Automatic Scaling Range Selection for Long-range Dependent Network Traffic

Xiangdong Xia, Georgios Y. Lazarou, Thomas Butler
Department of ECE, Mississippi State University, Mississippi State, MS 39762
Email: {glaz}@ece.msstate.edu

Abstract— **In this paper, we present an adaptive search algorithm to automatically select the scaling range in the wavelet-based Hurst parameter estimation method. This algorithm is recursive and adaptive in nature, and it can select a scaling range consistent with human visual selection. In addition, it can be easily extended to automatically find the (approximately) linear regions of any curve. We tested our algorithm on 13 NLANR network traffic traces. The results show that our algorithm works well.**

I. Introduction

It is believed [1] that the high variability in Internet traffic is due to the *long-range dependence* (LRD) property of the traffic processes. In general, a (weakly) stationary discrete-time real-valued stochastic process $Y = \{Y_n, n = 0,1,2,\ldots\}$ with mean $\mu = E[Y_n]$ and variance $\sigma^2 = E[(Y_n - \mu)^2] < \infty$ exhibits LRD if $\sum_{k=1}^{\infty} r(k) = \infty$, where $r(k)$ measures the correlation between samples of Y separated by k units of time. If $\sum_{k=1}^{\infty} r(k) < \infty$, then Y is said to exhibit *short-range dependence* (SRD).

Common traffic models with LRD are based on self-similar processes. In traffic modeling, the term self-similarity is usually used to refer to the *asymptotically-second order self-similar* processes [1]: assume that Y has an autocorrelation function of the form $r(k) \sim k^{-\beta} L(k)$ as $k \to \infty$, where $0 < \beta < 1$ and the function L is slowly varying at infinity, i.e., $\lim_{k\to\infty} \frac{L(kx)}{L(k)} = 1 \ \forall x > 0$. For each $m = 1,2,3,\ldots$, let $Y^{(m)} = \{Y_n^{(m)}, n = 1,2,3,\ldots\}$ denote a new aggregated time series obtained by averaging the original series Y over non-overlapping blocks of size m, replacing each block by its sample mean: $Y_n^{(m)} = \frac{Y_{nm-m+1} + \cdots + Y_{nm}}{m}$.

The new aggregated discrete-time stochastic process $Y^{(m)}$ is also (weakly) stationary with an autocorrelation function $r^{(m)}(k)$. Then Y is called asymptotically second-order self-similar with self-similar parameter $H = 1 - \frac{\beta}{2}$ if, for all k large enough, $r^{(m)}(k) \to r(k)$ as $m \to \infty$. That is, Y is asymptotically second-order self-similar if the corresponding aggregated processes $Y^{(m)}$ become indistinguishable from Y at least with respect to their autocorrelation functions. By definition, asymptotically second-order self-similarity implies LRD and vice versa [1]

The parameter H is called the *Hurst parameter*. For general self-similar processes, it measures the degree of "self-similarity." For random processes suitable for modeling network traffic, the Hurst parameter is basically a measure of the speed of decay of the tail of the autocorrelation function. If $0.5 < H < 1$, then the process is LRD, and if $0 < H \leq 0.5$, then it is SRD. Hence, H is widely used to capture the intensity of long-range dependence in a traffic process. The closer H is to 1 the more long-range dependent the traffic is, and vice versa.

There are several methods for estimating H from a traffic trace. One of the most widely used is based on wavelets [2]. Given a traffic trace Y_n, H can be estimated as follows:

- First, for each *scale* j and *position* k, compute the wavelet coefficients: $d(j,k) =< Y_n, \Psi_{j,k}(n) >= \sum_{n=1}^{\infty} Y_n \Psi_{j,k}(n)$ where $\Psi_{j,k}(n) = 2^{-j/2}\Psi_0(2^{-j}n - k)$ and Ψ_0 is the (Daubechies) mother wavelet [2].
- Then, compute the wavelet energy μ_j for each scale j: $\mu_j = \frac{1}{N_j}\sum_{k=1}^{N_j} d^2(j,k)$ where N_j is the total number of wavelet coefficients at scale j.
- Next, make a plot of $log_2(\mu_j)$ versus scale j and apply linear regression over the curve region that *looks* linear. Compute the slope α.
- Finally, estimate the Hurst parameter as $\hat{H} = \frac{\alpha+1}{2}$.

