Simulation of self-similarity in network utilization patterns as a precursor to automated testing of intrusion detection systems

David A. Nash†
Department of Electrical Engineering and Computer Science

Daniel J. Ragsdale
Information Technology Operations Center

United States Military Academy, West Point, NY, USA

Abstract—
The behavior of a certain class of automatic intrusion detection systems (IDS) may be characterized as sensing patterns of network activity which are indicative of hostile intent. An obvious technique to test such a system is to engage the IDS of interest, and then use human actors to introduce the activities of a would-be intruder. While having the advantage of realism, such an approach is difficult to scale to large numbers of intrusive behaviors. Instead it would be preferable to generate traffic which includes these manifestations of intrusive activity automatically.

While such traffic would be difficult to produce in a totally general way, there are some aspects of network utilization which may be reproducible without excessive investment of resources. In particular, real network loading often exhibits patterns of self-similarity, which may be seen at various levels of time scaling. These patterns should be replicated in simulated network traffic as closely as is feasible, given the computational ability of the simulator. This motivates interest in an efficient way to detect multi-scale phenomena in network traffic, as well as a means to create simulated traffic that exhibits the desired characteristics.

We propose the use of multiresolution wavelet analysis as a technique which may be used to accomplish the desired detection, and subsequent construction of self-similarity in the simulated traffic. Following a multiresolution decomposition of the traffic using an orthogonal filter bank, the resulting wavelet coefficients may be filtered according to their magnitude. Some of the coefficients may be discarded, yielding an efficient representation. We investigate the effect of compression upon the reconstructed signal's self-similarity, as measured by its estimated Hurst parameter.

I. Introduction

On a local area network (LAN), computers communicate with one another by exchanging messages in a particular format. Every such message is referred to as a *packet*. The semantics of a particular exchange of packets are determined by various message exchange protocols. The structure of these protocols and the physical attributes of the communications medium over which the packets are transmitted give rise to a set of characteristics which are common to all network traffic which uses the same apparatus, essentially independent of the traffic content.

We wish to address the problem of generating traffic for the purpose of simulating the performance of a local area network. Generally, this equates to the problem of describing a distribution of message arrivals which matches the distributions observed in actual computer networks. For purposes of analytic tractability, an often-used approach is to assume that the arrivals of individual messages are independent of one another, and consequently the use of an exponential distribution to model this phenomenon is used. It has become clear however, that this assumption may not be warranted in the case of computer network traffic [1]. In fact, it turns out that the very characteristic which the exponential distribution precludes (non-zero autocorrelation of a time series of observations) is one which is desirable to preserve. This is because one often observes a tendency of traffic patterns to be "bursty," as opposed to varying evenly about the mean.

There are two factors which may account for this phenomenon:
1. The initiation of a packet sequence is often directly related to human-generated input. These inputs are rarely continuous, as the modern paradigm of computing is to employ computers in an interactive mode, as opposed to a batch-processed mode. Consequently, patterns of activity and quiescence are observed, which correspond roughly to periods of requests for computer processing of some sort, followed by an interval during which the operator observes and digests the results.
2. The characteristics of the message protocols themselves tend to produce highly correlated transmissions, as they often follow a request-acknowledgement-acknowledgement format [2], [3].
Therefore, among the parameters that we consider

†Contact author: david-nash@usma.edu.

to be important when producing simulated network traffic, we include autocorrelation and self-similarity [4]. The need to reproduce the characteristic of self-similarity in simulated network traffic has been addressed before [5], [6], [7]. In particular we are interested in the possibility of using a wavelet transform for the purpose of efficiently representing network traffic. By describing the data in terms of a limited number of wavelet coefficients, we may hopefully reconstruct the behavior of the traffic of interest.

II. Motivation

The simulation of computer networks is of considerable interest to the academic, military, and commercial communities. The realization that computers can be employed as high-speed communications devices, as well as increased abilities to employ networks of computers to expand the problem sizes manageable by present algorithms has produced a dramatic growth in the number and complexity of LANs. Whereas network design and maintenance was formerly something of an experimental process, the magnitude of these tasks currently calls for a more sophisticated approach. Consequently simulation has evolved as a method of testing the effects of changes to the network, without requiring actual adjustment of the in-place topology.

The generation of traffic is central to a network simulation. It can be troublesome, however, to create simulated traffic in a way that is computationally efficient, and at the same time exhibits characteristics which one would expect to observe in an actual network. An approach which may prove effective involves the discrete wavelet transform. Decomposition of a traffic data set using the transform yields a set of coefficients whose magnitudes reflect the contribution of basis functions across varying scales. It is well known that this technique may be used to compress the data signal by removing coefficients which are smaller than some established threshold. This approach underlies a class of compression techniques called *transform coding*. Theoretically, one should be able to recapture the majority of the energy of the signal with some smaller percentage of the total number of coefficients which would be required to reconstruct the signal exactly.

