ESTIMATING LONG-RANGE DEPENDENCE IN IMPULSIVE TRAFFIC FLOWS

Xueshi Yang, Athina P. Petropulu

ECE Department, Drexel University
Philadelphia, PA 19104, USA
{yxs, athina}@cbis.ece.drexel.edu

Jean-Christophe Pesquet

Université de Marne-la-Vallée, Cité Descartes 5
Boulevard Descartes, Champs sur Marne 77454
Marne la Vallée Cedex 2, France
pesquet@univ-mlv.fr

ABSTRACT

Traffic flow in high-speed data network systems is often impulsive and long-range dependent.[1] Impulsiveness implies a heavy-tailed marginal distribution, thus lack of finite second-order statistics. Hence, traditional methods for quantifying the long-range dependence of traffic based on its second-order statistics are not applicable. Long range dependence and self-similarity play an important role in traffic engineering. We have recently shown that the generalized codifference can quantify the dependence structure of impulsive self-similar processes, such as high-speed network traffic. In this paper, we propose an estimator for the generalized codifference and provide the conditions for it to be asymptotically consistent. We show that these conditions are satisfied for the EAFRP which is a process proposed for modeling high-speed network traffic. We provide simulations results to demonstrate the properties of the proposed estimator, and show how it can be a useful tool in maintaining fairness among users sharing limited network resources.

1. INTRODUCTION

Self-similar traffic flow exhibits time rescaling invariability, in sharp contrast to Markovian flow, which loses dependence in coarsified time scales. Since bandwidth provisioning is operated upon rescaled traffic to some extent, self-similarity carries profound importance in resource allocation strategies for network design. For example, previous queueing analysis [3] showed that under self-similar traffic input, buffer overflow probability decreased hyperbolically instead of exponentially as the buffer size increased, the Hurst parameter determining the decay rate. Nowadays, as traffic sources continue to diversify, it is important to identify the Hurst parameter of distinct sources, so as to provide a reliable benchmark for resource allocation, e.g. buffer space assignment and bandwidth reservation. Provided that the self-similar process has finite second-order statistics, various method have been proposed to evaluate the Hurst parameter, such as the log-variance time plot and periodogram based methods. However, in state-of-art high-speed data networks, traffic flows often exhibit strong impulsiveness, in other words, are marginally heavy-tail distributed. In such scenarios, where second-order statistics do not exist, the generalized codifference can provide a measure of dependence structure, playing a similar role to the autocorrelation.

For a stationary stochastic process $x(k), k \in \mathbb{Z}$, the generalized codifference (GC) [7, 8] is defined as

$$\begin{aligned}\tau(n) &= -\ln E\{e^{is(x_{k+n}-x_k)}\} + \ln E\{e^{isx_{k+n}}\} \\ &\quad + \ln E\{e^{-isx_k}\}, \qquad s \in \mathbb{R}. \qquad (1)\end{aligned}$$

We define $\{x(k)\}$ to be long-range dependent (LRD) in the generalized codifference sense, if

$$\tau(n) \sim Cn^{-\beta}, \quad \text{as } n \to \infty, \qquad (2)$$

where β and C are real positive numbers. β is the LRD index, which is related to the Hurst parameter H by $\beta = 2 - 2H$. The smaller β is, the stronger long-range dependence $\{x(k)\}$ possesses.

The EAFRP (Extended Alternating Fractal Renewal Process) was proposed in [8] for constructive modeling of high-speed network traffic, and was shown to capture both traffic impulsiveness and LRD in a GC sense. It is defined as:

$$x_k = A_k V_k \qquad (3)$$

where V_k is a stationary renewal process alternating between 0 and 1. The renewal time intervals and reward process A_k (corresponding to each 1 state) are drawn from independent heavy-tailed distributions.

Let us define an empirical estimator for GC as follows:

$$\begin{aligned}\widetilde{\tau}_K(n) &= -\ln \frac{1}{K}\sum_{k=1}^{K} e^{is(x_{k+n}-x_k)} + \ln \frac{1}{K}\sum_{k=1}^{K} e^{isx_{k+n}} \\ &\quad + \ln \frac{1}{K}\sum_{k=1}^{K} e^{-isx_k}. \qquad (4)\end{aligned}$$

First we will provide the conditions for $\widetilde{\tau}_K(n)$ to be a consistent estimator of $\tau(n)$. We will show that these conditions are satisfied for the EAFRP [8].

