Loss of Self-Similarity Detection with Second Order Statistical Model And Multi-Level Sampling Approach

Mohd Fo'ad Rohani [1], Mohd Aizaini Maarof [2], Ali Selamat [3] and Houssain Kettani [4]

[1, 2, 3] *Faculty of Computer Science and Information Systems*
Universiti Teknologi Malaysia
81310 Skudai, Johor
Tel:+607-5532377, Fax: +607-5565044, E-mail: foad@.utm.my,
Tel: +607-5532002, Fax: +607- 5565044, E-mail: aizaini@.utm.my,
Tel:+607-5532099, Fax: +607-5565044, E-mail: aselamat@.utm.my

[4] *Electrical & Computer Engineering and Computer Science Department*
Polytechnic University of Puerto Rico
P. O. Box 192017
San Juan, PR 00919, USA
Tel:+787-6228000 ext. 340, 472 , Fax:+787-2818342 , E-mail: hkettani@pupr.edu

Abstract

Recent studies demonstrate that malicious packets introduce distribution error and perturb the self-similarity property of network traffic. Consequently, loss of self-similarity (LoSS) is detected which indicates poor Quality of Service (QoS). Previous works on LoSS detection typically estimate the self-similarity parameter at normal fixed sampling rate such as 10ms or 100ms. However, this is not sufficient to expose the distribution error of self-similarity model effectively hence increases false alarm rate detection. This paper proposes a multi-level sampling (MLS) approach to estimate self-similarity parameter in order to increase the accuracy of LoSS detection performance. The proposed LoSS detection method defines LoSS with second order self-similarity statistical (SOSS) model and estimates the self-similarity parameter using the Optimization Method (OM). The method has been tested using simulation of fractional Gaussian noise (FGN) traces and FSKSMNet datasets. The simulation results show that the MLS approach has significantly improved the LoSS detection accuracy from 50% to 100% for malicious traces and from none to 17% for legal Internet traffic traces, when compared to typical fixed sampling rate at 100ms.

Keywords

Loss of Self-Similarity (LoSS), Second Order Self-Similar Statistical (SOSS) Model, Multi-Level Sampling Hurst Estimation

1. Introduction

The concept of self-similarity and the related concept of long-range dependence (LRD) in local area network (LAN) traffic were initially introduced in [9] and brought the concepts into the field of network traffic and performance analysis. [9]The findings had challenged the validity of the Poisson assumption and shifted the community's focus from assuming memoryless and smooth behavior network traffic to assuming LRD and bursty behavior. Several causes of the self-similarity phenomenon had been pointed out such as the mixed behavior of TCP services model [11], the mixture of actions from individual users, hardware and software in networks [3] and the heavy-tailed distribution of file sizes transferred [3]. The work in [4] showed that congestion due to uncontrolled self-similarity structure degrades Quality of Service (QoS) performance by drastically increasing queuing delay and packet loss. Malicious packets such as Denial of Service (DoS) attacks can dominate the traffic protocol and produce distribution error, and hence disturb the self-similarity property [12]. As a result, Loss of Self-Similarity (LoSS) behavior is detected [12] and as shown in [1], [13] this can be used as a flag to alert security analysts of the possible presence of malicious actions, provided that the normal traffic background is self-similar, which is a common network traffic attribute.

The work in [1] had presented a new technique for detecting the possible presence of new DoS attacks without a template of the background traffic. The method used LoSS definition with the self-similarity or Hurst parameter H beyond normal long-range dependence self-similarity behavior ($0.5<H<1$) using the Periodogram and the Whittle methods. A new method of estimating Hurst parameter which is more accurate and faster, known as the Optimization Method (OM), was developed in [6], [7] and the method is based on second order self-similarity statistical (SOSS) model. The drawback of previous LoSS detection methods is limited to Hurst estimation at fixed sampling time scale as

applied in [1] and [13]. The work done by [5] has shown that different behavior dependence structure could exist at different time scales. The results show that loose dependence structure (or complex scaling) could exist at smaller time scale while strong dependence structure (or mono-fractal) could exist at higher time scale and the change point is usually associated with round trip time (RTT). This is a good indicator to investigate LoSS detection with multi-level sampling (MLS) to uncover the hidden property of loose and strong dependence structure of the self-similarity property.

