Dhaka Univ. J. Eng. & Tech. Vol. 1(1) 55-59, 2010 (July)

An Experimental Demonstration of Self-similarity in 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 is to study and analyze the behavior of the network data traffic of the University of Dhaka and to characterize it by measuring some specific parameters of the self similar traffic model of University of Dhaka (DU). A number of tests and analyses were performed on the data collected from the University of Dhaka Internet Gateway router and both the busy hour traffic and non-busy hour traffic are brought under experimentation. Visual tests as well as statistical experimentations are performed on the collected data and the conclusions are supported by rigorous statistical analyses of about 75 millions of data packets of high quality Ethernet traffic collected between Aug'07 and March'08. Both the visual test and statistical experimentations are successfully able to estimate several parameters of the Internet traffic and characterizing it thereby. The Matlab codes written during this work are also proposed to be used as an excellent tool for analyzing any other network traffic.

Keywords: Self-similarity, variance-time plot, long range dependancy, fractal nature.

I. INTRODUCTION

Researches and studies have shown that present high speed network traffic has some form of fractal nature which none of the traditional traffic models can capture. Hence, a new model is suggested, named, 'self-similar traffic model' in order to describe this fractal nature.

According to Manfred Schroeder [1] "The unifying concept underlying fractals, chaos and power laws is self-similarity. 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. In the case of stochastic objects like time series, self-similarity is used in the distributional sense: when viewed at varying scales, the object's distribution remains unchanged [2].

The organization of this paper is as follows. Section II summarizes the theoretical aspects of self similarity which are brought under experimentation for University of Dhaka (DU) network traffic. Section III provides the experimentation methodology, the network model and collected data. The results as well as discussions for the various tests are also presented in the same section. Finally, the concluding remarks are presented in section IV.

II. THEORETICAL ASPECTS OF SELF-SIMILARITY

A. *Traffic burstiness*

The most obvious property of a self-similar distribution is 'traffic burstiness' at large scales as well as at small scales. The traffic burstiness at different scales can be easily demonstrated through visual observations. In case of a self-similar traffic, the data remains bursty at smaller time scales (higher resolution) as well as larger time scales (lower resolution) whereas, for a non self-similar traffic (e.g., Poisson traffic), the data smoothen out (becomes non-bursty) at larger time scales (lower resolution). By observing any data traffic as a function of time, its self-similar nature can be verified which is the 'pictorial verification' of self similarity.

B. *Slowly decaying variance.*

From a 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 V-T plot 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 the self-similarity. For most processes the slope of the V-T plot is close to -1 and diminishes quite rapidly as the sample size is increased but for self-similar process, the variance decreases very slowly even when the sample size grows quite large with a slope value far greater than -1.

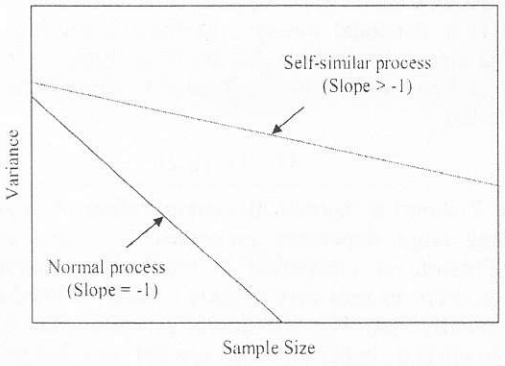

Fig. 1: Decay of Variance with Sample size between a Normal process and a Self-similar process.

As shown in the figure 1, for 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. The value of β is an important characteristic of any self-similar model.