We will provide simulation results to demonstrate the properties of the proposed estimator, and show how it can

This work was supported by National Science Foundation under grant MIP-9553227

[1] In traffic engineering, self-similarity and long-range dependence are improperly used as interexchangeable terms although they correspond to mathematically different concepts.

be a useful tool in classifying traffic flows according to their LRD index. Quantifying the traffic LRD is important in network resources provisioning, such as buffer allocation. Given certain traffic load, the stronger the LRD, the larger the buffer requirement in order to maintain a predefined Quality of Service (QoS). Using the LBNL Network Simulator, it is shown that the proposed estimator can be used as a metric in classifying traffic sources into categories with different network resource requirements.

2. CONSISTENCY OF THE GC ESTIMATOR

To prove the consistency of $\widetilde{\tau}_K(n)$, the following results are needed. Note that we will be mainly interested in the convergence in probability of the considered random sequences.

Lemma 1 *Let x_k, $k \in \mathbb{Z}$, be a stationary LRD process with first order characteristic function Φ. If $s \in \mathbb{R}$ and $\Phi(s) \neq 0$, then*

$$\Psi_K = \ln \frac{1}{K} \sum_{k=1}^{K} e^{isx_k}$$

is a consistent estimator of $\ln \Phi(s)$.

Proof : Let $y_k = e^{isx_k}$. It is clear that y_k is a second-order stationary process as its magnitude is equal to 1. A necessary and sufficient condition for y_k to be mean ergodic, i.e.

$$\frac{1}{K} \sum_{k=1}^{K} y_k \to E\{y_k\} \tag{5}$$

in the mean square sense is that its covariance function tends to zero as time lag tends to ∞. The covariance function can be expressed as

$$c(n) = |\Phi(s)|^2 \Big(\frac{E\{e^{is(x_{k+n} - x_k)}\}}{E\{e^{isx_{k+n}}\}E\{e^{-isx_k}\}} - 1 \Big). \tag{6}$$

Due to the characterization of the LRD in (2) and the expression of $\tau(n)$ in (1),

$$\frac{E\{e^{is(x_{k+n} - x_k)}\}}{E\{e^{isx_{k+n}}\}E\{e^{-isx_k}\}} \sim e^{-Cn^{-\beta}}, \qquad n \to \infty \tag{7}$$

which implies that $c(n) \to 0$. As mean square convergence entails convergence in probability, $K^{-1} \sum_{k=1}^{K} y_k$ converges in probability to $\Phi(s)$. Furthermore, as $\ln(\cdot)$ is a continuous function on $\mathbb{C} \setminus \{0\}$, we deduce that

$$\Psi_K \to \ln \Phi(s), \qquad K \to \infty. \tag{8}$$

As shown in [8], the assumptions in this lemma are satisfied for the EAFRP.

By proceeding in a way similar to the previous proof, it can be shown that

Lemma 2 *Let x_k be a stationary process and, for $m \in \mathbb{N}$ and $s \in \mathbb{R}$, let $\Phi^{(2)}(s, -s; m) = E\{e^{is(x_{k+m} - x_k)}\}$ be its second-order characteristic function evaluated at $(s, -s)$. As long as the process e^{isx_k} is autocorrelation ergodic in the mean square sense and $\Phi^{(2)}(s, -s; m) \neq 0$, it holds:*

$$\ln \frac{1}{K} \sum_{k=1}^{K} e^{is(x_{k+m} - x_k)} \to \ln \Phi^{(2)}(s, -s; m) \quad \textit{as } K \to \infty. \tag{9}$$

A necessary and sufficient condition for the process e^{isx_k} to be autocorrelation ergodic is [6]:

$$E\{e^{is(x_k - x_{k+m} - x_{k+n} + x_{k+n+m})}\} \to |\Phi^{(2)}(s, -s; m)|^2 \text{ as } n \to \infty. \tag{10}$$

Proposition 1 *Let x_k, $k \in \mathbb{Z}$, be an EAFRP process. Its generalized codifference estimator, $\widetilde{\tau}_K(n)$ is a consistent estimator of $\tau(n)$.*

Proof:

It is sufficient to show that the assumptions of Lemma 1 and 2 are satisfied by the EAFRP process. Indeed, in this case, we have, as $K \to \infty$:

$$\widetilde{\tau}_K(n) \to -\ln \Phi^{(2)}(s, -s; n) + \ln \Phi(s) + \ln \Phi(-s) = \tau(n). \tag{11}$$

In Appendix A we show that the conditions of Lemma 2 are satisfied for x_k, while the applicability of Lemma 1 is a direct consequence of the results in [8].

