

Comparison of Two Sampling-Based Data Collection Mechanisms for Intrusion Detection System

Kuo Zhao, Liang Hu, Guannan Gong, Meng Zhang, Kexin Yang

Department of Computer Science and Technology

Jilin University

Changchun City, Jilin Province, P.R.China

Abstract - Data collection mechanism is a crucial factor for the performance of intrusion detection system (IDS). Simple random sampling and Stratified random sampling techniques of statistics are introduced to the procedure of data collection for IDS, and formulas used to calculate the sample size of packets based on these sampling techniques are presented. The implementation of packets sampling is provided, and efficiencies of these data collection mechanisms for IDS are compared in this paper. Experimental results show these two mechanisms both can improve the efficiency of data collection and strengthen the processing performance of IDS, while stratified random sampling technique performs better especially in the large-scale high-speed network.

Keywords: intrusion detection system, sampling, data collection, performance.

1 Introduction

As we increasingly rely on information infrastructures to support critical operations in defense, banking, telecommunication, transportation, electric power, and many other systems, intrusions into information systems have become a significant threat to our society with potentially severe consequences [1], [2]. It is very difficult to defend all sorts of network attacks with traditional security strategies or mechanisms (such as firewall). Even the best security countermeasure will be inevitably invaded, and intrusion detection system (IDS) has been an indispensable part of computer security [3]. Intrusion detection techniques mainly consist of misuse detection and anomaly detection, and there are many well-known detection methods such as state transition analysis, expert system, neural network and data mining. Both detection techniques have inherent limitations. Recently, there appears a mixed intrusion detection technique that attempts to integrate the advantages of preceding techniques. At

present, researches on IDS mainly focus on attack (intrusion) detection, while data collection mechanism is the crucial factor affected on the performance of intrusion detection system. It is necessary for IDS to collect data with an effective and reliable manner [4]. While it is impossible to collect all packets especially in the large-scale high-speed network (such as 1000M Ethernet), so a more reasonable network traffic collection mechanism should be applied to intrusion detection system.

Sampling is an effective technique to monitor network traffic in real time, and has been applied to various network engineering and management applications [5], such as traffic report [6], traffic characterization [7] and attack (intrusion) detection or analyses [8], [9]. At present, all the IDSs attempt to collect the whole network traffic, whereas it is inevitable for IDS to drop out some packets due to limitation of computer resources, such as capture, memory and analyses. So IDS can't execute complete real-time data collection. In addition, the proportion of network traffic with attack (intrusion) signature is commonly small, and capturing the whole traffic will degrade the efficiency of network bandwidth utilization.

In this paper, Simple random sampling and Stratified random sampling techniques of statistics are introduced to the procedure of data collection of IDS, and formulas used to calculate the sample size of packets based on these two sampling techniques are presented. The implementation of packets sampling is provided. Experimental results show these two mechanisms can both improve the efficiency of data collection and strengthen the processing performance of IDS, while stratified random sampling technique performs better especially in the large-scale network with heavy traffic.

2 Data Collection Mechanisms Based on Sampling Technique

It is a very important matter for sampling technique to calculate the sample size. Given the same condition, sampling error will decline with the increasing sampling

Supported by the Outstanding Youth Foundation of Jilin Province of China under Grant No.20040119.

size. There is an inverse proportion relation between sampling error and the square root of sample size [10].

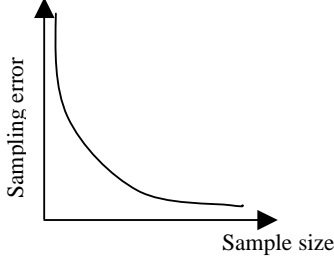


Fig. 1. Relation between sampling error and sample size

Figure 1 illustrates that sampling error tends to stable after given phases. It is obviously unworthy to reduce sampling error by increasing the sample size. At this point, we can greatly reduce the sample size with slightly degradation of precision.

In this paper, it is the key matter for IDSs to effectively reduce data collection amount as possible on condition that they can't capture the whole network traffic. The following parts present formulas used to calculate the sample size of packets based on simple random sampling technique and stratified random sampling technique.