This paper presents LoSS detection method with a multi-level sampling approach to expose the distribution error of self-similarity model effectively. The propose method defines LoSS detection using SOSS statistical model and OM. The sequel of this paper is as follows: Section 2 presents mathematical definitions and properties of SOSS and how to estimate its parameter. Section 3 on the other hand, discusses the concept of LoSS detection and related work. Section 4 discusses the datasets that were used in the simulations while Section 5 presents our experiment procedure and the results. Finally our conclusions and future work directions are summarized in Section 6.

2. SOSS Statistical Model and OM

Let $X = \{X(t), t = 0,1,2,....,N\}$ be a second-order stationary process with constant mean μ, finite variance σ^2, and autocorrelation function $\rho(k)$ that depends only on the integer *k*. Their definitions are given as follows:

$$\mu = E[X(t)], \quad \sigma^2 = E[(X(t)-\mu)]^2$$

$$\rho(k) = E[(X(t)-\mu)(X(t+k)-\mu)]/\sigma^2$$

Let $X^{(m)} = \{X^{(m)}(t), t > 0\}$ denote the aggregate process of X at $m, m = 1,2,3,....,N)$. That is, for each m, $X^{(m)}$ is given by:

$$X^{(m)}(t) = \frac{1}{m}\sum_{l=m(t-1)+1}^{mt} X(l)\ , t > 0\,.$$

Let $\gamma^{(m)}(k)$ and $\rho^{(m)}(k)$ denote the variance and autocorrelation function of $X^{(m)}$ respectively. X is called *exactly second-order self-similar (ESOSS)* if

$$\rho(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}]\ , \text{and}\,.$$

$$0 < \beta < 1,\ k = 1,2,3,....,N$$

X is called *long-range dependent (LRD)* if its autocorrelation function satisfies $\rho(k) \square ck^{-\beta}, k \to \infty$, where c is a positive constant, $0<\beta<1$ and $H = 1 - \frac{\beta}{2}$.

X is called *asymptotical second-order self-similar (ASOSS)* if

$$\lim_{m\to\infty} \rho^{(m)}(k) = \rho(k),\ k \geq 1$$

It can be shown that ESOSS implies $\rho(k) = \rho^{(m)}(k)$ for all m≥1. Thus, *second order self-similarity* captures the property of correlation structure preserving under time aggregation and is represented by:

$\rho(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$ for ESOSS or

$\lim_{m\to\infty} \rho^{m}(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$ for ASOSS.

In second-order stationary for $0<H<1$, $H\neq 0.5$, autocorrelation function satisfies $\rho(k) = H(2H-1)k^{2H-2}, k \to \infty$. In particular, if $0.5<H<1$, $\rho(k)$ asymptotically behaves as $ck^{-\beta}$, $\rho(k) = c_r k^{-\beta}$ for $0<\beta<1$ where $c_r > 0$ is a constant, and $\beta=2-2H$.

There are several methods to estimate H. In this paper we will be using the Optimization Method (OM) which was developed in [6], [7] and was shown to be comparatively fast and accurate with respect to other methods. The method is based on how near sample autocorrelation measure fits to *ESOSS* model. The estimation method defines error fitting function $E_K(H)$ as

$$E_K(H) = \frac{1}{4K}\sum_{k=1}^{K}(\rho(k) - \rho_n(k))^2$$

where $\rho(k)$ denotes the autocorrelation function of the model with parameter H that OM would like to fit the data to, $\rho_n(k)$ is the sample autocorrelation function of the data, k is autocorrelation lag and K is the largest value of k for which $\rho_n(k)$ is to be computed to reduce edge effects. The estimation of parameter H is based on optimizing $E_K(H)$ with threshold value $\leq 10^{-3}$ is chosen empirically [6].