To estimate the LRD index, ie. β in (2), a least mean square error line is fitted in the log-log scale plot of $\widetilde{\tau}_K(n)$. The slope of the fitted line is a consistent estimation of β as $n \to \infty$ and $K \to \infty$. This is due to the fact that $\widetilde{\tau}_K(n)$ is a consistent estimator of $\tau(n)$, the consistency of integrated squared error (ISE) estimators[5] implies the consistent estimation of β.

3. SIMULATIONS

This section is divided into two parts. In part (A) we demonstrate the statistics of the proposed estimator for EAFRP process using simulations. In part (B) we consider real traffic traces fed in a queuing system simulated via the LBNL Network Simulator, and show that the LRD index estimated through the proposed estimator is an important parameter in network resource provisioning.

A) *Evaluation of estimator performance*

The EAFRP is a renewal processes, alternating between ON(1) and OFF(0) states, the durations of which are heavy-tail distributed with indices α_1 and α_0, respectively. The rewards during ON states are also heavy-tail distributed with tail index α_A. As it was shown in [8] the GC is a power-law decaying function, with exponent $-(\min(\alpha_0, \alpha_1) - 1)$. EAFRPs were constructed based on $\alpha_0 = 1.9$, $\alpha_1 = 1.1, ..., 1.9$ and $\alpha_A = 1.4$. $\widetilde{\tau}_K(n)$ is computed, and a least squares line is fitted on the log-log plot of $\widetilde{\tau}_K(n)$.[2] The estimated mean slope and the corresponding variance based on 50 Monte Carlo simulations is shown in Fig. 1.

Although in theory the variance of a marginally heavy-tailed process is infinite, it is occasionally argued in the literature that the empirical variance of a finite length segment would be finite, thus a log-variance plot could still be used to estimate the LRD index. The least squares line fitted in the log-variance plot gives the LRD index, which is equal to the exponent of the power-law decaying autocorrelation function of the data. The log-variance results are also included in Fig. 1. It clearly indicates that the log-variance plot method is of no use in this case, while

[2] The estimation is thus sensitive to the fitting range. In our experiments, the first 50 points of $\widetilde{\tau}_K(n)$ are used.

the proposed estimator performs reasonably well. Fig. 2 shows the variance of the GC estimator decreases as the data length increases.

B) *Classification of real traffic based on the proposed estimator*

In self-similar traffic engineering, long-range dependence plays an important role in network resources provisioning. Traffic flows with long-range dependence causes buffer overflow more frequently than sources with only short-range dependence. On the other hand, most of the existing resources allocation strategies were developed based on the assumption that the traffic sources are following arriving patterns such as Markovian modulated traffic. They do not take the long-range dependence effect into account. Hence, traffic flows with long-range dependence are often being treated unfairly, in comparison with Markovian sources. For example, our initial simulations show that the Random Early Drop (RED)[2] algorithm often punishes self-similar sources more severely than necessary by marking or dropping packets from them. Hence, identifying the long-range dependence in already diversified traffic streams is essential to network resources allocation strategies design.

The generalized codifference estimator could be an useful tool in classifying traffic streams with long-range dependence.

Figure 3(a) shows a typical example of real traffic traces collected from the 100Mbps Ethernet located at ECE department, Drexel university. Traffic traces are displayed as bytes per second. The impulsive outlook of the data is evident from that figure.

The traffic trace is formatted as a $n \times 2$ array. The first column and the second column correspond to the time (in microseconds) until next packet to be generated and its size respectively.

To generate another traffic trace with the same degree of impulsiveness but with weaker LRD, we modified the original data trace as follows. We partitioned the data trace into large blocks, e.g 200×2, and randomly re-ordered these blocks. This operation partially removed the long-range dependence structure of the original data. The re-ordered trace is shown in Fig. 3(b).