2.1 Symbol Token

In this paper, all the packets in the network are treated as population, and individual packet as population unit.

Assume the number of population unit is N . The total number of packets with attack signature in the population is A . Let $Y_i=1$ if the packet (Y_i) with attack signature, otherwise $Y_i=0$ ($i=1,2,\dots,N$), then $P = \frac{A}{N}$ is the population proportion.

Assume the sample size is n . The total number of packets with attack signature in the sample is a . Let $y_i=1$ if the packet (y_i) with attack signature, otherwise $y_i=0$ ($i=1,2,\dots,n$), then $p = \frac{a}{n}$ is the sample proportion.

2.2 Simple Random Sampling

Simple random sampling is the basic sampling technique where we select a group of subjects (a sample) for study from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample [10].

For simple random sampling, the variance of p is

$$V(p) = \frac{P * Q}{n} * \frac{N-n}{N-1} \quad (1)$$

where $Q = 1 - P = \frac{N-A}{N}$, and the absolute error

$$d = t * \sqrt{V(p)} = t * \sqrt{\frac{P * Q}{n} * \frac{N-n}{N-1}} \quad (2)$$

or the relative error

$$r = t * \frac{\sqrt{V(p)}}{P} = \frac{t}{P} * \sqrt{\frac{P * Q}{n} * \frac{N-n}{N-1}} \quad (3)$$

so

$$n = \frac{t^2 * \frac{P * Q}{d^2}}{1 + \frac{1}{N} * (\frac{t^2 * P * Q}{d^2} - 1)} \quad (4)$$

or

$$n = \frac{t^2 * \frac{Q}{r^2 * P}}{1 + \frac{1}{N} * (\frac{t^2 * Q}{r^2 * P} - 1)} \quad (5)$$

we can first calculate

$$n_0 = \frac{t^2 * P * Q}{d^2} \text{ or } n_0 = \frac{t^2 * Q}{r^2 * P} \quad (6)$$

if $\frac{n_0}{N} < 0.05$, then n is approximately equal to n_0 ,

otherwise n_0 needs to be modified as

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \quad (7)$$

2.3 Stratified Random Sampling

Before sampling, the whole population is first divided into mutually exclusive subgroups, called stratum. Let N be the number of population unit, L be the number of strata and N_1, N_2, \dots, N_L represent the size of each stratum, then $N = \sum_{h=1}^L N_h$. If the sample is taken randomly from each stratum, the procedure is known as stratified random sampling [10].

Assume N_h is the size of h th stratum ($h=1,2,\dots,L$), $W_h = \frac{N_h}{N}$ is the stratum weight of h th stratum, and the number of units in a sample falling in h th stratum is n_h . Let $f_h = \frac{n_h}{N_h}$ be the sample proportion of h th stratum, p_h represent the subsample proportion of h th stratum. Then the estimate of population proportion P is:

$$p_{st} = \sum_{h=1}^L W_h p_h \quad (8)$$

For the general stratified sampling, p_{st} is the unbiased estimate of P if p_h is the unbiased estimate of P_h ($h=1,2,\dots,L$), and the variance of p_{st} is:

$$V(p_{st}) = \sum_{h=1}^L W_h^2 V(p_h) \quad (9)$$

Since the sample is taken randomly from each stratum, p_h is the unbiased estimate of P_h ($h=1,2,\dots,L$), and

$$V(p_h) = \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h} \quad (Q_h = 1 - P_h) \quad (10)$$

so

$$V(p_{st}) = \sum_{h=1}^L W_h^2 V(p_h) = \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h} \quad (11)$$

For stratified sampling, it is supposed to solve the problem of sample size allocation in strata given fixed population size. There are two main allocation methods. One is optimum allocation, and the other is proportional allocation. In this paper, we tend to select the latter because the optimum allocation of one variable may be not proper to the other variables. Proportional allocation refers to allocate the size within strata based on each stratum weight. That is

