Statistical Error Analysis on Recording LRD Traffic Time Series

Ming Li

School of Information Science & Technology, East China Normal University
Shanghai 200026, PR China
ming_lihk@yahoo.com

Abstract. Measurement of LRD traffic time series is the first stage to experimental research of traffic patterns. From a view of measurement, if the length of a measured series is too short, an estimate of a specific objective (e.g., autocorrelation function) may not achieve a given accuracy. On the other hand, if a measured series is over-long, it will be too much for storage space and cost too much computation time. Thus, a meaningful issue in measurement is how to determine the record length of an LRD traffic series with a given degree of accuracy of the estimate of interest. In this paper, we present a formula for requiring the record length of LRD traffic series according to a given bound of accuracy of autocorrelation function estimation of fractional Gaussian noise and a given value of H. Further, we apply our approach to assessing some widely used traces in the traffic research, giving a theoretical evaluation of those traces from a view of statistical error analysis.

1 Introduction

The Internet is a complex system such that conventionally scientific computations are quite limited in the performance research of the global Internet. Therefore, measurement plays a key role in the performance research because measured data of real traffic reflect the information about real-life situations of the global Internet under current protocols and infrastructure.

By analyzing measured data, findings regarding traffic were achieved in the last decade. In summary, 1) traffic is of long-range dependence (LRD), and 2) traffic is asymptotically self-similar [1]. The research of this paper will show that the particularity of LRD is also reflected in measurement.

Recording traffic is the first stage for the experimental research of traffic patterns. Here, we ask for a question how to validate the reasonableness of measured traffic data. To explain this question, we ask for another question whether there was another global Internet that was superior to the current one we are using so that it could be used for measurement validation, e.g., data validation/assessment, in the standardization sense. Unfortunately, the answer is NO. The global Internet has the property of *uniqueness*. In addition, simulating the Internet encounters painful difficulties [2]. For those reasons, conventional approaches for validation/assessment of measurement

H. Jin et al. (Eds.): NPC 2004, LNCS 3222, pp. 403-406, 2004.

data in the field of measurement (e.g., [3]) fail for the Internet traffic measurement. Hence, the theoretical research in measurement of LRD traffic is expected.

For measuring a random sequence, an important thing is that a measured sequence should have enough length so as to provide an enough accurate estimate of an objective (e.g., autocorrelation function (ACF)). In the field of measurement, however, length requirements of a measured random sequence are traditionally for those with short-range dependence (SRD), e.g., [4]. Intuitively, length requirements of LRD sequences should be distinctly different from those of SRD sequences because LRD processes evidently differ from SRD ones. However, we have not seen any reports about record length requirements for traffic measurement, to our best knowledge (except Li's early note [5]). This paper will show that the length requirement of a measured LRD sequence does drastically differ from that of SRD one. Note that the result in this paper is based on ACF estimation of fractional Gaussian noise (FGN). However, parameters to be considered in practice may not be ACF of FGN in monofractal but others, e.g., the Hurst function [6]. Therefore, the result in this paper may be conservative but it may yet be a reference guideline for record length of traffic in academic research and practice.

The rest of paper is organized as follows. In Section 2, we present the formula for requiring record length of measured LRD traffic with a given accuracy and a given value of H based on ACF estimation of FGN. Discussions are given in Section 3 and conclusions in Section 4.

2 Upper Bound of Standard Deviation

Denote $x(i) = x(t_i)$ ($i = 0, 1, 2, \ldots$) as a traffic trace, representing the number of bytes in a packet on a packet-by-packet basis at the time t_i. Mathematically, $x(i)$ is LRD if its ACF $r(k)$ is non-summable while $x(i)$ is called asymptotically self-similar if $x(ai)$ ($a > 0$) asymptotically has the same statistics as $x(i)$.

In mathematics, the true ACF of $x(i)$ is computed over infinite interval. However, any physically measured data sequences are finite in record length. Let a positive integer L be the data block size of $x(i)$. Then, $r(k)$ is estimated over finite interval. As known, a useful (actually widely used) model of traffic is FGN [7] [8]. Its normalized ACF is given by $0.5[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$, where H is the Hurst parameter. We take it as a representative of LRD traffic for our research about record length.

Suppose $r(\tau)$ is the true ACF of FGN and $R(\tau)$ is its estimate with L length. Then, R is a random variable. Let $M^2(R)$ be the mean square error in terms of $R(\tau)$. Then, $M^2(R) = \mathrm{Var}(R)$ [4]. We aim at finding a relationship that represents $M^2(R)$ as a two-dimension function of L and H so as to establish a reference guideline for requiring record length for a given degree of accuracy. We represent this relationship by the following theorem.