We applied the proposed estimator on both data traces, and plotted the estimated GC in log-log scale (see Fig. 4). Comparing the minimum mean-square fitted lines in Fig. 4 (a) and (b), the original data set and the re-ordered one has an estimated slope of -0.2971 and -0.4465 respectively. As expected, the re-ordered data trace exhibits weaker dependence.

The queueing simulation is performed using the LBNL Network Simulator platform. The simulated network topology is shown in Fig. 5. It consists of 5 nodes. Source nodes 0, 1, 2 are competing the buffer space (20 packets) in node 3, and bottleneck link between node 3 and node 4 to reach destination node 4. Link configurations are shown in Fig. 5. Numbers annotated above links are the link bandwidth and delay respectively. Nodes 0, 1, 2 are UDP senders, and node 4 functions solely as a sink. Traffic generated from nodes 0 and 1 correspond to the traffic trace shown in Fig. 3(a), and (b) respectively. Real traffic collected from another UNIX terminal forms the traffic stream stemming from node 2. Queue objects used in node 3 are the DropTail type, which implements FIFO scheduling and drop-on-overflow buffer management typical of most present-day Internet routers.

The packet dropping process is shown in Fig. 6(a)(b) for node 0 and node 1 respectively. Note that to shorten the simulation time, all the traffic traces are compressed by a factor of 10, i.e. the first column in the data trace file is divided by 10 before they are input to the simulator. Thus, the total simulation time shown is 1000 seconds, instead of the original 10^4 seconds. The total packets dropped from node 0 is 59529 Bytes, corresponding a bit loss ratio of 1.28%, vis-á-vis 54776 from node 1, corresponding 1.17%. It is easily seen that the traffic source with stronger long-range dependence gets smaller throughput than its counterparts with weaker dependence structure. Hence, it is necessary to allocate more buffer space to sources with stronger long-range dependence, as to maintain fairness among different users when they are sharing limited network sources.

4. APPENDIX A

In [8] it was shown that:

$$\Phi(s,-s;m) = [(\Phi_A(s)-1)\eta+1][\Phi_A(-s)-1)\eta+1] + L(m) \qquad (12)$$

where $\Phi_A(s)$ denotes the characteristic function of the amplitude of the ON states of the EAFRP process (heavy-tailed); η is the mean of the AFRP model amplitude, and $L(m)$ denotes a function that decays in a power-law fashion with m.

Also in [8] it was shown that:

$$\Phi(s) = 1 + \Phi_A(s)\eta - \eta \qquad (13)$$

Based on (12) and (13) it can be seen that $c(m)$ is a function of $L(m)$, thus tends to 0 as $m \to \infty$.

To give a sketch of the proof of sufficient condition (10), let us consider the residue life T of the EAFRP at $t = k+m$. It is shown in [8] that $P\{n-m < T\}$ is a power-law decaying function of $(n-m)$. It can be shown that:

$$E\{e^{is[(x_k - x_{k+m})-(x_{k+n}-x_{k+n+m})]}\} = |\Phi(s,-s,m)|^2 P(n-m > T) + L_1(n-m) \qquad (14)$$

where $L_1(n-m)$ is a power-law decaying function of $n-m$. The proof is omitted due to lack of space. For arbitrary finite m, the last term tends to zero while $n \to \infty$.

REFERENCES

[1] J. Beran, *Statistics for Long-Memory Processes*, Champman & Hall, New York, 1994.

[2] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", *IEEE/ACM Trans. on Networking*, Vol.1, No.4:397-413, Aug. 1993.

[3] D. Heath, S. Resnick, and G. Samorodnitsky, "Patterns of buffer overflow in a class of queues with long memory in the input stream", *Annals of Applied Probability*, Vol.7:1021-1057, 1997.

[4] A.P. Petropulu, J-C. Pesquet, X. Yang, "Power-law shot noise and relationship to long-memory processes," *IEEE Trans. on Sig. Proc.*, Vol.48, No.7:1883-1892, July 2000.

[5] C.R. Heathcote, "The Integrated Squared Error Estimation of Parameters,", *Biometrika*, Vol.64, No.2:255-264, 1977.

