j,k)$ known as *details*. Intuitively, the dyadic grid samples the wavelet domain at a resolution appropriate to the scale. The *octave* j is simply the log base 2 of scale $a = 2^j$. For finite data of length n, j will vary from $j = 1$, the finest scale in the data, up to some $j_{\max} \approx \log_2(n)$.

The estimator has excellent computational properties due to the fast 'pyramidal' filter-bank algorithm [5] for calculation of the discrete wavelet transform, which has a complexity of only $O(n)$. The number of wavelet coefficients $d_x(j,k)$ thus generated is also of order n, and subsequent computations of the estimator have only this complexity.

The main feature of the wavelet approach which makes it so effective for the statistical analysis of scaling phenomenon such as LRD is the fact that the wavelet basis functions themselves possess a scaling property, and therefore constitute an optimal 'co-ordinate system' from which to view such phenomena. The main practical outcome is that the LRD in the time domain representation is reduced to residual **short** range correlation in the wavelet coefficient plane $\{j,k\}$, thus removing entirely the special estimation difficulties.

We can now outline the estimator as consisting of the following four stages:

1. **Wavelet decomposition:** A discrete wavelet transform of the data is performed, generating the details $d_x(j,k)$.
2. **Detail variance estimation:** At each fixed octave j the mean squared detail μ_j is computed giving an estimate of the variance of the details[2]. For LRD processes the μ_j follow a power-law in j with exponent α.
3. **Analysis using the Logscale Diagram:** Form the plot of $y_j = \log_2(\mu_j)$ against j, the *Logscale Diagram*[3], the scaling range (j_1, j_2) where scaling occurs is determined[4].
4. **LRD parameters estimation:** The LRD parameters H and c_f are estimated by weighted linear regression over the scaling region[5][6][7].

The joint AV estimator offers excellent statistical performance: negligible bias and close to optimal variance, and known confidence intervals, independent of the unknown H value. The asymptotic variance is $\mathrm{var}(\hat{H}(n)) \sim \frac{2^{j_1-3}}{\ln^2 2} n^{-1}$. The $1/n$ decrease is a non-trivial property for an estimator of LRD as it is normally a characteristic of estimates in a short range dependent context. Also note that the AV estimator is semi-parametric and therefore not dependent on a specific model for the data. It is also intrinsically robust with respect to non-stationarities in the mean and variance of the underlying process [3,11].

The AV estimator is gaining acceptance as the method of choice for measuring LRD in traffic [4,6]. Until recently however it has been used as an off-line, or batch estimator. It is ideally suited to on-line use however [12], making it suitable for real-time estimation. In the following section we will see that this method allows a traffic stream to be monitored continuously for months at a time, without a large memory requirement.

3. Practical implementation of real-time estimation

This section describes the analysis of Ethernet data on the local area network at the Software Engineering Research Centre (SERC) at RMIT University with simple low-cost hardware. Ethernet was initially tested (though we have now extended our monitoring to 155 Mbps ATM systems) for two reasons apart from the obvious convenience. First it was the first type of data network where self-similar traffic was shown to exist [7]. Second, it is relatively easy to extract traffic from an Ethernet because of the broadcast nature of the medium. The method is described in detail in [12], though the parameters used in this experiment are described below.

The SERC LAN was based mainly around a standard passive hub (the hub has no switching capability) 10baseT Ethernet with a file server, compute server, ~3 X-terms, ~6 Windows PC's, ~6 FreeBSD Unix boxes, 2 printers, and 1 MacIntosh attached. The numbers of user boxes and their type changed over the period of measurement as a number of boxes were dual-boot Windows/Unix boxes, or laptops (which were not always attached

[2]Since the expectations of the details are all identically zero [8,5]
[3]In computing y_j small corrective terms $g(n_j)$ are in fact subtracted from $\log_2(\mu_j)$.
[4]If the data is truly LRD then the upper cutoff scale $j_2 = j_{\max} \approx \log_2(n)$, however scaling in a finite range is also observed in data [1].
[5]H is related to the slope of the plot, and c_f to a power of the intercept.
[6]The weights are functions of the known variances of the y_j and do **not** depend on the data.
[7]Confidence intervals for H are derived from the standard variance formulae for weighted linear regression with mutually independent y_j, and so again are **not** functions of the data.

to the network), and several new boxes were attached during the time period. In addition there was a second 10Base2 leg of the network connected by a Router (which also provided the gateway to the Internet) to which a small number of X-terms were attached.

