Dhaka Univ. J. Eng. & Tech. Vol. 1(1) 61-64, 2010 (July)

Estimation of the Hurst Parameter for the Dhaka University Network Data Traffic

Shihan Sajeed[*], **Dewan Lutful Kabir**[**], **Nigar Sultana and Shahida Rafique**

Department of Applied Physics, Electronics & Communication Engineering, University of Dhaka, Dhaka, Bangladesh
shihan.sajeed@gmail.com[*], *dlkabir@gmail.com*[**]

Abstract

The main goal of this work was to study and analyze the network data traffic of the University of Dhaka and to estimate the Hurst parameter to assess the degree of self similarity in the data traffic. For this verification a number of tests and analyses were performed on the data collected from the University Gateway router. The conclusions were supported by a rigorous statistical analysis of 75 millions of data packets of high quality Ethernet traffic measurements collected between Aug '07 and March'08 and the data were analyzed using both visual and statistical experimentation. Busy hour traffic and non-busy hour traffic, both were considered. All the program codes were written using MATLAB and can be used as an excellent tool to determine the degree of self-similarity in a network's data traffic.

Keywords: self-similarity, Hurst parameter, variance-time plot.

I. Introduction

According to Schroeder [1]: 'The unifying concept underlying fractals, chaos and power laws is self- similarity (SS). Self-similarity or invariance, against changes in scale or size, is an attribute of many laws of nature and innumerable phenomena in the world around us. Self-similarity is, in fact, one of the decisive symmetries that shape our universe and our efforts to comprehend it.'

A phenomenon that is self-similar looks the same or behaves the same when viewed at different degrees of "magnification" or different scales on a dimension. The dimension can be space (length, width) or time. For stochastic objects like time series, SS is used in the distributional sense: when viewed at varying scales, the object's distribution remains unchanged [2]. A stochastic process x (t) is statistically self-similar with parameter H (0.5 $\le$ H $\le$ 1) if for any real a > 0 the process $a^{-H}x($ has the same statistical properties as x. This relationship may be expressed by the following three conditions:

- $E[x(t)] = {}^{E[x(at)]}/_{a^H}$ (mean)
- $Var[x(t)] = {}^{Var[x(at)]}/_{a^{2H}}$ (variance)
- $R_x(t,s) = {}^{R_x(at, as)]}/_{a^{2H}}$ (autocorelation)

The parameter *H is* known as the 'Hurst parameter' and is a key measure of self-similarity (SS). More precisely, *H* is a measure of the persistence of statistical phenomenon and is a measure of the length of long range dependence [3] of a stochastic process. In this paper, the self-similarity parameter or the H parameter of University of Dhaka network data traffic is measured. Besides this, a similar network was considered that generates random data and the characteristics of the randomly generated data was compared to the actual network to prove its failure of showing the Hurst effect.

The organization of this paper is as follows: section II summarizes the theoretical demonstration of various methods to estimate the Hurst parameter. Section III provides the experimentation methodology, the network model and the collected data. In section IV, the results for various tests are summarized. Finally, the concluding remarks are presented in section V.

II. Theoretical background

There are a number of ways to measure the H parameter of a self similar process. In this work two different techniques have been used.

A. Variance time plot:

From statistical point of view, the most salient feature of a self-similar process is that the variance of the aggregated sample decreases more slowly than the reciprocal of the sample size as shown in Figure 1. The figure, also called 'variance-time (V-T) plot', shows a theoretical comparative study between a self-similar and a normal process.

The so-called *variance-time* plots can be obtained by plotting log [var($X(m)$)] against log(m), where m denotes the level of aggregation, and by fitting a simple least squares line through the resulting points in the plane, ignoring the small values for m [3], [4].

The slope of the straight line is a characteristic of SS. For most processes such as Poisson, the slope of the variance time plot is approximately equal to -1 and diminishes quite rapidly as the sample size is increased but for SS process, the variance decreases very slowly even when the sample size grows quite large with a slope value far greater than -1.