It is clear that there is a direct relationship between the computational complexity of signal generation, the number of coefficients used, and the error induced in the reconstruction. It is not obvious, however, what the effect of this method of compression would be upon the self-similarity attributes of the reconstruction. It is this effect which we propose to investigate.

III. Data set

The data set used for this study was a collection of observations made of a network in place at the Lawrence Berkeley Laboratory in January, 1994. The observations were made over the course of one hour, and recorded all packets arriving at or originating from the host site. This dataset has been used in a previous study of wide-area traffic [1], titled therein as LBL-PKT-5. We synthesized this data so that each observation in the data set represents a measurement of the number of packets which were transmitted during the interval since the preceding observation. The sampling rate was 1 sample per second. A portion of the total of $3,600$ observations in the data set is shown in Figure 1.

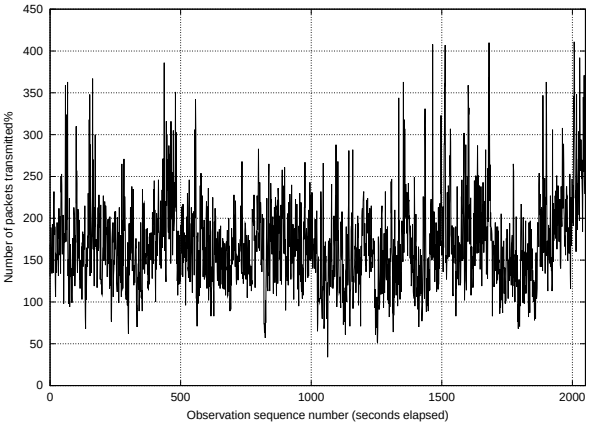

Fig. 1. A portion of the Ethernet traffic data set.

Our investigation is preliminary in the sense that we analyzed a single data set. We hypothesize that the intrinsic characteristics of the Ethernet protocol dominate the signal so far as self-similarity is concerned, and that usage patterns play a secondary role. Whether or not this is the case will determine the utility of our proposal to use the decomposition technique to generate traffic in a general setting. Our intention for the future is to apply this method to traffic from different environments to quantify the importance of this effect, and to use those observations to further refine the model. Since our analysis is essentially in the frequency domain, we do not anticipate that the observed relationship will show a strongly correlated effect in the time domain. In particular, we are purusing data sets from the 1998 and 1999 DARPA offline intrusion detection evaluations, as well as data captured during a Department of Defense Advanced Warfighter Experiment conducted in October, 2000.

IV. The discrete wavelet transform

In general, the transformation of a discrete signal $S[n]$ is accomplished by finding the coefficients c_k which satisfy:

$$S[n] = \sum_{k=0}^{\infty} c_k \varphi_k[n], \tag{1}$$

where the functions φ_k are functions whose union over the integers represents a basis for the vector space containing the signal of interest.

The efficient representation of signals in terms of basis functions has its roots in Fourier analysis. In classical Fourier analysis, a signal is represented in terms of a sum of sinusoidal basis functions, each of which is orthogonal to the other. The orthogonality of the basis functions results in a representation which is highly efficient, and which has practical implications with regard to the use of the magnitude of the coefficients to characterize the "strength" of the contribution made by each basis element.

In 1910, Haar showed that it was possible to represent a signal using *compact* basis functions [8]. Since that time, a rich theory has grown up around the use of compact basis functions (known as *wavelets*). The fact that Haar's basis functions did not extend to infinity in the time/space domain (as opposed to sinusoids, for example) promised the potential of representation techniques analagous to Fourier's, which would provide resolution in both the time and frequency domains. Daubechies [9] generalized this theory to include orthonormal bases, and in so doing laid the foundations for the application of wavelet analysis to essentially any finite-energy signal (i.e., any signal $f(x) : \int_{\infty}^{\infty} |f(x)|^2 \, dx < \infty$).

V. Measurement of self-similarity

Self-similarity is an assessment of a process' tendency to exhibit related behaviors or characteristics over different scales of time and/or space [10]. With regard to network traffic, it has been shown that the frequency of certain categories of packets exhibit the property of self-similarity across different partitionings of time. Leland *et al* provided an extensive examination of the self-similar nature of Ethernet traffic [4]. Their analysis employed several statistical metrics of self-similarity. One measure which is especially straightforward to apply is the *rescaled range statistic*, abbreviated R/S. Mandelbrot gave a detailed description of the calculation of the R/S statistic in [11]. Letting $X^*(t) = \sum_{k=1}^{t} X_k$, the R/S statistic $R/S(s,t)$ for some lag s and start time t is given by:

$$R/S(s,t) = \frac{R(s,t)}{S(s,t)}, \tag{2}$$

where

$$\gamma(s,t,u) = X^*(t+u) - X^*(t) - \left(\frac{u}{s}\right)\left[X^*(t+s) - X^*(t)\right], \tag{3}$$

$$R(s,t) = \max_{0 \le u \le s}\{\gamma(s,t,u)\} - \min_{0 \le u \le s}\{\gamma(s,t,u)\}, \tag{4}$$

and

$$S(s,t) = \sqrt{(1/s)\sum_{k=t+1}^{t+s} X_k^2 - \left[(1/s)\sum_{k=t+1}^{t+s} X_k\right]^2} \tag{5}$$