3. LoSS Detection with MLS Approach

It is proven in [9] that normal Internet traffic activities follow the ESOSS model. However, in the presence of malicious packets such as DoS attacks, the self-similarity property is disturbed and consequently LoSS is detected as shown in [1], [12] and [13]. LoSS detection in [12] used the abrupt change property of distribution ratio of higher scale to lower scale as an indicator to the presence of distribution error. Nevertheless the work was not suggesting at what level of aggregation scale to be used for revealing the distribution error significantly. Alternatively, the work in [1] defined LoSS as Hurst value beyond normal range of LRD which is $0.5\leq H\leq 0.99$ using Periodogram and Whittle method. The results show that the method can detect new DoS attack pattern without specific normal template. The results also demonstrate that the method has high detection rate with an average of 60% to 84% which depends on the intensity of the attack packets.

Recently, a new method of estimating Hurst parameter which is more accurate and faster was developed in [6] and

[7]. It is referred to as the Optimization Method (OM) and it provides a technique to identify whether the data tend toward the self-similarity model according to the curve-fitting error value calculated. The work in [13] used OM to detect anomaly traffic based on the curve-fitting error value. Nevertheless the technique only considered fixed sampling which is not efficient enough to reveal the hidden distribution error accurately. Due to the complexity of modern Internet applications, the need for different sampling strategy is necessary to estimate the self-similarity behavior accurately [2]. Therefore, a multi-level sampling (MLS) approach is required to reveal any hidden distribution error of self-similarity property efficiently.

The proposed LoSS detection method considers three parameters: estimated Hurst (H), estimated curve fitting error and sampling level *m,* to define normal and abnormal behavior of Internet traffic. To this end, let define the following statements:

A: ($H \in 0.5<H<1$)

B: *curve fitting error <Threshold (at normal m)*

C: *mean (multi-level curve fitting error) < Threshold (at multi-level m)*

Then, normal Internet behavior is defined as LoSS is not detected at both normal fixed sampling and multi-level sampling such as: $A \cap B \cap C$. On the other hand, abnormal behavior is defined as LoSS is detected at either normal fixed sampling or multi-level sampling such as: $B \cup C$. Normal sampling here refers to *m*=10ms or 100ms which is used for Hurst estimation in [6], [7], [9] and [13] while multi-level sampling consider shorter time-scales such as *m*= 10ms, 50ms, 100ms, 200ms, 500ms, 700ms and 1000ms that represent engineering factors which have stronger impact than human behaviors [2].

4. Data Preparation

The experiments use two simulation datasets. The first is synthetic data that is ESOSS and is generated at random by using fractional Gaussian noise (FGN) model developed in [8] for 0.5<H<1. The length of the trace is equivalent to 15-30 minute at normal traffic Ethernet LAN as used in [9]. The second dataset is FSKSMNet Internet traffic simulation on September 29, 2006 at Faculty of Computer Science and Information (FSKSM) local area networks (LANs). Internet Monitoring Laboratory (InMonLab) has been setup with baseline 100BaseFX Fast Ethernet as LAN FSKSM backbone and connected to main university Gigabit backbone. The network design at FSKSM is constructed with ten proxies of Virtual LANs (VLANs) for students, administrators and academic staffs. The FSKSMnet Internet traffic simulation activities are divided into normal and abnormal traffic. Normal Internet activities are defined as legal Internet activities that abide by faculty network policy. On the other hand, abnormal traffic contains at certain rate simulated injection of DoS flooding packets includes TCP Reset, TCP SYN, UDP, ICMP and IGMP flooding packets. Details of FSKSMnet simulation are shown in Table 1 with each of capturing session is about 30 minutes. Normal traces are labeled as N while abnormal traces are labeled as AB.