The output from a packet filter, which read all of the transmitted packets, was passed through a pre-filtering program which generates a time series corresponding to the number of bytes transferred during each sampling interval. This series is the raw data $x(t)$ analyzed by the on-line estimator. A sampling interval of 1ms was used in the experiments described here. All of the possible scales were used in the estimation, i.e. j_1 was set to 1, though we considered the log-scale diagrams themselves to inform our conclusions.

The use of commodity PCs and NICs allowed us to build very cheap monitoring systems, i.e. $< \$5000$ AUS. Such a low capital outlay is a requirement if such monitors are to become common enough to be useful.

4. Long term measurements

As already stated, a major advantages of on-line measurement is that measurements can be collected over long periods. For instance, data from the SERC LAN has been collected from March the 4th to August the 26th of 1998. A major reconfiguration of our local network occuring at the end of August was the trigger for our current study.

The data was collected by running a set of monitors almost continuously over the above period. However, the monitors were also used for other purposes, and therefore could not be used continuously. Therefore the data has some gaps over the time period.

We performed monitoring using blocks of data 1, 4 and 24 hours long, with the intention of studying the diurnal and weekly variations of both the load and H values of the traffic. An important aim was to determine over what time scales the Hurst parameter might vary, to determine an appropriate practical measurement interval. Another aim was to investigate structural features of the variation. For example if the variation was dominated by the diurnal behaviour of the system, it may be sufficient to measure this diurnal cycle thoroughly once and then to use it to predict H estimates at given times of day, or to choose the worst case as in the traditional "busy hour" used by Telcommunications providers. Naturally such an idea relies on a concept of having a time interval over which it is reasonable to consider that H does not vary, so that it be well defined and therefore measurable. Measurements over such an interval, say an hour, can then be used to track changes in H over longer intervals, such as days or weeks. We rely on the robustness properties of the AV estimator for reliable H estimates where aspects of the data vary.

Figure 1 shows example sample paths for the Hurst parameter over one week in April 1998. The graph shows that, typically, there would seem to be considerable variation in the Hurst parameter over these time periods. (Space limitation prevent the presentation of the sample paths over the entire measurement interval. For a larger time-set please see [10].) The broad correspondance between 1 and 4 hour estimates (within the limitations of the measurements) seems to indicate that H can be taken to be constant at times scales up to 4 hours. The disagreement between the 4 and 24 hour measurements (at least in some cases) is evidence that H cannot be taken as constant over 24 hour intervals. The hypothesis test reported in [13] would enable such questions to be answered in a more objective fashion, however its use is beyond the scope of this study.