C. *Long range dependency*

Long range dependency is defined in terms of the behavior of the auto covariance function C (τ) of a stationary process as τ increases. For many processes, C (τ) rapidly decays with τ. For example, for the Poisson increment process with mean λ and increment L, the auto covariance for values of $\tau > L$ is [2]:

$$C(\tau) = R(\lambda) - \lambda^2 = \lambda^2 - \lambda^2 = 0 \quad (1)$$

In general, a short range dependant process satisfies the condition that its auto covariance decays at least as fast as exponentially and hence it can be written as [2]:

$$C(\kappa) \sim \alpha^{|\kappa|} \text{ as } |\kappa| \to \infty \quad \text{where } 0 < \alpha < 1 \quad (2)$$

Here ~ denotes that the expressions on the two sides are asymptotically proportional to each other. The types of data traffic models typically considered in the literature employ only short-range dependent processes. Using the equality,

$$\sum_{K=0}^{K=\infty} x^k = 1/(1-x) \text{ while } |x| < 1 \quad (3)$$

It can be observed that $\Sigma_k C(k)$ for a short-range process is finite. In contrast, a long-range dependent process has a hyperbolically decaying auto covariance:

$$C(\kappa) \sim |\kappa|^{-\beta} \text{ as } |\kappa| \to \infty \quad \text{where } 0 < \beta < 1 \quad (4)$$

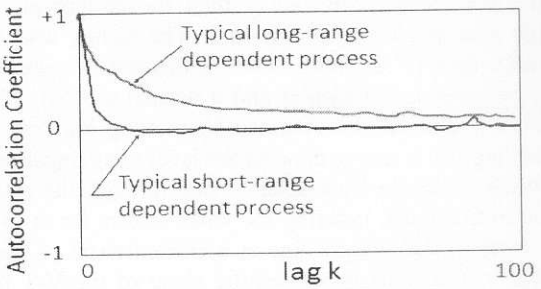

Fig. 2: Autocorrelation Function for typical short-range and long-range dependent processes.

In this case, $\Sigma_k C(\kappa) = \infty$. This long-range dependency can be explained with the help of the auto correlation function (AF) which is a statistical measure of the relationship, if any, between a random variable and itself, at different time lags. Here β is a parameter which is related to Hurst parameter by the relation:

$$H = 1 - (\beta/2) \quad (5)$$

Figure 2 shows a theoretical demonstration of short range and long range dependent processes. For most processes (e.g., Poisson, or compound Poisson), the autocorrelation function drops to zero very quickly (usually immediately, or exponentially fast). For self-similar processes, the AF drops very slowly (i.e., hyperbolically) toward zero, but may never reach zero.

III. Experimental Detail

A. *Traffic monitoring tool*

The monitoring system that is 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.

B. *Network model of University of Dhaka*

There are 10 faculties, 51 departments, 9 institutes, 18 residential halls and hostels and 18 different research centers in the University of Dhaka (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.

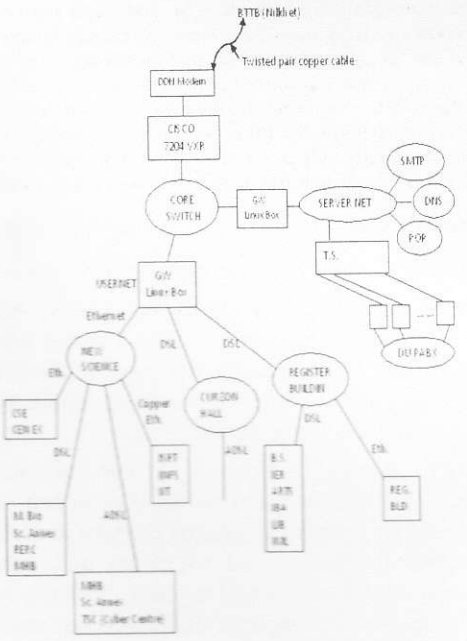

Fig. 3: Backbone network of University of Dhaka Campus.

C. *Collection of data:*

Data were collected for two different time durations at two different times of the year and are summarized in Table 1. The '35 minutes duration' data containing approximately 1 million data packets was considered 'non-busy hour traffic' since the University was closed during the time of collection and only a few of the departmental offices were open, while all other educational activities were suspended. On the contrary, the '3 hour duration' data containing a massive 75 millions packets was considered as 'busy-hour traffic' since the campus activities were on full swing during that time. Moreover, 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.

Table 1. Traffic measurement records of University of Dhaka (taken 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	10,74588
8th March 2008 Start of trace: 10:30 am End of trace: 01:30 pm Total 3 hours	745,58988

D. Pictorial verification

Codes were written using MATLAB for the pictorial verification of self-similarity and the resulting graphs for the busy hour traffic (3 hour duration) are shown in Figure 4.