[6] J.G. Proakis, D.G. Manolakis, *Digital Signal Processing*, Prentice Hall, 1996.

[7] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random processes: Stochastic Models with Infinite Variance*, New York: Chapman and Hall, 1994.

[8] X. Yang, A.P. Petropulu, "The extended alternating fractal renewal process for modeling traffic in high-speed communication networks", *IEEE Trans. on Sig. Proc.*, submitted in 2000.

Estimation of LRD index

Theory	Proposed GC estimator	Log-Variance graph
0.1	0.1268(0.0029)	0.6709(0.0507)
0.2	0.1942(0.0106)	0.7983(0.0831)
0.3	0.3016(0.0139)	0.7810(0.0574)
0.4	0.3671(0.0252)	0.8659(0.0478)
0.5	0.5032(0.0253)	0.8722(0.0295)
0.6	0.5776(0.0376)	0.8881(0.0198)
0.7	0.6530(0.0634)	0.9168(0.0251)
0.8	0.7318(0.0495)	0.9806(0.0128)
0.9	0.8252(0.0749)	0.9618(0.0136)

Figure 1: Mean (variance) of the LRD index obtained based on 50 independent realizations of an EAFRP of length $18,000$.

Estimation of LRD index at different lengths

N	3000	6000	9000	12000	15000	18000
Mean	0.2730	0.2941	0.3070	0.3016	0.3136	0.3034
Var	0.0129	0.0116	0.0085	0.0062	0.0059	0.0060

Figure 2: Mean and variance of the proposed GC estimator obtained based on 50 different realizations of an EAFRP with $\alpha_0 = 1.9$, $\alpha_1 = 1.3$.

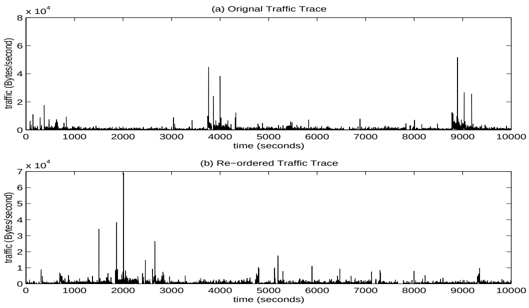

Figure 3: (a) Real traffic trace (b) re-ordered trace.

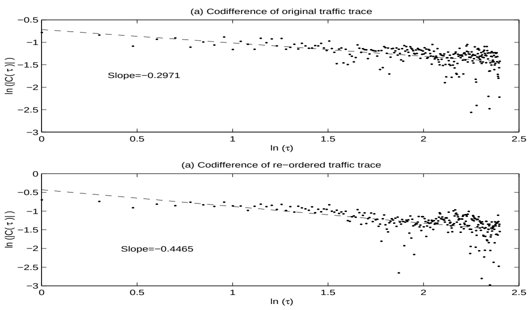

Figure 4: Codifference of (a) real traffic trace, (b) re-ordered trace.

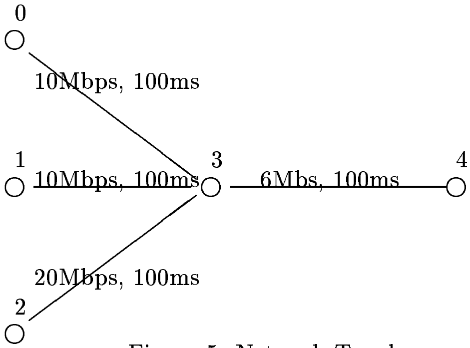

Figure 5: Network Topology

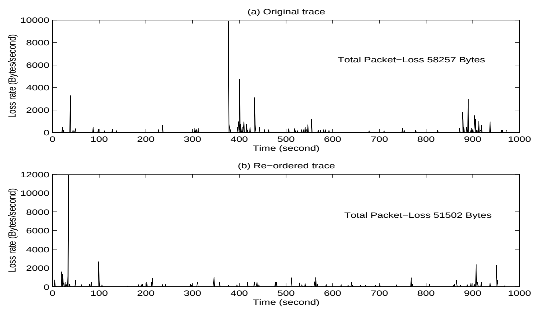

Figure 6: Loss-rate of (a) real traffic trace, (b) re-ordered trace.