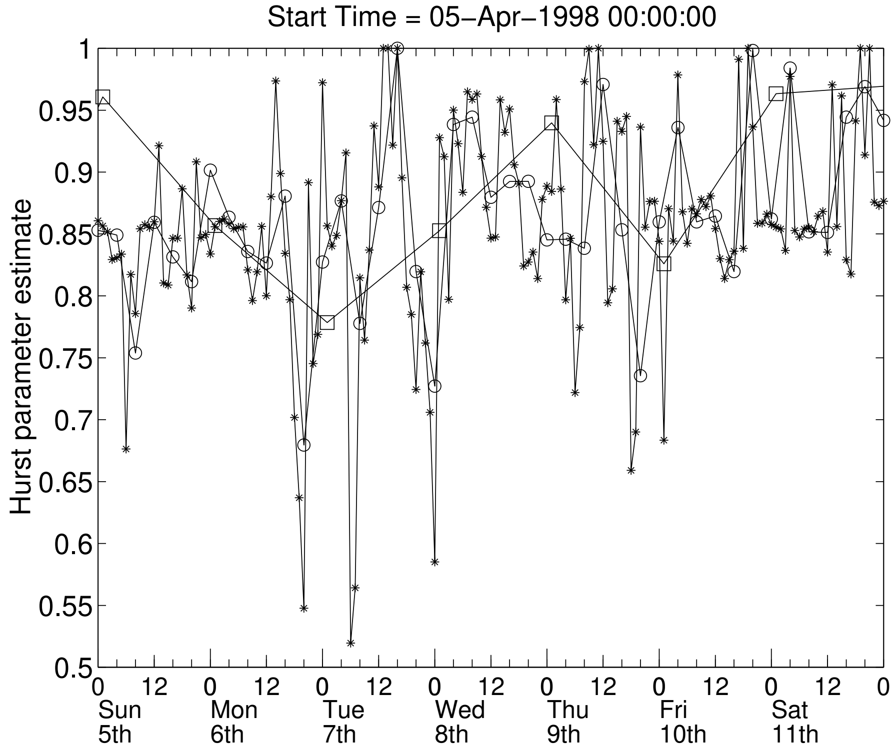


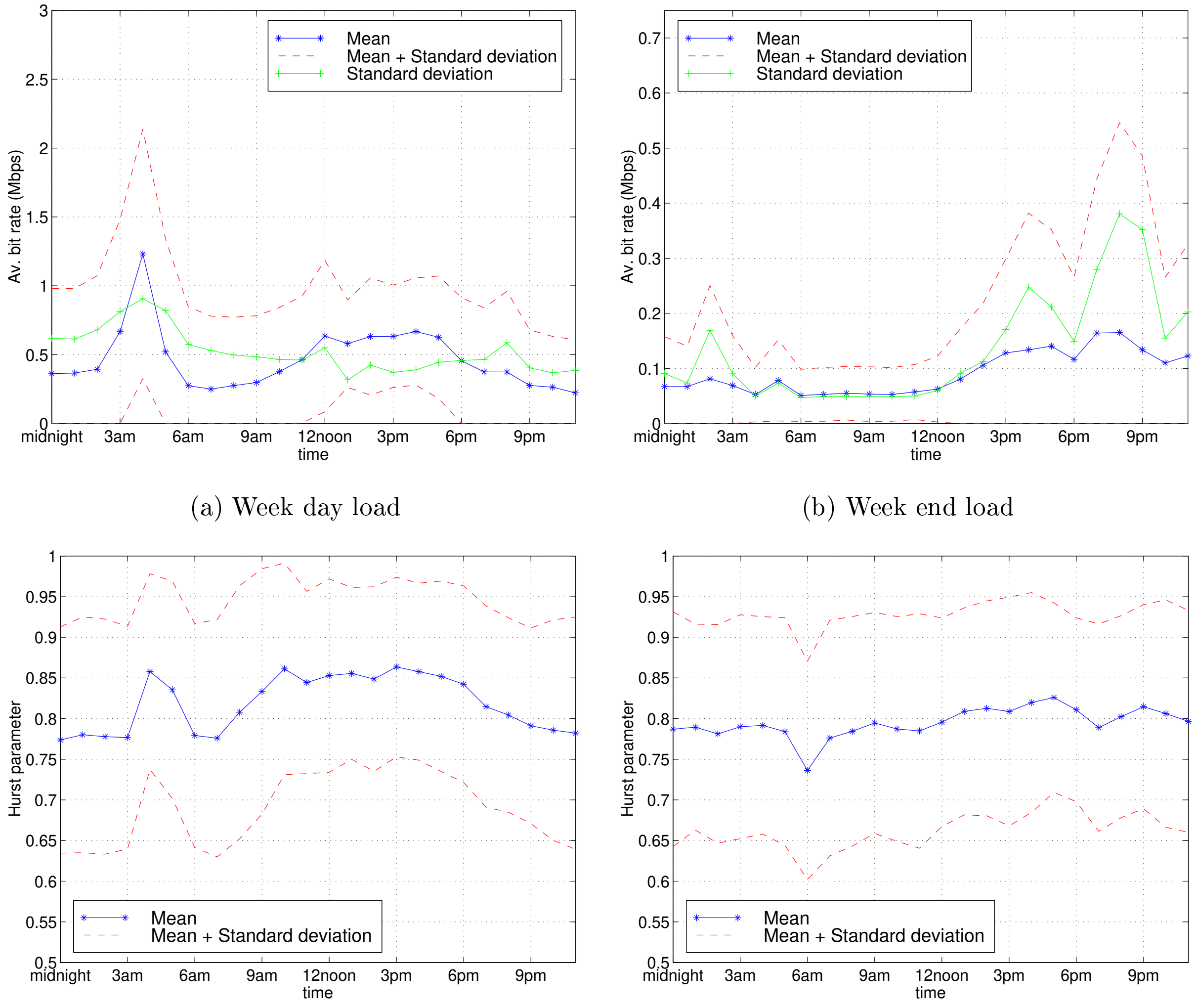
Figure 1. Example sample paths for the Hurst parameter. The three curves are based on 1 hour blocks of data ($*$), 4 hour blocks of data ($\circ$) and 24 blocks of data ($\square$).