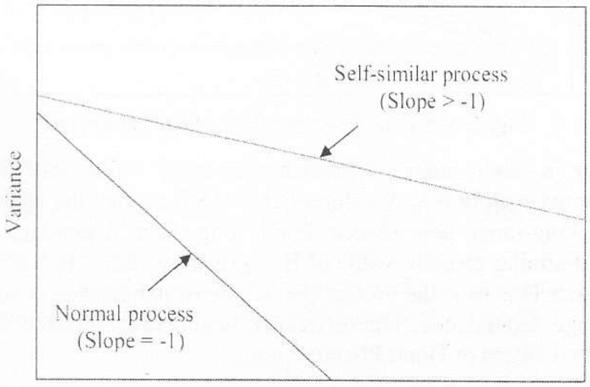

Fig. 1: Decay of Variance with respect to Sample size between a Normal and a Self-similar process.

The variance decay like $n^{-\beta}$ for some $\beta \in (0, 1)$, instead of like n^{-1} for the processes whose aggregated series converge to second-order pure noise. Slope values close to -1 indicates weak self similarity while the value close to 0 indicates a stronger self similarity [5]. Here β is a parameter which is related to Hurst parameter by the relation:

$$H = 1 - (\beta/2) \qquad (1)$$

B. *R/S statistics*

For a given set of observations (X_k : k = 1, 2, ..., n) with sample mean X(n) and sample variance $S^2(n)$, the rescaled adjusted range or R/S statistic is given by [5]:

$$\frac{R(n)}{S(n)} = \frac{1}{S(n)} \quad [\max(0, w_1 \ldots w_n) - \min(0, w_1 \ldots w_n)] \qquad (2)$$

Where,

$$w_k = (X_2 + X_2 + \ldots + X_k) - K X_n \text{ for } k = 1, 2\ldots n \qquad (3)$$

Hurst found that many naturally occurring time series can be well represented by the relation [6 , 7]:

$$E[R(n)/S(n)] \sim c\, n^H, \quad \text{as} \quad n \to \infty \qquad (4)$$

Here 'H' is called Hurst parameter & its value is about 0.73, and 'c' is a finite positive constant that does not depend on n. On the other hand, if the observations X_k come from a short-range dependent model, then Mandelbrot and Van Ness [8] showed that:

$$E[R(n)/S(n)] \sim d\, n^{0.5}, \quad \text{as} \quad n \to \infty \qquad (5)$$

Here, d is a finite positive constant, independent of n.

R/S plot is an excellent way of testing self-similarity & estimating the Hurst parameter in which the R/S statistic for different values of n (block size) is plotted with a log scale on each axis. If time series is self-similar, the resulting plot will have a straight line shape with a slope H that is greater than 0.5 and less than 1 as shown in Figure 2.

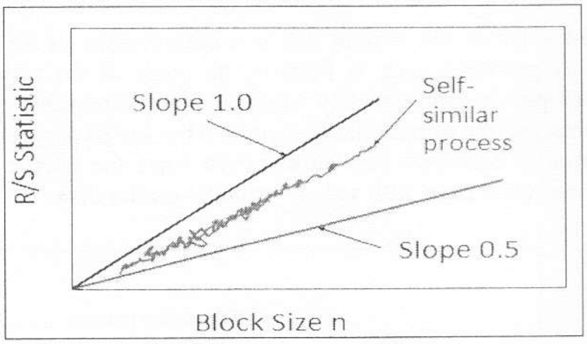

Fig. 2: R/S plot (The slope gives the H parameter)

For a short-range dependent process, value of H is approximately 0.5. A value of H = 0.5 indicates the absence of long-range dependence. For a long-range dependent i.e., self-similar process, value of H is given by: 0.5 < H < 1. The closer H is to 1, the greater the degree of persistence or long-range dependence. The difference is generally referred to as Hurst Effect or Hurst Phenomenon.