Mandelbrot's method is a graphical technique, whereby successive calculations of R/S are plotted on a log-log scale for various values of lag s and start time t. The slope of a straight-line fit for these values of R/S forms an estimate of the *Hurst parameter*, which is itself a measurement of tendency in the data to be self-similar [12]. An extensive discussion of techniques for the estimation of H may be found in [13].

VI. Procedure

The data was obtained in an ASCII file which included a number of extraneous data elements, such as IP addresses of source and destination hosts, TCP ports and number of bytes in the packet. These were removed using a Unix shell script. The data in its original format was at the scale of 1μsec. We aggregated this data to resolve at the 1 sec level for computational convenience. The self-similarity of the data set was measured according to the procedure described in section V using a lag increment and a start time increment of 100. A plot of the result is shown in Figure 2. Software which accomplishes the calculation of R/S was written in the C++ programming language, and the estimation of the Hurst parameter was done in Excel. Code for the program is included in the appendix. WaveLib, a public-domain library of subroutines [14], was called from this code to accomplish the wavelet decomposition, compression, and decompression.

We next decomposed the data by discrete wavelet transform, using a Daubechies D_8 wavelet. From this decomposition, we obtained reconstructions of the original data set using successively smaller subsets of the sorted list of all coefficients, such that the largest wavelet coefficient was always included. This was accomplished by setting a variable threshold T, and excluding from a particular reconstruction those coefficients which were smaller than T% of the largest one. In this fashion we obtained 25 reconstructed signals which were of progressively inferior quality in terms of their RMSE, but correspondingly superior

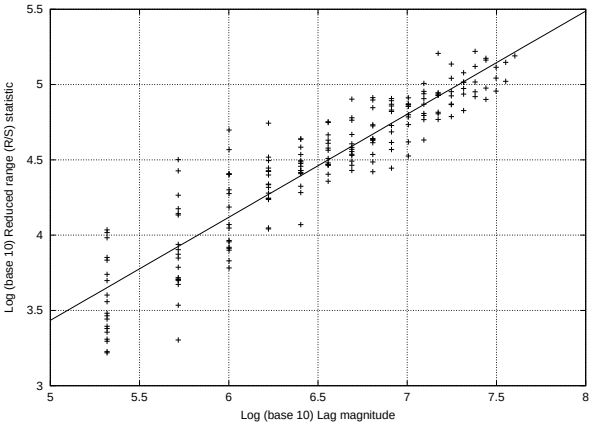

Fig. 2. Estimation of $\hat{H}$ for the original data set.

in terms of the efficiency (i.e., smaller numbers of coefficients) of their representation. Some samples of these reconstructed signals for various values of T are shown in Figures 3 through 6. The Hurst parameter for these reconstructed signals was then estimated in the same fashion as was the original data. It is apparent that even with substantial compression, the reconstructions capture a substantial amount of the original signal's character, and thus should be useful for the purpose of simulation.

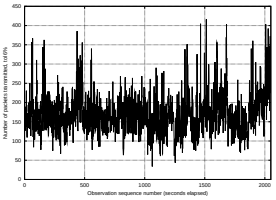

Fig. 3. Threshold = 6%, compression = 50.8%, 1007 coefficients.

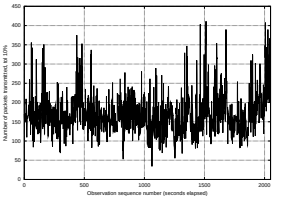

Fig. 4. Threshold = 10%, compression = 70.3%, 609 coefficients.

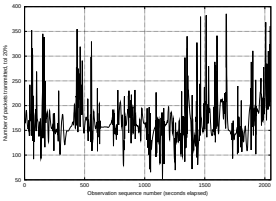

Fig. 5. Threshold = 20%, compression = 91.7%, 170 coefficients.

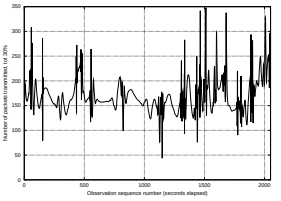

Fig. 6. Threshold = 30%, compression = 96.8%, 65 coefficients.