The next set of figures shows the diurnal, or daily cycle of load at SERC. In *Figure 2*(a) and (b) we display the average from Wednesday, the 4th of March to Wednesday the 26th of August, 1998, of the load on the Ethernet during each one hour period. Figure (a) shows the weekday load, while Figure (b) shows the weekend load[8] – the two are substantially different, and it does not seem appropriate to combine them. The figure also shows the standard deviation of the load, and the mean $\pm$ the standard deviation.

The two most notable features of the mean load during the week are firstly a weak busy cycle we refer to as the *user busy cycle*. We refer to the user busy cycle as weak, because its magnitude is not large compared with the natural variation during the day – this view is supported by quantile plots (not shown here) of the data. Only by averaging measurements obtained over a large number of days, does this gentle diurnal variation emerge from the highly variable background. The peak of this cycle appears to occur at 4pm. The second notable feature is a large peak early in the morning. This is the result of the nightly backups starting at $\sim$ 3 am following each weekday.

Also notable in *Figure 2*(a) is the fact that the standard deviation of the results (over the days in the average) does not appear to correlate well with the user busy cycle, but that it does seem to be correlated with the backup peak.

[8]Backups on our system begin at approximately 3am and may extend past 6am after weekdays and therefore occur on Saturday morning but not Sunday or Monday morning. We consider the backups to be part of the week day workload, and hence we have adjusted the measured beginning and end of the weekend to 7am on Saturday morning, 7am on Monday morning respectively.



(a) Week day load

(b) Week end load

(c) Week day Hurst parameter

(d) Week end Hurst parameter

Figure 2. Diurnal cycle in load and Hurst parameter.

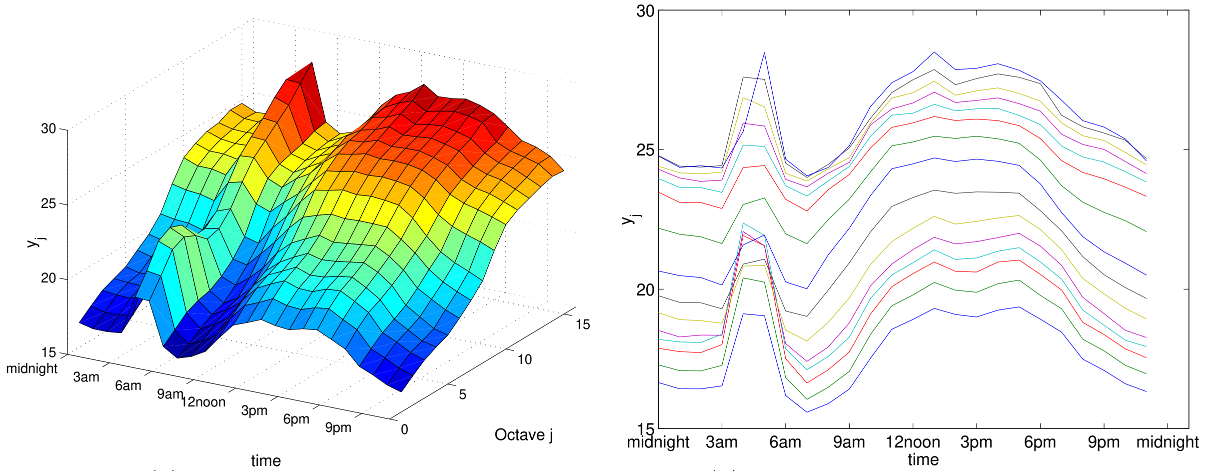
The traffic on the weekends is significantly lower than during the week – not surprisingly. However there is also a user busy cycle during the weekends which has a later peak, around 7 or 8 pm, which is not inconsistent with the observed behaviour of workers at SERC – they tend to work later in the day on the weekends.