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \quad \text{or} \quad f_h = \frac{n_h}{N_h} = \frac{n}{N} = f \quad (12)$$

then the variance of p_{st}

$$V(p_{st}) = \frac{1-f}{Nn} \sum_{h=1}^L \frac{N_h^2 P_h Q_h}{N_h - 1} \approx \frac{1-f}{n} \sum_{h=1}^L W_h P_h Q_h \quad (N_h - 1 \approx N_h) \quad (13)$$

If the estimation precision is presented by error bound,

then $V(p_{st}) = \left(\frac{d}{t}\right)^2$ or $V(p_{st}) = \left(\frac{rP}{t}\right)^2$, where d is absolute error, r is relative error and t is the α quantile on both sides of standard normal distribution.

so the formula used to calculate the sample size is

$$n = \frac{\sum_{h=1}^L W_h P_h Q_h}{V(p_{st}) + \frac{\sum_{h=1}^L W_h P_h Q_h}{N}} \quad (14)$$

3 Implementation of packets sampling

There are two kinds of packets sampling techniques as to large-scale high-speed network. One is integrated sampling technique such as Poisson model defined in RFC2330 [11]. Sampling algorithms produce random events. For example, fixed timers or counters are used to trigger sampling events. Packets have been chosen to

sample before they arrive computer systems, and this method needs continuous sampling events. The other is distributed sampling technique. On this occasion, sampling events are defined in advance. Packets won't be chose to sample before they arrive. Whether these packets are sampled or not depends on their contents. I.Cozzani [12] and Nick Duffield [13] make use of this sampling technique.

In this paper, we use the second sampling model. Considering the randomness of identification segment of IP header, packets sampling appears on condition that sampling mask matches certain bits string of IP header. This matching mechanism is based on bit. We utilize a bit mask with random content to compare the certain bits of IP header of every packet. The precision and reliability of this sampling method lie in the offset and length of bit mask.

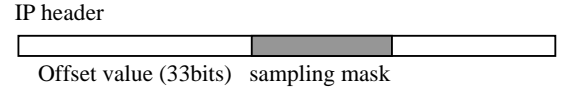


Fig. 2. Packets sampling model

Fig. 2 illustrates sampling mask model used in this paper. Identification segment of IP header presents strong randomness, and it isn't related to the characteristics of network traffic. Also, this segment won't change during network transmission. So we select partial bits of this segment to match sampling mask bits string.

4 Experiments

The whole traffic on a chief node of region network of CERNET (China Education and Research Network) in a day is used for background traffic, where some new typical attacks (intrusions) as well as other attacks in [14] are inserted, and then we have the test traffic. The IXIA1600T (a special network test facility of IXIA Crop) is used to transmit the test traffic in 1000M Ethernet. ISS RealSecure Network Gigabit is selected to detect attacks (intrusions). As the first commercial IDS, RealSecure has been playing an important role in this field, and it provides network intrusion detection and response capabilities that monitor Gigabit network.

In the experiments, the test traffic is treated as population. Considering that the procedure of sampling in each stratum of stratified random sampling technique is identical to simple random sampling, we calculate the sample size of packets with the formulas of stratified random sampling.

The stratification scheme design is based on the packet type (TCP, UDP and others). From Thompson and Park [15], [16], the proportion of TCP packets in high speed backbone network traffic is about 75%, UDP packets

is approximately 20% and other packets is around 5%. Proportional allocation method is used in this paper, then $h=3$, $W_1=0.75$, $W_2=0.20$, $W_3=0.05$. Since the amount of attack traffic is usually very small, from sampling theory, relative error r is better than absolute error d here. Let confidence is 95%, corresponding $t = 1.96$, to achieve $r = 10\%$, we can make a conservative estimate of the sample size according as $P*Q$ will be maximum value given $P=Q=0.5$. Assume $P=0.1\%$, from (14), the sample size is about $9.62E+7$. For stratified random sampling, the sample consist of $7.22E+7$ TCP packets, $1.92E+7$ UDP packets and $0.48E+7$ other packets, while the proportion of different kind of packets in the test traffic is not concerned for simple random sampling.