Table 1 Simulation of FSKSMnet Dataset on September 29, 2006

FSKSMNet-Normal			FSKSMNet-Abnormal		
Trace	Capture	Total Packet	Trace	Capture	Total Packet
N1 (Tr4)	12.15pm-12.45pm	IP=3846328: TCP(97.94%), UDP(1.91%), ICMP(0.11%), IGMP(0.01%), Others(0.03%)	AB1 (Tr1)	10.45am-11.15am	IP=7468026: TCP(85.60%), UDP(14.35%), ICMP(0.04%), IGMP(0.01%), Others(0.005%)
N2 (Tr5)	12.45pm-1.15pm	IP=3502111: TCP(97.31%), UDP(2.50%), ICMP(0.16%), IGMP(0.01%), Others(0.03%)	AB2 (Tr2)	11.16am-11.46am	IP=9429765: TCP(58.41%), UDP(1.38%), ICMP(16.33%), IGMP(23.85%), Others(0.0043%)
N3 (Tr6)	1.15pm-1.45pm	IP=3715632: TCP(97.40%), UDP(2.45%), ICMP(0.11%), IGMP(0.01%), Others(0.033%)	AB3 (Tr3)	11.46am-12.16pm	IP=9707011: TCP(69.17%), UDP(30.77%), ICMP(0.04%), IGMP((0.005%), Others(0.02%)
N4 (Tr7)	1.45pm-2.15pm	IP=4197509: TCP(97.87%), UDP(1.69%), ICMP(0.12%), IGMP(0.01%), Others(0.31%)	AB4 (Tr10)	3.15pm - 3.45pm	IP=10081214: TCP(48.76%), UDP(47.09%)) , ICMP(0.05%), IGMP(0.004%), Others(4.09%)
N5 (Tr8)	2.15pm-2.45pm	IP=7371721: TCP(92.17%), UDP(0.93%), ICMP(0.07%), IGMP(0.004%), Others(6.83%)	AB5 (Tr11)	3.45pm-4.15pm	IP=8932254: TCP(93.73%), UDP(1.20%), ICMP(0.04%), IGMP(0.003%), Others(5.021%)

5. Empirical Analyses

5.1 LoSS Detections with MLS

The aim of the experiments is to investigate the need of multi-level sampling approach in order to improve the accuracy of LoSS detection efficiently. Figure 1 illustrates the results of Hurst parameter estimation and curve fitting error for the FGN and FSKSMnet traces at different levels of *m*.

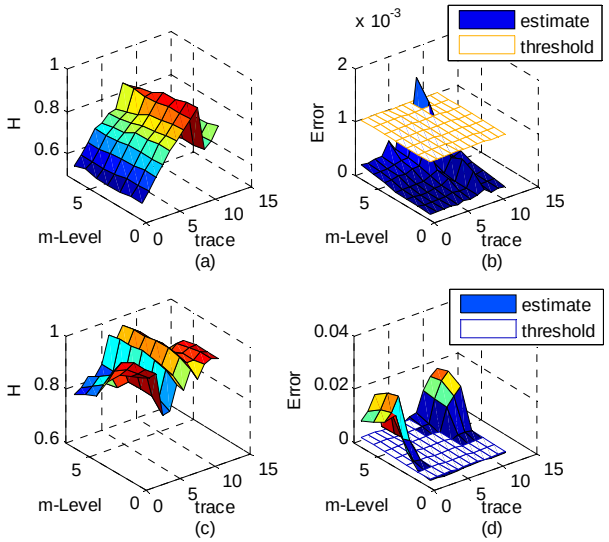

Figure 1 FGN- (a) Hurst Estimation (b) Curve fitting Error, FSKSMNet- (c) Hurst Estimation (d) Curve fitting Error

The results show that all the traces exhibit LRD with 0.5<H<1 as shown in Figure 1(a) and (c). It is obvious from Figure 1(b) and Table 2 that more than 91% of the FGN traces preserved the self-similarity property at all levels of *m*, but for *m*≥700 a small portion of LoSS is detected which is less than 8.5%. On the other hand, Figure 1(d) illustrates that for the FSKSMnet traces, almost more than 80% of the traces have LoSS occurrence and Table 2 shows that the percentage of traces with the LoSS occurrence differs at different level of *m*. The results also show that less than 10% of the traces have LoSS detection at *m*≥50ms and almost 25% of the traces have LoSS detection at *m*≥200ms. The detail results are shown in Table 2.