Figure 2 (c) and (d) show the equivalent picture for Hurst parameter estimates. The standard deviation itself has not been plotted in the results, but it remains roughly constant with a value slightly larger than 0.1.

Figure 2(d) seems to indicate that the weekend traffic has a variable Hurst estimates (witness the size of the confidence band), but that the Hurst parameter does not depend strongly on the time of day. On the other hand *Figure 2*(c) indicates that during the week the Hurst parameter does depend on time of day, though not strongly. In particular there is a peak at about 4am which appears to be correlated with the backup peak, and a cycle corresponding to the user busy cycle, *i.e.*the Hurst parameter seems to have some correlation to the network load. Interestingly the peak due to the backups fits the backup load peak, while the busy cycle behaviour of the load seems to lag the Hurst parameter. This behaviour is thought to relate to the type of traffic present – for instance, SNMP vs X11 traffic. It would be interesting in future work to partition the data into types before study, though this would require separate monitors for each applications class considered.

Note that in all cases natural variation is greater than that observed across the diurnal cycle – and sample paths confirm that the variation occurs between continuous measurements in a single day, rather than being quasi-constant within a day and highly variable across days. Thus, estimates change quickly enough to justify fine scale ongoing monitoring rather than, for example, weekly measurements.

There are many potential causes for the observations above apart from the ideal case where each measurement is giving a true indication of the nature of the system. For instance, the AV estimator, while remarkably robust, is not completely immune to the effects non-stationarity, and is dependent of a good choice of j_1. Thus in order to assess the validity of the results we examine the Logscale Diagrams as well as the Hurst parameter estimates. The Logscale Diagram is in fact first of all an analysis tool to examine the average second order behaviour (energy) in the data as a function of scale, independently of the desire to measure scaling behaviour. Recall that long range dependence is detected in the Logscale Diagram if a region of alignment is detected, with lower cuttoff scale j_1.



(a) Week day values of y_j.

(b) Week day values of y_j by time.

Figure 3. Diurnal cycle in log-scale diagrams.

In *Figure 3* we show Logscale Diagram averaged by time of day for week days. Figure (a) shows the values of the log-scale averages y_j by scale j and time of day. Figure (b) shows the same set of data by time of day with one curve per scale. The plots show that the averaged y_j vary over the course of a day. Note that one kind of non-stationarity which might effect the H estimates would show up here in a certain characteristic form of the Logscale Diagram (for details see [11]) which is not apparent here. It seems then that there is real variation in the Logscale Diagram over the course of a day.

However, the changes in the Logscale Diagram cannot be captured solely by the estimate of H. Consider *Figure 4* which shows the log-scale coefficients y_j by scale for three sets of times respectively dominated by – low load (8am-10am and 10pm-3am), high load (11am-6pm) and backup loads (4am-5am). Each set of data has a different charateristic shape. During the high load periods the curves are approximately straight lines meaning the Hurst parameter estimates are useful in describing the range of scales in the data. However, at night, and in the early morning when the load is low, the log-scale curves are not straight, and hence the Hurst parameter estimates are only relevant for the asymptotic behaviour of the data, or not relevant at all.

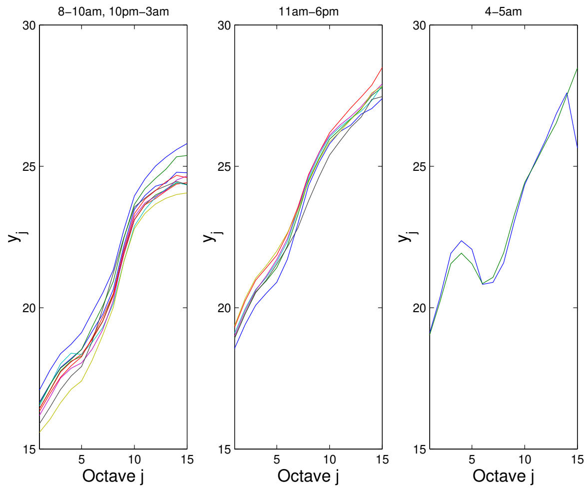


Figure 4: Diurnal cycle in log-scale diagrams — week day values of y_j by scale j.