The scaling behavior of internet traffic is not strictly self-similar, but rather more complex [5]. By using the wavelet method above, we will usually see that the $log_2(\mu_j)$ versus j curve is not strictly linear (for example, see Figures 1-3). Below a certain cut-off scale j_1, the scaling behavior is not self-similar. (It is believed that internet traffic below the lower cut-off scale j_1 exhibits multi-fractal scaling behavior.) Above a certain higher cut-off scale j_2, there will be few transformed wavelet coefficients, and the estimation of the Hurst parameter will be quite noisy [3]. When estimating the Hurst parameter, it is better to discard data in the fine scales less than j_1 and coarse scales greater than j_2.

So far, to estimate the Hurst parameter using the wavelet-based method, a visual inspection of the $log_2(\mu_j)$ vs. j plot is necessary to identify the linear trend region of the curve (see step 3 above). This becomes a problem because it is not objective. Furthermore, there are situations in which visual inspection is not possible such as real-time automatic Hurst parameter estimation. This problem is very important in the process of determining the Hurst parameter by the wavelet method.

In this paper, we present a robust, adaptive and recursive algorithm that automatically searches and determines the linear region of the curve. The algorithm recursively finds the scales j_1 and j_2 for which the curve over the range $[j_1, j_2]$ seems linear. This algorithm is based on the greedy algorithm principle. The global optimum solution can be obtained by using local optimum subsolutions. The algorithm is simple and fast, and thus it can be incorporated into the present wavelet-based method for real-time Hurst parameter estimation of network traffic. We tested our algorithm with 13 long traffic traces from the Auckland-II, Auckland-IV, NZIX-II, Bell Labs-I, and Abilene-I data archives [6].

The rest of this paper is organized as follows. Section II presents our proposed algorithm. Section III shows some scaling range results on real traffic traces using our algorithm. Our conclusions are presented in section IV.

II. The Algorithm

To determine the linear region, a natural approach is to use linear regression to model the observational data. Then, using a model to simulate the experimental data, it is important to measure how well that model matches the experimental data. Rather than the usual residual (the sum of the squares of the deviations), we use Theil's inequality coefficient U [4]. Because it is normalized, U is a data-independent measure of the goodness-of-fit and therefore better for our purposes. Theil's U is defined as follows:

$$U = \frac{\sqrt{(1/N)\sum_{j=1}^{N}(Y_j^s - Y_j^a)^2}}{\sqrt{(1/N)\sum_{j=1}^{N}(Y_j^s)^2} + \sqrt{(1/N)\sum_{j=1}^{N}(Y_j^a)^2}} \qquad (1)$$

where Y_j^s represents the simulated value, and Y_j^a represents the actual value. N is the total number of data points. The numerator is the square root of mean square error, and the denominator is used to scale U such that it will always be in the range of 0 and 1. U is 0 when the model's simulation results match the observational data exactly. U is 1 when the model's performance is worst.

The algorithm searches and determines the linear region of the curve. Given a $log_2(\mu_j)$ vs. j curve for a traffic trace Y_n, we begin by choosing $j_1 < j_2$ in the middle region[1] of the curve. Then, while the Theil's parameter for the range $[j_1, j_2]$ is below some given threshold γ_0, we expand the range either left to $[j_1 - 1, j_2]$ or right to $[j_1, j_2 + 1]$, always choosing the expansion with the smaller value of Theil's parameter. So as long as this process continues, the linear range $[j_1, j_2]$ will stretch to its optimal width.

The following shows the algorithm in detail:

1) Set the initial values of j_1 and j_2 as follows: $j_1 = \frac{M}{2} - 1$, $j_2 = \frac{M}{2} + 1$, where M is the maximum wavelet scale. Let L be the length of Y_n, then $M = log_2(L)$.
2) Using the least square method, fit a linear line between j_1 and j_2: $y(j) = kj + b + \varepsilon_j$. where $y(j) = log_2(\mu_j)$ and ε_j is the error associated with point j. Letting $N = j_2 - j_1 + 1$, the slope k and constant b are computed as follows:

$$k = \frac{N\sum_{j=j_1}^{j_2} jy(j) - \sum_{j=j_1}^{j_2} j \sum_{j=j_1}^{j_2} y(j)}{N\sum_{j=j_1}^{j_2} j^2 - (\sum_{j=j_1}^{j_2} j)^2}$$