Table 3 shows the LoSS detection at each sampling level *m* that represents normal and abnormal behavior of the traces. Result in Table 3 shows that for the synthetic traces almost 90% follow ESSOS and less than 10% of LoSS is detected at *m*=700ms and 1000ms. This is a good indication that the probability of FGN generator to generate ESOSS model is very high. For normal FSKSMnet traces, no LoSS occurrence is detected at *m*<500ms. But LoSS is appearing at higher level of *m*>500ms. This demonstrates that normal Internet traffic with legal activities can contribute to the distribution autocorrelation error and the LoSS is detected at higher levels of *m*.

Table 2 LoSS occurrences for simulated traces

m-level(ms)	n(LoSS)	FSKSMNet(%)	FGN(%)
All	L0(no LoSS)	16.67	91.70
1000 (1)	L1(1)	8.33	0.00
700 (2)	L2(1,2)	16.67	8.30
500 (3)	L3(1,2,3)	8.33	0.00
200 (4)	L4(1,2,3,4)	25.00	0.00
100 (5)	L5(1,2,3,4,5)	16.67	0.00
50 (6)	L6(1,2,3,4,5,6)	8.33	0.00

e.g. L0=at all level m no LoSS is detected, L1- at m=1000ms LoSS is detected.

Table 3 LoSS detection at each sampling level m

m-level	FGN	FSKSMnet	
	Synthetic	AB	N
	LoSS(%) /*m*	LoSS(%)/*m*	LoSS(%)/*m*
10	0.00	0.00	0.00
50	0.00	16.67	0.00
100	0.00	50.00	0.00
200	0.00	100.00	0.00
500	0.00	100.00	16.67
700	8.33	100.00	50.00
1000	8.33	100.00	66.67

The result in Table 3 also shows that at *m*=100ms, less than 50% of the abnormal traces are detected which indicates poor LoSS detection. This demonstrates that at normal sampling rate the distribution of autocorrelation error is hidden for certain malicious traffic. However, for m>100ms LoSS is 100% detected for the abnormal traces. The inconsistent of detecting LoSS behavior at different levels of *m* as shown in Table 2, creates a big challenge to choose an optimum level of *m* that LoSS is accurately detected.

5.2 LoSS Detections Performance with MLS

A new approach of LoSS detection with MLS is proposed to eliminate the ambiguity of LoSS behavior at different level of *m*. This can be achieved by using the average value of multi-level curve fitting error. Figure 2 demonstrates that with MLS approach, all FGN traces are detected as following the ESOSS model, all malicious traces are detected as abnormal traces and almost 85% of normal traces are considered as following the ESOSS model. The results in Table 3 demonstrate that LoSS occurrence is unable to be detected at *m*=10ms. However, by comparing normal fixed sampling at *m*=100ms with MLS, the LoSS detection with

MLS has improved the detection accuracy from 50% to 100% for abnormal traces, from none to almost 17% for normal traces and none is detected for FGN traces.

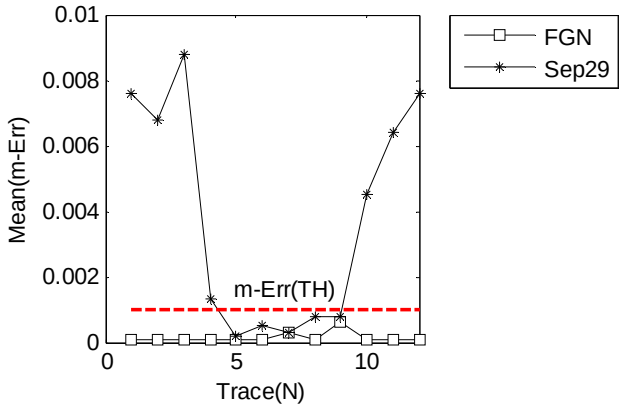

Figure 2 Average estimation of multi-level curve fitting error (average (m-Error)).