III. TESTS & ANALYSIS

A. *The traffic monitoring tool*

The monitoring system used to collect the data for the present study is software named 'WIRESHARK', which copies all packets seen on the Ethernet under study with accurate timestamps, and will do so for very long runs without interruption. The monitor was downloaded from the website www.wireshark.com

B. *Network environment in the University of Dhaka*

There are 10 faculties, 51 departments, 9 institutes, 18 residential halls and hostels and 18 different research centers present in DU. All of these are equipped with the internet facility. So, it is a large enough network to handle massive amount of data packets for such kind of experimentation. The backbone of the University of Dhaka network, obtained with the special permission of the director of the IIT of University of Dhaka is presented in Figure 3. The CISCO 7204VXR router is used as the backbone router. The traffic is mainly divided into three parts between new science building, Curzon Hall and the register building. With further investigations additional information were obtained about the campus network as shown in the following Figure 3.

C. *Collection of data:*

Data were collected for two different time durations at two different times of the year as summarized in table 1. The '35 minute duration' data was considered 'non-busy hour traffic' since the University was closed during time of collection and only a few departmental offices were open, while other educational activities were suspended. The '3 hour duration' data was considered 'busy-hour traffic' since the campus activities were on full swing during that time and the data was captured during the "busiest time" of the day. As a result the two captured data series acted as an excellent candidate to represent two extreme conditions for the university network traffic.

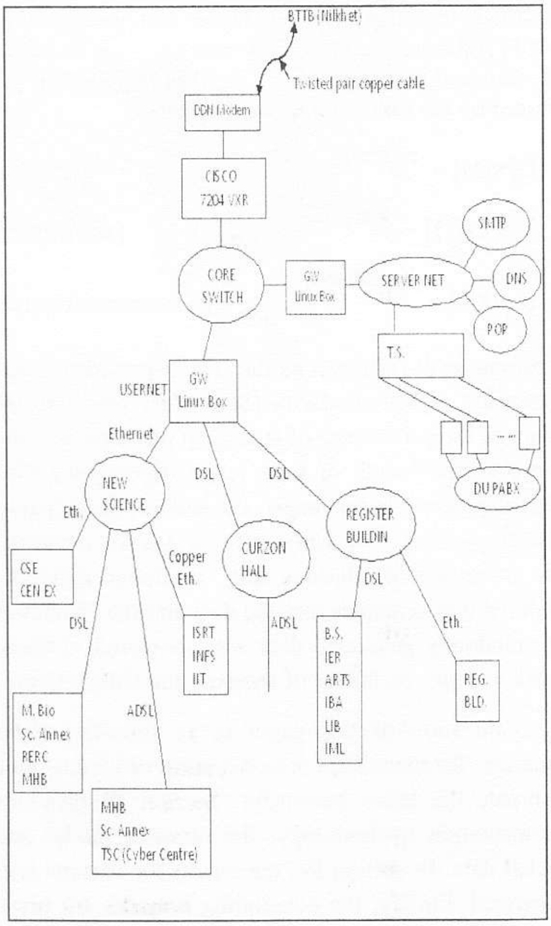

Fig. 3: The backbone of DU Campus Network.

Table 1: Traffic Measurement Records of University of Dhaka Network (Captured from Institute of Information Technology)

Measurement period	Total no. of packets
22nd August 2007 Start of trace: 11:17 am End of trace: 11:52 am Total 35 minutes	1,074,588
8th March 2008 Start of trace: 10:30 am End of trace: 01:30 pm Total 3 hours	74,558,988

Table 1 summarizes the captured data from which it is seen that the '35 minute duration' data contained almost 1 million data packets while the data which was collected for a duration of 3 hours contained a massive 75 millions data packets and hence could be used for any rigorous analysis. MATLAB program codes were written for the analysis and estimation of the collected data and various experiments were performed (described in section II).

D. *Test number 1: Variance-Time plot test*

For the non busy hour traffic, the variance of the aggregated series is plotted against the aggregated sample from the collected data. A random data of the same size was also generated to compare the variance time plot of the random data with that of the experimental data traffic. The random data traffic is obtained by using a random number generating program. Both the variance time plots are shown in Figure 4.