ROC (Receiver Operating Characteristic) technique is used to evaluate the detection effect of ISS RealSecure. The ROC approach analyzes the tradeoff between false alarm and detection rates for detection systems. It was originally developed in the field of signal detection. More recently, it has become the standard approach to evaluate intrusion detection systems [17]. In this paper, we mainly take into account the detection rates of RealSecure when its false alarm rate is under 0.2. Too many false alarms will make administrators consume unnecessary time and energy analyzing these alarms, which compromises the usability and validity.

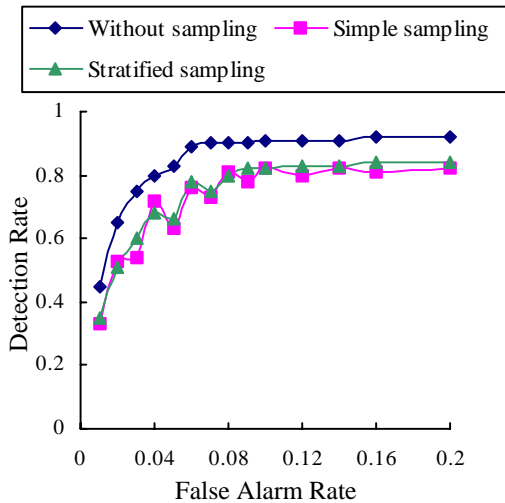


Fig. 3. The ROC curves at 1Gbps speed

It can be seen from figure 3 that the detection rates of RealSecure remarkably increase during initial phases, and then they tend to stable. The curvilinear trends of simple sampling and stratified sampling are also ascending, but the incremental extents of detection rates are relatively less. The ROC curves of sampling show definite fluctuant trends, and the fluctuant trend of simple sampling are more evident.

The possible cause is that the randomness introduced by random sampling affects detection results of RealSecure. From (14), the sample size of packets is only about 0.0183 percent of the whole test traffic, while detection rates with sampling are closed to the real detection rates by capturing the whole traffic. It is obvious that we can greatly reduce the data collection amount of IDS with slightly lower detection rates when false alarm rates tend to stable.

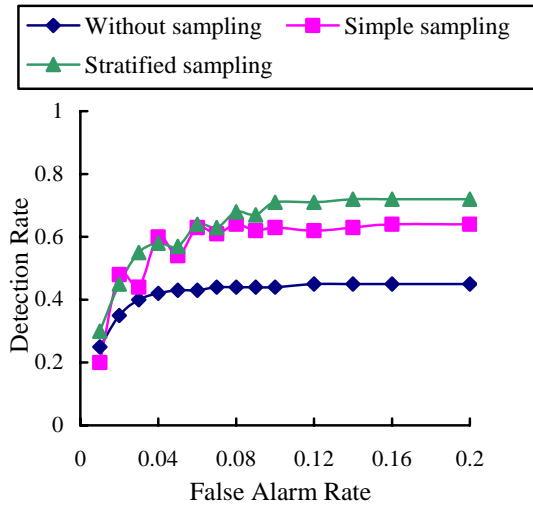


Fig. 4. The ROC curves at 10Gbps speed

To test the data collection effect of RealSecure in the large-scale high-speed network, we use IXIA1600T to transmit the test traffic at 10Gbps speed. In the circumstances, RealSecure can't capture the whole traffic and drop a large number of packets. From figure 4, we can see that the detection rates sharply decline. While using sampling methods to collect data, RealSecure can still make response to high-speed traffic. With the increases of false alarm rates, detection rates remain high and steady. Meanwhile, we can also observe that the detection rates of stratified sampling are relatively higher than simple sampling. As a matter of fact, most cyber attacks (intrusions) make use of uncommon packets (except TCP and UDP packets) such as ICMP, IGMP or other kind of packets. Though the proportion of these packets is generally small, they play an important role in detecting attacks (intrusions) for IDS. For stratified sampling, the sample contains definite proportional these packets, while few these packets may be chosen for simple sampling due to randomness. In practice, we can improve detection rates of IDS by raising the stratum weight of other types of packets.