Theorem. Let $x(t)$ be a FGN function with $H \in (0.5, 1)$. Let $r(\tau)$ be the true ACF of $x(t)$. Let L be the block size of data. Let $R(\tau)$ be an estimate of $r(\tau)$ with L length. Let $\mathrm{Var}[R(\tau)]$ be the variance of $R(\tau)$. Then,

$$\mathrm{Var}[R(\tau)] \le \frac{\sigma^4}{L(2H+1)}[(L+1)^{2H+1} - 2L^{2H+1} + (L-1)^{2H+1}], \quad (1)$$

where σ^2 is the variance of FGN.

The proof of Theorem is omitted due to the limit space. Without losing the generality, we consider $\sigma = 1$. Denote $s(L, H)$ as the bound of standard deviation in the normalized case. Then, one has

$$s(L, H) = \sqrt{\frac{1}{L(2H+1)}[(L+1)^{2H+1} - 2L^{2H+1} + (L-1)^{2H+1}]}. \quad (2)$$

Following (2), we see that $s(L, H)$ is an increasing function of H.

3 Discussions

From (2), it is seen that a large L is required for a large value of H (strong LRD) for a given s. In engineering, accuracy is usually considered from the perspective of order of magnitude. When H = 0.55, 0.75 and 0.95, one has $s(L,H)\big|_{L=2^7,\,H=0.55}$ = 0.118, $s(L,H)\big|_{L=2^8,\,H=0.75}$ = 0.306, and $s(L,H)\big|_{L=2^{23},\,H=0.95}$ = 0.621. These show that Ls vary in orders of magnitude when H = 0.55, 0.75 and 0.95 for a given s, implying a series with larger value of H requires larger L for a given s.

An exact value of $s(L, H)$ usually does not equal to the real accuracy of the correlation estimation of a measured LRD-traffic sequence because FGN is only an asymptotical expression for real traffic [9] and traffic is multi-fractal in nature. On the other hand, there are errors in data transmission, data storage**,** measurement, numerical computations, and data processing. In addition, there are many factors causing errors and uncertainties due to the natural shifts, e.g., various shifts occurring in devices, or some purposeful changes in communication systems. Therefore, the concrete accuracy value is not as pressing as accuracy-order for the considerations in measurement design. For that reason, we emphasize that the contribution of $s(L, H)$ lies in that it provides a relationship between s, L and H for a reference guideline in the design stage of measurement.

Table 1 lists some well known traces on WAN. Now, we evaluate 1Lbl-pkt-4.TCP of 1.3×10^6 length, which is the shortest one in Table 1. For H = 0.90 (strong LRD) and s being in the order of 0.1, we can select $L = 2^9$. Because Theorem provides a conservative guideline due to inequality used in the derivations and the assumption of mono-fractal model of FGN, we verify that those traces are quite lengthy for ACF estimation as well as general patterns/structures of traffic.

Table 1. Six TCP packet traces

Dataset	Date	Duration	Packets
dec-pkt-1.TCP	08Mar95	10PM-11PM	3.3 million
dec-pkt-2.TCP	09Mar95	2AM-3AM	3.9 million
dec-pkt-3.TCP	09Mar95	10AM-11AM	4.3 million
dec-pkt-4.TCP	09Mar95	2PM-3PM	5.7 million
Lbl-pkt-4.TCP	21Jan94	2AM-3AM	1.3 million
Lbl-pkt-5.TCP	28Jan94	2AM-3AM	1.3 million

4 Conclusions

We have derived a formula representing the accuracy of the correlation estimation of FGN as a 2-D function of the record length and the Hurst parameter. It may be conservative for real traffic but it may yet serve as a reference guideline in measurement. Based on the present formula, the noteworthy difference between measuring LRD and SRD sequences has been pointed out.

Acknowledgments

Special thanks go to Vern Paxson for his experienced help. This research is in part sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Sate Education Ministry, PRC.

References

1. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, *IEEE/ACM Trans. on Networking*, 2 (1), Feb. 1994, 1-15.
2. S. Floyd and V. Paxson, *IEEE/ACM Trans. on Networking*, 9 (4), Aug. 2001, 392-403.
3. M. Li, translated, *Measurement uncertainty: ANSI/ASME PTC 19.1-1985 (Instruments and Apparatus)*, Chinese version, Ship Mechanics Information Editorial Board, 1993.
4. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedure*, 3rd Edition, John Wiley & Sons, 2000.
5. M. Li, and et al., *IEEE ICII2001*, vol. 2, 2001, 45-49.
6. S. C. Lim and S. V. Muniandy, *Physics Letters* A, 206 (5-6), 1995, 311-317.
7. J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.
8. M. Li, W. Zhao, and et al., *Applied Mathematical Modelling*, 27 (3), 2003, 155-168.
9. M. Li, and et al., *Electronics Letters*, 36 (19), 2000, 1168-1169.