The behaviour of the Logscale Diagram during the backup load is not characteristic of scaling behaviour at all, but rather there is a strong phenomena occuring on scales 3 to 5 (time scales from $2^3 = 8$ to $2^5 = 32$ ms). This is interesting, but not at all unreasonable. We know that during these times the load is due to machine driven processes (the backup software AMANDA) which may well generate quite regular traffic. Note that the system may still be asymptotically self-similar, however on the times scales observed the dominant behaviour is the fixed scale behaviour.

In the same vein note that during the busy hours of the day human interaction dominates the workload, whereas at times of low load (night, the weekend), computer interactions dominate. Thus the workload type seems to have a strong effect on the nature of scaling. Thus it is unsurprising that the weekend Logscale Diagrams (not shown here) are all qualitatively similar to the low load diagrams from the weekdays. Hence it would seem that human interaction plays an important role in self-similarity in traffic.

Therefore, although the Hurst parameter may not be capturing enough of the traffic behaviour at certain times, the scaling behaviour certainly does exist and has some diurnal variation, which is related to the level of human interaction with the system.

5. Conclusion

This study has applied the data reduction of the on-line AV estimator to nearly six months of LAN traffic. It it not the intention of this paper to bring out all of the information in this data, but rather it is intended to provide a taste of the possibilities created by the cheap, ubiquitous monitoring allowed by an on-line estimator, and to give some interesting results on interpretion of time variation in H.

The limitations of the data also restrict the utility of the data and it is clear that further study, ideally broken down by traffic type, will be required before drawing firm conclusions about the daily variation of the parameters of LRD, however, the data is strongly suggestive of a number of features described in detail above.

A conclusion of practical importance is that for real time network needs such as call admission control or rate adaptive applications one should make continuous measurements of the Hurst parameter (and other parameters of LRD). This is because, although diurnal variations exist, local changes are also very significant and may dominate. Our results suggests that in the case of Ethernet traffic that such measurements should be based around time scales from 1 to 4 hours: any smaller and the data sets are not sufficient to obtain an accurate estimate, any larger and the parameters of interest may change substantially over the measurement interval. Ideally a measurement scheme will be developed which is adaptive, and can choose the best possible measurement interval from the data.

REFERENCES

1. P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch. Wavelets for the analysis, estimation and synthesis of scaling data. submitted to Self Similar Network Traffic Analysis and Performance Evaluation, K. Park and W. Willinger, Eds., 1999.
2. P. Abry, P. Gonçalvès, and P. Flandrin. *Wavelets and Statistics*, volume 105 of *Lecture Notes in Statistics*, chapter Wavelets, Spectrum estimation, $1/f$ processes., pages 15–30. Springer-Verlag, New York, 1995.
3. P. Abry and D. Veitch. Wavelet analysis of long-range dependent traffic. *IEEE Trans. on Info. Theory*, 44(1):2–15, 1998.
4. A.Feldmann, A.C.Gilbert, W.Willinger, and T.G.Kurtz. Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic. Preprint, 1997.
5. I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia (PA), 1992.
6. Judith L. Jerkins and Jonathan L. Wang. A measurement analysis of ATM cell-level aggregate traffic. In *IEEE GLOBECOM'97*, 1997.
7. Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, Feb 1994.
8. S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
9. V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modelling. In *Proceedings of SIGCOMM '94*, 1994.
10. Matthew Roughan and Darryl Veitch. A study of the daily variation in the self-similarity of real data traffic. Technical Report 0070, SERC, Software Engineering Research Centre, Level 3, 110 Victoria St, Carlton Vic, 3053, AUSTRALIA, 1998.
11. Matthew Roughan and Darryl Veitch. Measuring long-range dependence under changing traffic conditions. In *IEEE INFOCOM'99*, NY, NY, March 1999. IEEE Computer Society Press, Los Alamitos, California.
12. Matthew Roughan, Darryl Veitch, and Patrice Abry. On-line estimation of parameters of long-range dependence. In *IEEE GLOBECOM'98*, pages 3716–3721, Sydney, Australia, November 1998.
13. Darryl Veitch and Patrice Abry. Testing the stationarity of the Hurst parameter using wavelets. submitted to Trans. Info. Theory.
14. Darryl Veitch and Patrice Abry. A wavelet based joint estimator of the parameters of long-range dependence. *to appear in IEEE Transactions on Information Theory special issue on "Multiscale Statistical Signal Analysis and its Applications"*, 1998.