VII. Results

The results of the analysis are summarized in Table I. It would appear that there is a roughly linear relationship between the tolerance and the estimate of $\hat{H}$. A plot of a linear regression for this relationship is shown in Figure 7. As might be expected however, there is some question as to the independence of the residuals of this model. Figure 8 reveals the possible presence of a sinusoidal pattern in the residuals, which suggests that an autoregressive model may be more appropriate.

TABLE I
Estimates of the Hurst parameter $\hat{H}$.

Tolerance (%)	Output: Input	RMSE	$\hat{H}$
30	1:31.5	39.37	0.718
29	1:30.1	39.08	0.722
⋮	⋮	⋮	⋮
8	1:2.6	13.54	0.684
7	1:2.3	11.54	0.687
6	1:2.0	9.58	0.686
⋮	⋮	⋮	⋮
Original	1:1	0	0.686

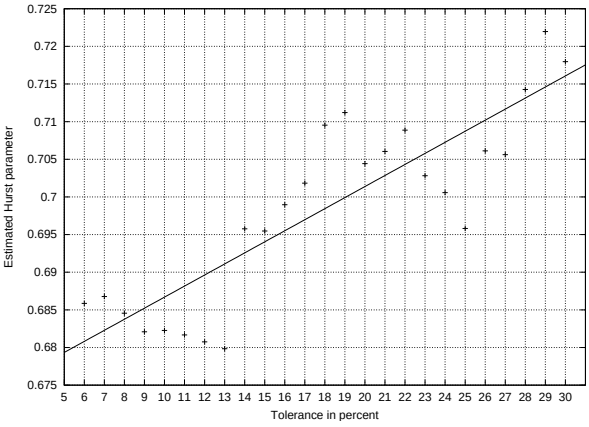

Fig. 7. Relationship of $\hat{H}$ with degree of compression.

VIII. Conclusions

It would appear that the estimated Hurst parameter is affected by reduction of the wavelet coefficients used to reconstruct the signal, to the extent that there is a general trend of increase in $\hat{H}$ with increasing compression. Further investigation to determine usable models seems reasonable, given that the residuals from a strictly linear model have zero mean.

References

[1] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modelling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, June 1995.

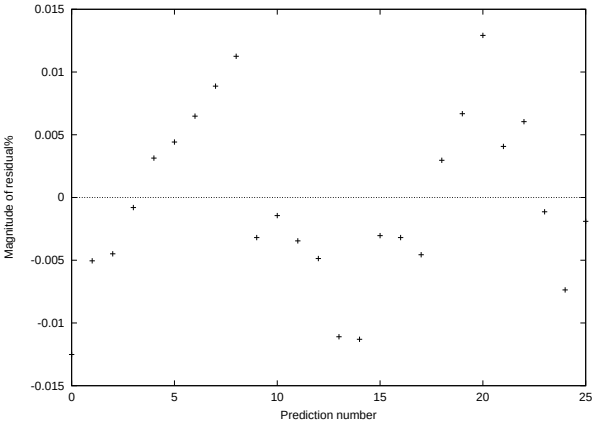

Fig. 8. Residuals from the linear model.

[2] M. Schwartz, *Telecommunication Networks: Protocols, Modeling, and Analysis*. Reading, Massachussetts: Addison-Wesley, 1987.

[3] J. H. B. Deane, C. Smythe, and D. J. Jeffries, "Long range order in network traffic dynamics." `http://www.ee.surrey.ac.uk/Personal/D.Jefferies/Selfsim/htmlpaper.html`, June 1996.

[4] W. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, February 1994.

[5] S. Robert and J. LeBoudec, "A Markov modulated process for self-similar traffic," technical report, Laboratoire de Reseaux de Communication, September 1995.

[6] V. Paxson, "Fast approximation of self-similar traffic," Technical report LBL-36750, Lawrence Berkeley Laboratory, April 1995.

[7] M. W. Garret and W. Willinger, "Analysis, modeling and generations of self-similar VBR video traffic," in *Proceedings of the 1994 ACM SIGCOMM Conference*, (London, UK), pp. 269–280, 1995.

[8] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, pp. 331–371, 1910.

[9] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, pp. 909–996, October 1988.

[10] H. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*. Rennselaer, New York: Springer-Verlag, 1992.

[11] B. B. Mandelbrot, "Some long-run properties of geophysical records," *Water Resources Research*, vol. 5, pp. 321–340, April 1969.

[12] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, vol. 116, pp. 770–799, 1951.

[13] O. Rose, "Estimation of the Hurst parameter of long-range dependent time series," Technical report TR-137, Department of Computer Science, University of Würzburg, February 1996.

[14] M. Bourges-Sévenier, "Réalisation d'une bibliothèque C de fonctions ondelettes," Technical report 864, Institut de Recherche en Informatique et Systémes Aléatoires, September 1994.