Table 3 Performance of LoSS Detection with multi-level sampling (ML) approach

Traffic	Class	Sampling		
		10ms	100ms	MLS
FGN	Synthetic	0%	0%	0%
FSKSMnet	Normal	0%	0%	16.67%
FSKSMnet	Abnormal	0%	50%	100%

6. Conclusion and Future Work

This paper presents LoSS detection using SOSS model and proposed a multi-level sampling approach to improve LoSS detection accuracy. The experimental simulation results have demonstrated that the proposed LoSS detection method has improved the accuracy of LoSS detection significantly when compared to fixed sampling method. It is clearly shown in the experiments that at sampling rate 10ms it is very hard to detect LoSS which is indicated by zero detection. On the other hand, LoSS detection accuracy for abnormal traffic at sampling rate 100ms is increased up to 50%. However, with multi-level sampling approach the accuracy of LoSS detection for abnormal traffic is 100% detected. Furthermore, the proposed method is also capable to detect LoSS for legal Internet traffic activities from none to 17% and verifies that synthetic FGN traces is always following ESOSS model. Future work will consider more real Internet traffic datasets to test the reliability and robustness of the proposed method

Acknowledgements

This work was funded by Universiti Teknologi Malaysia (UTM). The authors are grateful to Dr. Sulaiman Mohd Noor at CICT, UTM and Mr. Firoz at Unit IT, FSKSM for their helps in conducting the simulation of real traffic FSKSMNet dataset.

References

[1] Allen, W. H. and Marin, G.A. 26-29 March 2004. The LoSS technique for detecting new Denial of Service attacks. SoutheastCon, 2004. Proceedings. IEEE, pp. 302-309.

[2] Cairano-Gilfedder, C. and Clegg, R.G. Oct. 2005. A decade of Internet research -- advances in models and practices. *BT Technology Journal 23*, vol.4, pp. 115-128.

[3] Crovella, M.E. and Bestavros, A. December 1997. Self-similarity in World Wide Web traffic: Evidence and possible causes networking. *IEEE/ACM Transactions on Volume 5*, Issue 6, pp. 835 – 846.

[4] Erramilli, A., Narayan, O. and Willinger, W. 1996. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Networking*, 4:209–223.

[5] Feldmann, A., Gilbert, A.C., Willinger, W. and Kurtz, T.G. April 1998. The changing nature of network traffic: scaling phenomena. *ACM Computer Communication* Vol.28(2), pp. 5-29.

[6] Kettani, H. 2002. A Novel Approach to the Estimation of the Long-Range Dependence Parameter. University of Wisconsin – Madison : PhD. Thesis.

[7] Kettani, H. and Gubner, J. A. June 2006. A Novel Approach to the Estimation of the Long-Range Dependence Parameter. *IEEE Transactions on Circuits and Systems II*, Volume 53, Issue 6, pp. 463-467.

[8] Ledesma, S. and Liu, D. April 2000. Fractional Gaussian noise power spectrum synthesis using linear approximation for generating self-similar network traffic. *ACM Computer Communication Review*, vol.30, no.2, pp. 4-17.

[9] Leland, W., Taqqu, M., Willinger, W. and Wilson, D. 1993. On the self-similar nature of Ethernet traffic. Proc. of ACM SIGCOMM 23(4), pp. 183–193.

[10] Park, K., Kim, G. and Crovella, M. November 1997. On the effect of traffic self-similarity on network performance. SPIE International Conference on Performance and Control of Network Systems.

[11] Paxson, V. and Floyd, S. June 1995. Wide-area traffic: The failure of Poisson modeling. *IEEE-ACM Transactions on Networking*, 3(3).

[12] Schleifer, W. and Mannle, M. Feb. 2001. Online error detection through observation of traffic self-similarity. IEE Proceedings on Communications, 148(1).

[13] Yazid, M., Hanan, A. and Aizaini, M. 2004. Iterative window size estimation on self-similarity measurement for network traffic anomaly detection. *International Journal of Computing and Information Science, (IJCIS)*, vol. 2(2), pp. 83-91.