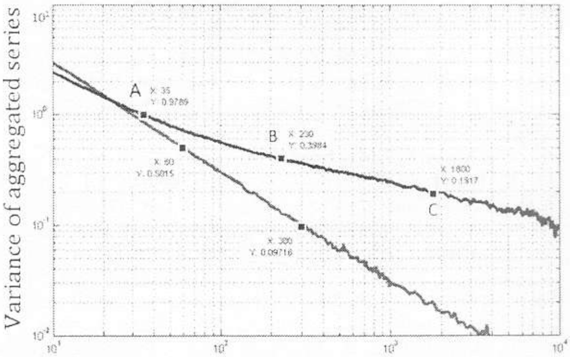

Fig. 4: V-T plots for non busy hour traffic (35 minute duration data) along with the V-T plot for random data of the same size.

In Figure 4, the upper line (blue) represents the variance-time plot for the experimental data obtained from the DU network while the lower line (red) represents variance-time depiction for the random data generated using random number generator code. It is clearly seen from Figure 4 that the V-T plot for real data decays (i.e., reaches towards zero) more slowly than the random data.

The slope of the lower (red) line is calculated and the value is approximately - 1.0197 which is a clear indication of random data. The result agrees with the theory as the data was indeed generated using random number generator code. On the contrary, for the upper line in Figure 4 the slope was calculated for three different portions.

Between points A and B the slope is approx = - 0.477

Between points B and C the slope is approx = - 0.355

Between points A and C the slope is approx = - 0.412

In this case the slope had different value for different region and though it was seen that the real data had a slowly decaying variance but the parameter β could not be readily estimated due to the nonlinearity in the curve. However, this variance time plots were generated for the non busy hour traffic containing almost 1,074,588 numbers of packets.

In order to obtain a more accurate value of β, the analysis was repeated for the 3 hour data containing 74,558,988 packets and the resulted V-T plot is shown in Figure 5. In this case, the slope of the lower line (random data) is approximately - 0.99 as is expected from random data. For the experimental data (red line in this case):

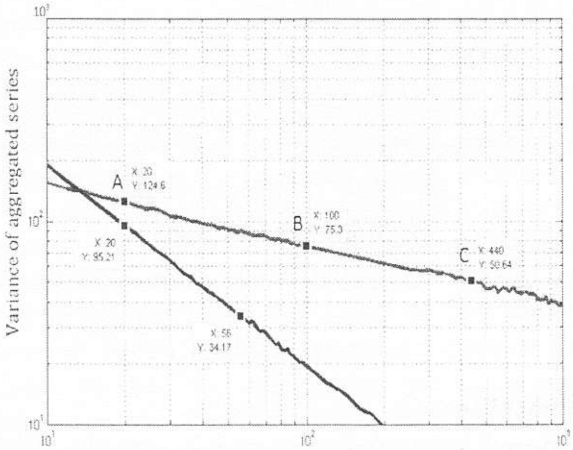

Fig. 5: V-T plots for busy hour traffic (3 hour duration data) along with the V-T plot for random data of the same size.

In this case, for the experimental data (upper line in figure 5):

Between points A and B the slope is approx = - 0.31

Between points B and C the slope is approx = - 0.27

Between points A and C the slope is approx = - 0.29

Hence, the actual value of β is approximately -0.29 and is more accurate because the line is more linear and almost any region in the line had a slope of value close to – 0.29.

Estimation of the H using V-T plot:

'H' is estimated by putting the value of β in equation (1):

H = (1 – 0.412 / 2) = 0.794 (for 30 minutes data)

H = (1 – 0.290 / 2) = 0.855 (for the 3 hour data)

E. *Test number 2: R/S statistics*

Matlab codes where written for performing the R/S analysis on the experimental data and the result is shown in Figure 6.