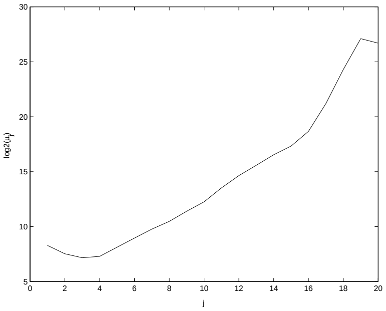

Fig. 1. $log_2(\mu_j)$ vs. scale j for the Auckland-IV traffic trace 20010301-310-0

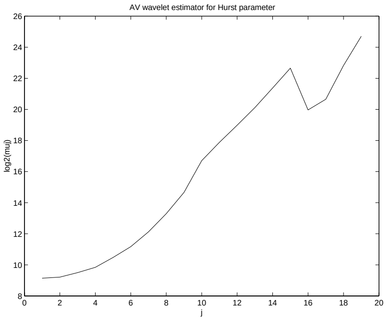

Fig. 2. $log_2(\mu_j)$ versus scale j for the Bell Labs-I traffic trace 20020519-151927.

$$b = \frac{\sum_{j=j_1}^{j_2} y(j)}{N} - k\frac{\sum_{j=j_1}^{j_2} j}{N}$$

3) Next, compute Theil's U parameter using (1) and data between j_1 and j_2. Note that here $Y_j^s = kj + b$, $Y_j^a = y(j), j = j_1, ..., j_2$: $\delta(j_1, j_2) = U$
4) If $\delta(j_1, j_2) \leq \gamma_0$ then compute $\delta(j_1 - 1, j_2)$ and $\delta(j_1, j_2 + 1)$.
    a) If $\delta(j_1 - 1, j_2) < \delta(j_1, j_2 + 1)$ and $\delta(j_1 - 1, j_2) < \gamma_0$ then set $j_1 \leftarrow j_1 - 1$ and go back to Step 2.
    b) If $\delta(j_1, j_2 + 1) < \delta(j_1 - 1, j_2)$ and $\delta(j_1, j_2 + 1) < \gamma_0$ then set $j_2 \leftarrow j_2 + 1$ and go back to Step 2.
5) else Stop.

Obviously, the value of the residual error threshold γ_0 determines the *goodness-of-fit*. We recommend using the threshold value γ_0=0.015. On one hand, Table II shows that this threshold produces linear ranges that closely match the ranges obtained by visual inspection. On the other hand, such γ_0 is close enough to zero to yield a good estimate of the Hurst parameter.

III. Results

We implemented the above algorithm in Matlab and obtained *automatically* the linear trend ranges $[j_1, j_2]$ for the 13 long traffic traces.

To illustrate how the algorithm works, we take traffic trace 20010301-310-0 as an example to demonstrate how the

[1] All traffic traces analyzed for this paper exhibit monofractal behavior that spans the middle region of the wavelet curve. Therefore, for simplicity, we let the algorithm to start progressing from the middle region of the curve. However, this is not a necessary condition of the algorithm.

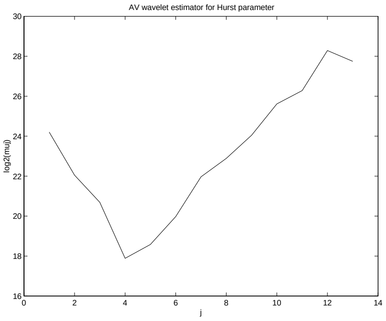

Fig. 3. $log_2(\mu_j)$ versus scale j for the Abilene-I traffic trace IPLS-KSCY-20020814-090000-1.

algorithm gradually searches for the optimal linear scaling range. Table I shows detailed search results step by step.

TABLE I

STEP BY STEP ILLUSTRATION OF LINEAR RANGE SELECTION BY ADAPTIVE SEARCH ALGORITHM FOR THE CURVE IN FIGURE 1.

number of points(N)	[j1,j2]	stretch direction	theil U
3	[9,11]		0.0039
4	[9,12]	right	0.0035
5	[9,13]	right	0.0032
6	[9,14]	right	0.0033
7	[8,14]	left	0.0036
8	[8,15]	right	0.0042
9	[8,16]	right	0.0041
10	[7,16]	left	0.0052
11	[6,16]	left	0.0064
12	[5,16]	left	0.0075
13	[4,16]	left	0.0084

In Table II, the linear scaling range obtained using visual inspection and the adaptive search algorithm are compared. Clearly, the algorithmically obtained linear regions for the $log_2(\mu_j)$ vs. j curves are consistent with ones obtained from visual inspection. To estimate the Hurst parameter we used the Matlab routine *LDestimate.m* which is available in [7]. We observe from Table II that the estimated Hurst parameter values for two of the IPLS traffic traces are greater than one. The explanation of this is beyond of the scope of this paper. However, the reader should note that there are well-defined self-similar processes with stationary increments, infinite second moments, and $H \geq 1$ [8].