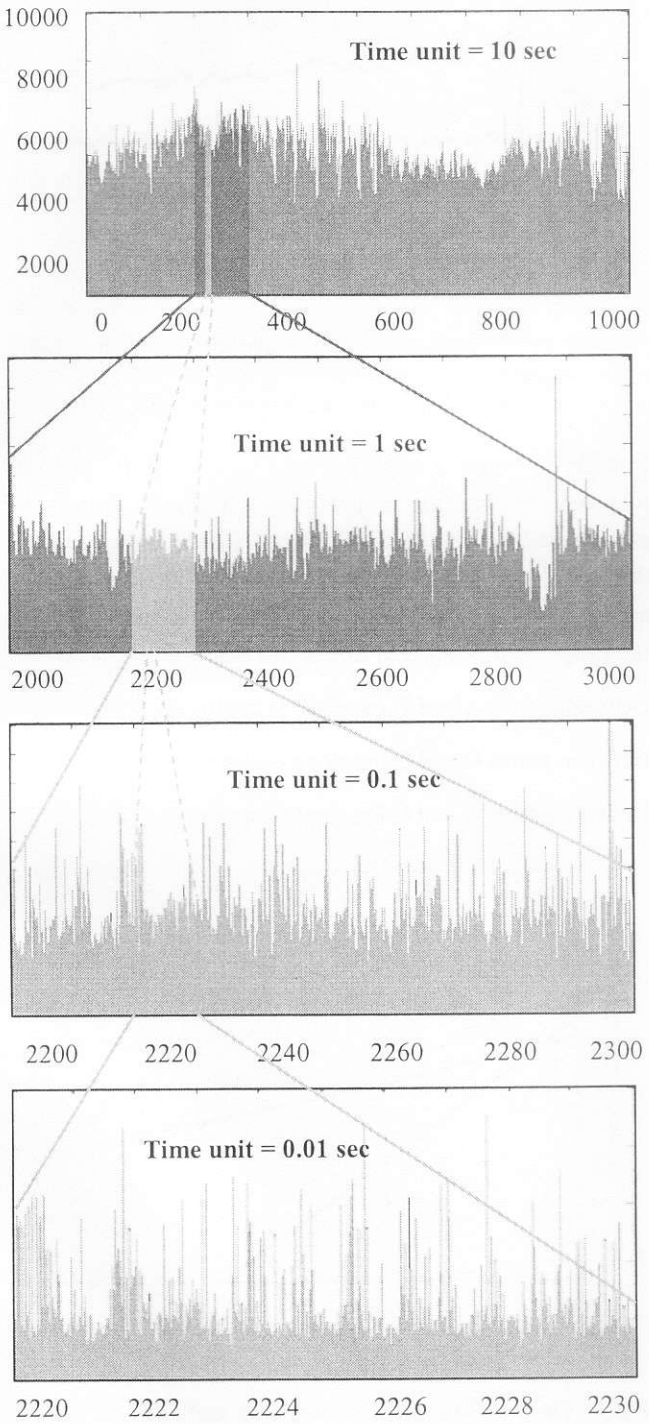

Fig. 4: Pictorial veriffication of self-similarity for the busy hour traffic. The top graph has a time unit of 10 seconds while the 2nd graph's time unit is 1 second. The time resolution is increased by a factor of 10 in the subsequent graphs.

Similar analyses were performed on the non-busy hour traffic (35 minutes duration data) and the resultant graphs are shown in figure 5.

Figure 4 and 5 illustrate the number of packets per time unit versus time unit for two different time durations which are obtained by running the MATLB codes with the help of the collected data of Table 1.