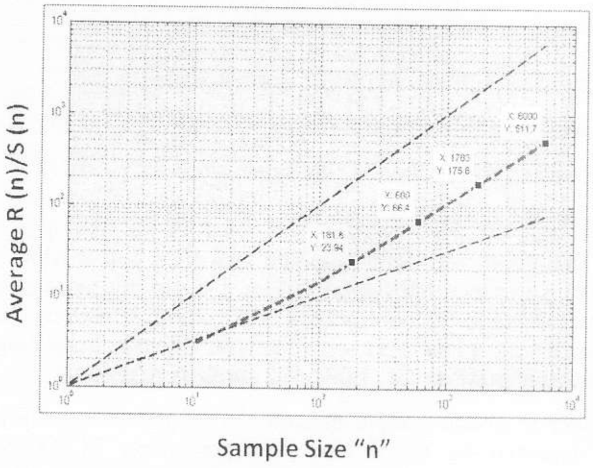

Fig. 6: Measurement of the slope from the R/S plot.

The value of the slope for larger samples, i.e. the value of the Hurst parameter was found to be,

$$H = 0.87$$

IV. RESULT

From the V-T plot test, for both the busy hour traffic and non busy hour traffic the data was seen to have a slowly decaying variance and the estimate of the HURST parameter was,

H = 0.794 (35 minutes traffic)

H = 0.855 (3 hour traffic)

From the R/S analysis, the value obtained for the Hurst parameter was,

$$H = 0.87$$

V. CONCLUSION

In this work, two methods were used for estimating the value of Hurst parameter and the value was found. The estimated values of the Hurst parameter from these tests were found to be perfectly in the predicted range. The network data traffic of University of Dhaka was characterized which may be further used to make improvements over this network. Following concluding remarks are made:

As seen from the two tests, the estimated value of the Hurst parameter is not same. For the V-T plot test the estimated value of H was in the range of 0.794 – 0.855 but for R/S statistics test the estimated value of H was 0.87. Thus, relying only on the results of any single test for the Hurst parameter is likely to draw a false conclusion and should never be used for planning and modeling, no matter how sound the theoretical backing is for the estimator in question. It is proposed to perform at least two tests for the estimation of Hurst parameter and to choose that value of H for design, modeling and planning, which is capable of dealing the worst case scenario.

- While simple filtering techniques are suggested in the literature for improving the performance of Hurst parameter estimation, they had little or no effect on the data analyzed in this paper.
- The value of the Hurst parameter is typically a function of the overall utilization of the Ethernet and can be used for measuring the "burstiness" of the traffic (namely, the burstier the traffic the higher *H*).
- H varies with the length of the data as well as with the time at which they are collected.
- Major components of Ethernet LAN traffic such as external LAN traffic or external TCP traffic share the same self-similar characteristics as the overall LAN traffic.

REFERENCES

[1] Schroeder, M, "Fractals, Chaos, Power Laws: Minutes from an infinite Paradise", New York: Freeman, 1991.

[2] Stallings, W., "High Speed Networks and Internets: Performance and Quality of Service", 2nd edition, Pearson Education Inc., ISBN 81-7758-569-X, pp. 219-244, 2006.

[3] Clegg, Richard G., "A Practical Guide To Measuring The Hurst Parameter", International Journal of Simulation: Systems, Science & Technology 7(2), pp. 3-14, 2006.

[4] Crovella, A., and Bestavros, A., "Self-similarity in World Wide Web Traffic: Evidence and possible causes", proceedings, ACM Sigmetrics conference on measurement and modeling of computer systems, May 1996.

[5] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D., "On the self-similar nature of Ethernet traffic (extended version)", IEEE/ACM Transaction on Networking 2, Feb. 1994.

[6] H. E. Hurst, "Long-Term Storage Capacity of Reservoirs", Trans. Amer. Soc. Civil Engineers 116, pp. 770-799,1951.

[7] H. E. Hurst, "Methods of Using Long-Term Storage in Reservoirs", Proc. Institution Civil Engineers, Part I, pp. 519-577, 1955.

[8] B. B. Mandelbrot, J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications", SIAM Review 10, pp. 422-437, 1968.