IV. CONCLUSION

In this paper, we presented a robust and adaptive algorithm that can be incorporated in the wavelet-based method of estimating the Hurst parameter for traffic traces with LRD, that is, traffic traces that exhibit monofractal behavior over a wide range of time scales. It provides a systematic and objective way to determine the linear trend region instead of subjective human visual inspection. It is based on the greedy algorithm principle and uses Theil's inequality measure. Our algorithm automatically searches and determines the linear region of the $log_2(\mu_j)$ vs. j curve. That is, the algorithm recursively finds the scales j_1 and j_2 for which the curve over $[j_1, j_2]$ looks linear.

TABLE II

COMPARISON BETWEEN ALGORITHMICALLY OBTAINED LINEAR TREND REGIONS AND VISUAL INSPECTION FOR THE AUCKLAND TRAFFIC TRACES.

Traffic Trace	by inspection	by algorithm	$\hat{H}$
19991129-134258-0	[4,20]	[3,20]	0.813
19991129-134258-1	[6,20]	[7,20]	0.968
19991201-192548-0	[3,19]	[3,19]	0.890
nzix-II	[4,15]	[4,16]	0.979
20010220-226-0	[3,15]	[4,14]	0.836
20010220-226-1	[4,15]	[4,15]	0.891
20010301-310-0	[4,16]	[4,16]	0.919
20010301-310-1	[4,16]	[4,16]	0.903
20020519-000000	[5,10]	[2,9]	0.564
20020519-151927	[2,15]	[4,15]	0.964
IPLS-CLEV-20020814-0	[4,12]	[5,12]	1.299
IPLS-CLEV-20020814-1	[4,12]	[4,12]	0.965
IPLS-KSCY-20020814-1	[4,12]	[4,12]	1.121

This algorithm is an improvement to the process of estimating the Hurst parameter in the wavelet domain. It can be easily integrated with various traffic control schemes that require *real-time* Hurst parameter estimations. We tested our algorithm with 13 long NLANR traffic traces and obtained satisfactory results. These traffic traces exhibit monofractal behavior that spans the middle region of the wavelet curve, otherwise the processes would have been classified as multifractal. Therefore, for simplicity, we let the algorithm to start progressing from the middle region of the curve. However, this is not a necessary condition of the algorithm. It can be easily modified such that, for example, the initial values of j_1 and j_2 are set to be the two end scale values (i.e., first and last). Also, this algorithm can be extended to automatically determine all the linear regions of a curve that have different slopes, such as wavelet curves of multifractal processes.

REFERENCES

[1] K. Park, W. Willinger, Editors, *Self-Similar Network Traffic And Performance Evaluation*, John Wiley & Sons, 2000.

[2] P. Abry and D. Veitch, " Wavelet analysis of long-range dependent traffic," IEEE Trans. Information Theory, vol.44, no.1,pp.2-15,1998.

[3] S. Giordano, S. Miduri, M.Pagano, F.Russo, S.Tartarelli, "A Wavelet-based approach to the estimation of the Hurst Parameter for self-similar data," Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on , Volume: 2, 1997 Page(s): 479 -482 vol.2.

[4] Theil, Henri. "Statistical Decomposition Analysis," Amsterdam, North-Holland Publishing Company, 1972.

[5] A.Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The Changing Nature of Network Traffic: Scaling Phenomena," *Computer Communication Review,* Vol.28, no.2, 1998.

[6] NLANR PMA. Special Traces Archive.[Online]. Available: http://pma.nlanr.net/Special/index.html

[7] D. Veitch. (2002, Nov.). Code for the estimation of Scaling Exponents: *LDestimate.m*. [Online]. Available: http://www.cubinlab.ee.mu.oz.au/ darryl/secondorder_code.html

[8] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall/CRC, 1994.