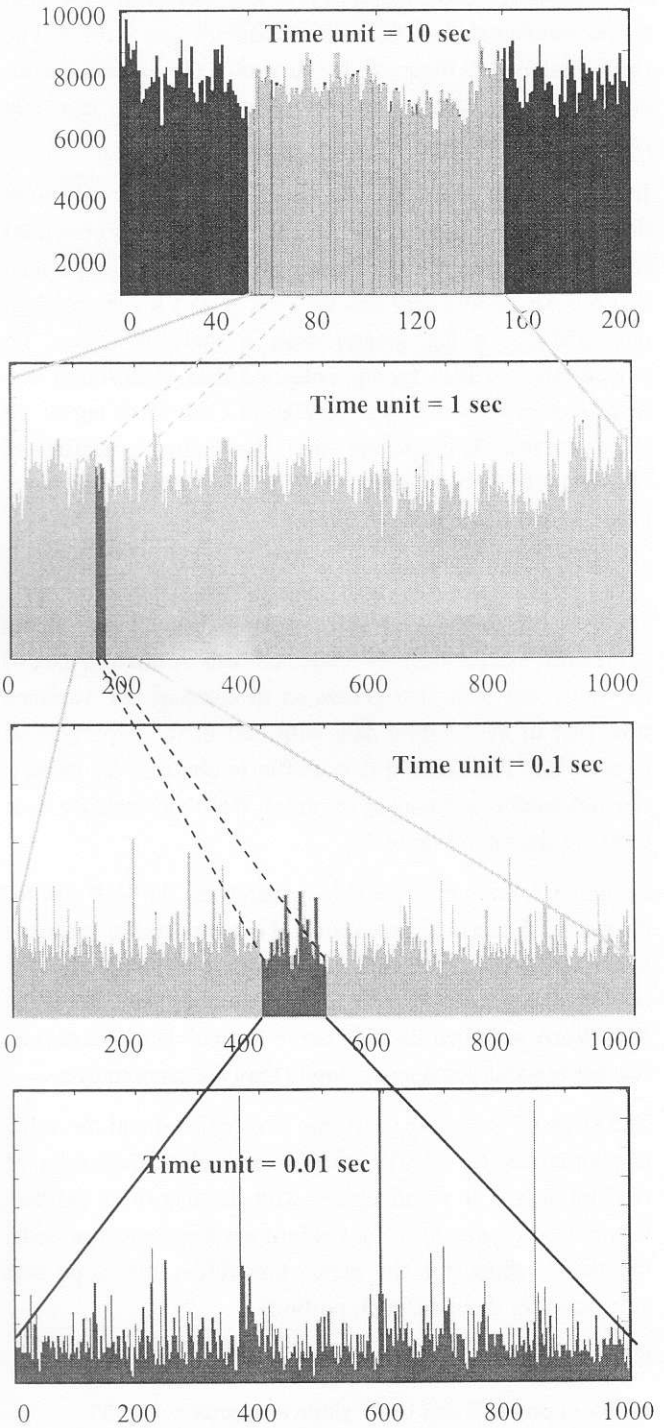

Fig. 5: Pictorial veriffication of self-similarity for the non-busy hour traffic. The top graph has a time unit of 10 seconds while the 2nd graph's time unit is 1 second. The time resolution is increased by a factor of 10 in the subsequent graphs.

In both figures, x-axis represents the "Time Unit" and y-axis stands for the "Packets/time unit". From the resulting graphs it is seen that the burstiness of traffic remains at all scales (from "time unit = 10 sec" to "time unit = 0.01 sec") for the captured traffics of both duration. Thus, the graphs are

successful verification and proof of the self-similarity in the University of Dhaka network traffic.

E. Long range dependency test

The long-range dependency test is also performed for both the experimental data traffic and random data traffic. The result is shown in Figure 6. In the y-axis, the values of auto-correlation function is placed whereas, time lags are represented in the x-axis.

In Figure 6, the blue curve represents the curve for the actual data and the red curve is the AF (Auto-correlation Function) for the randomly generated data. From Figure 6, it is seen that the AF decays very quickly (red curve) for the random data and for a lag greater than 5 there is almost no autocorrelation. But, for the collected data, it is found that there is some positive autocorrelation even for a lag of 50 (blue curve). This proves that the network traffic of University of Dhaka campus has the long range dependency property and hence is self-similar.

F. Variance-Time plot test

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 7.