For IDS deployed in the backbone network, they can't monitor the whole network traffic real time due to limitation of computer resources, such as capture, memory

and analyses. Especially in face of DDoS (distributed denial of service) attacks, IDS may be crashed due to overload of network traffic or computing resources. Our research will help IDS to resist against DoS attacks.

5 Conclusions

In this paper, Simple random sampling and Stratified random sampling techniques of statistics are introduced to the procedure of data collection of IDS, and formulas used to calculate the sample size of packets based on these two sampling techniques are presented. The implementation of packets sampling is provided, and efficiencies of these data collection mechanisms for IDS are compared. Experimental results show both these mechanisms can exceedingly strengthen the processing performance of IDS by the means of replacing dropping packets passively with sampling packets actively, while stratified random sampling technique performs better especially in the large-scale high-speed network.

Certainly, IDSs ought to capture the whole network traffic if they have the ability to deal with large-scale high-speed network. Our research will not only contribute to improving the efficiency of data collection for IDS but also help IDS to resist against DoS attacks.

References

- [1] Escamilla T. *Intrusion Detection: Network Security Beyond the Firewall*, John Wiley & Sons, 1998.
- [2] Jajodia S., Ammann P., McCollum C.D. "Surviving information warfare attacks," *Computer*, 32(3), 57-63, 1999.
- [3] Denning DE. "An intrusion-detection model," *IEEE Transactions on Software Engineering*, SE-13: 222-232, 1987.
- [4] Spafford E, Aamboni D. "Data collection mechanisms for intrusion detection systems," CERIAS Technical Report, Center for Education and Research in Information Assurance and Security, West Lafayette: Purdue University, IN 47909-1315, 2000.
- [5] Phaal P., Panchen S., McKee N. InMon Corporation's sFlow: "A Method for Monitoring Traffic in Switched and Routed Networks," IETF RFC 3176, 1998.
- [6] Duffield N.G., Grossglauser M. "Trajectory Sampling for Direct Traffic Observation," *IEEE/ACM Trans. on Networking*, 9(3), 280-292, 2001.
- [7] Claffy K.C., Polyzos G.C., Braun H.-W., "Application of Sampling Methodologies to Network Traffic Characterization," Proceedings of ACM SIGCOMM'93, San Francisco, CA, USA, 13-17, 1993
- [8] Murali Kodialam, Lakshman T.V., "Detecting network intrusions via sampling: A game theoretic approach," IEEE INFOCOM 2003 - The Conference on Computer Communications, 1880-1889, 2003.
- [9] Alpacan T. and Basar T. "A game theoretic approach to decision and analysis in network intrusion detection," in 42nd IEEE Conference on Decision and Control, Maui, 2003.
- [10] Cochran, W. G. *Sampling techniques*. Third Edition. New York: John Wiley and Sons, New York, USA. 1977.
- [11] Paxson V, Almes G, Mahdavi J, Mathis M. "Framework for IP performance metrics," IETF RFC 2330, 1998.
- [12] Cozzani I, Giordano S. "A passive test and measurement system: traffic sampling for QoS evaluation," In: IEEE Communications Society, ed. Proceedings of the Global Telecommunications Conference (GLOBECOM'98). Sydney: IEEE Press, 1236-1241, 1998.
- [13] Duffield N, Grossglauser M. "Trajectory sampling for direct traffic observation," In: Gunningberg P, Pink S, eds. Proceedings of the ACM SIGCOMM 2000. Stockholm: ACM SIGCOMM, 271-282, 2000.
- [14] Kendall, Kristopher. "A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems," Masters Thesis, MIT, 1999.
- [15] Thompson K., Miller G., Wilder R. "Wide area Internet traffic patterns and characteristics," *IEEE Network*, 11(6), 10-23, 1997.
- [16] Park JS, Lee JY, Lee SB, "Internet traffic measurement and analysis in a high speed network environment: Workload and flow characteristics," *Journal of Communications and Networks*, 2 (3): 287-296, 2000.
- [17] Richard P. Lippmann, David J. Fried, Isaac Graf, Joshua W. Haines, et al, "Evaluating Intrusion Detection Systems: the 1998 DARPA Off-Line Intrusion Detection Evaluation," in Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), IEEE Press, 2000.