In Figure 7, the upper line (blue) represents the V-T plot for the practical data obtained from real network traffic while the lower line (red) represents variance-time depiction for the random data generated using random number generator code. It is clearly seen that the V-T curve 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 (blue) 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.

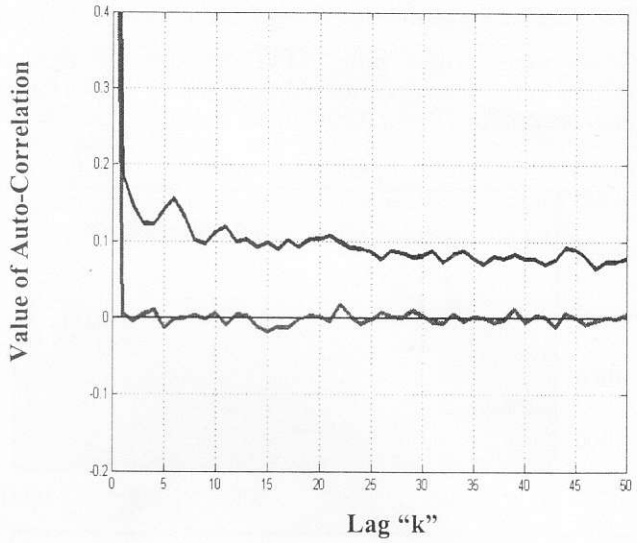

Fig. 6: Autocorrelation function vs. lag k.

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 8. 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):

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

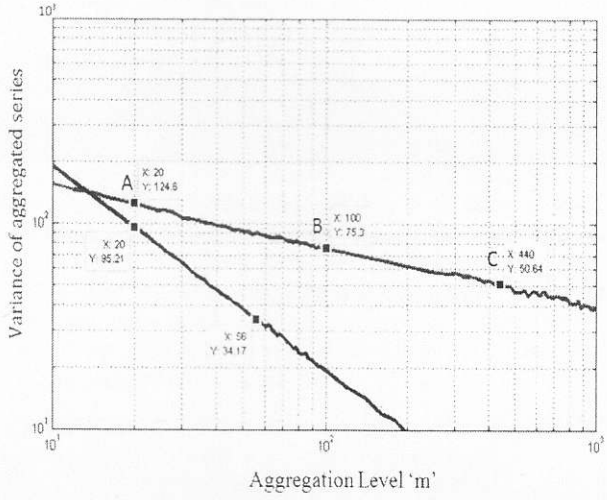

Fig. 7: Variance-Time plots for busy hour traffic (3-hour duration data) along with the Variance-Time plot for random data of the same size.

Hence, the actual value of β is much closer to -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.

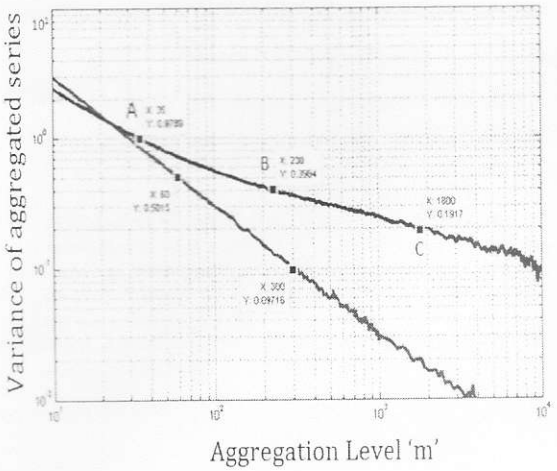

Fig 8: Variance-Time plots for busy hour traffic (3-hour duration data) along with the Variance-Time plot for random data of the same size.

The value of β can be used to estimate the 'Hurst parameter' which is another important attribute of self-similar model representing the degree of self similarity. The Hurst parameter is estimated by putting the value of β in equation (5):

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

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

IV. Conclusion

This work was intended for determining the nature of the self-similarity in the network traffic of the University of Dhaka. From the variance time plot the value of β was estimated -0.29 and the Hurst parameter was estimated to be approximately '0.8'.

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., 2006, ISBN 81-7758-569-X, pp. 219-244.

[3] 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.

[4] 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.