Effect of sampling rate and monitoring granularity on anomaly detectability

Keisuke Ishibashi*, Ryoichi Kawahara†, Mori Tatsuya†, Tsuyoshi Kondoh* and Shoichiro Asano‡
* Information Sharing Platform Labs. NTT Corporation 3-9-11 Midoricho Musashino-shi Tokyo, Japan
† Service Integration Labs. NTT Corporation 3-9-11 Midoricho Musashino-shi Tokyo, Japan
‡ National Institute of Informatics 2-1-2 Hitotsubashi Chiyoda-ku Tokyo, Japan

***Abstract*—In this paper, we quantitatively evaluate how sampling decreases the detectability of anomalous traffic. We build equations to calculate the false positive ratio (FPR) and false negative ratio (FNR) for given values of the sampling rate, statistics of normal traffic, and volume of anomalies to be detected. We show that by changing the measurement granularity, we can detect anomalies even with a low sampling rate and give the equation to derive optimal granularity by using the relationship between the mean and variance of aggregated flows. With those equations, we can answer for the practical questions that arise in actual network operations; what sampling rate to set in order to find the given volume of anomaly, or, if the sampling is too high for actual operation, then what granularity is optimal to find the anomaly for a given lower limit of sampling rate.**

I. Introduction

With threats against Internet security increasing, monitoring Internet traffic and detecting anomalous traffic, such as DDoS (distributed denial of service) attacks has become a critical task in network operations.

Monitoring techniques range from counting the volume of traffic on a link by using an SNMP MIB [1] to capturing packets transferred through networks by applying mirroring at switches and/or routers or tapping the link. The former method has a drawback in that it provides only the traffic volume and no information about the source and victim of the detected anomaly. On the other hand, the latter method gives rich information but lacks scalability because capturing a large volume of packets is still difficult and installing capturing devices throughout the whole network is prohibitively expensive.

Recently, because of its easy implementation and rich information for identifying and diagnosing anomalous traffic, flow monitoring at routers has been used for these purposes [2], [3]. This method mainly provides 5-tuple of flow information through the routers. There have been many studies on detecting and diagnosing anomalous traffic using flow monitoring results [4], [5]. However, in a high-traffic-rate environment, flow monitoring increases the load on the router. To decrease this load, sampling has been introduced for flow monitoring. It is naturally expected that sampling will introduce uncertainty into the measurement results, so it is necessary to determine how the detectability of anomalous traffic will be affected by sampling.

Two recent studies reported the effect of sampling on anomaly detection [6], [7]. While they comprehensively studied various sampling methods and anomaly detection methods using actual anomalies in their data, they don't give explicit equations for the relationship between detectability and sampling rate. Thus, there was still no answer for the simple question, "to detect a 10-Kpps anomaly in normal traffic whose baseline is 200-Kpps, what is the maximum sampling rate?" This is important for network operators. In [8], Kawahara quantitatively evaluate how sampling decrease the detectability of anomaly in the number of flows. However, to the best of our knowledge, there have been no studies that provide an answer to the above question for the anomalies with large number of packets, which is required for actual network operation.

In this paper, we simply focus on packet-volume based anomaly, and evaluate the effect of sampling on anomaly detection. We then derive the relationship among the size of anomalies to be detected, the statistics of normal traffic, and the sampling rate. By using the relationship, we provide the optimal sampling rate or maximum detectable anomaly with given parameters such as normal traffic statistics. In addition, we show that by changing the measurement granularity, we can detect anomalies even with sampled traffic.

The rest of the paper is organized as follows. In section II, the effect of sampling on anomaly detection is derived theoretically and an evaluation of actual traffic data is given. Then, in section III, under the sampling, which granularity of traffic monitoring is optimal with given evaluated to detect give volume of anomalies.

II. Effect of Sampling

A. False Negative/Positive Ratio with Packet Sampling

By definition, anomalous traffic is detected by its deviation from normal traffic behavior or statistics. If sampling changes the statistics of normal and anomalous traffic flows equally, then detectability does not depend on the sampling rate. However, as shown below, while the mean rates of normal and anomalous traffic decrease linearly as the sampling rate decreases, the variance of the normal traffic does not decrease as fast for a very small sampling rate. This is because, sampling itself introduces deviations into sampled traffic. Thus, sampling increases the relative variance of normal traffic, and it is possible that the detectability of anomalous traffic is decreased. Thus, we derive the relationship for this effect.

In the following of the paper, we assume that packets are sampled using a random packet sampling method, where

packets passing through a certain router are sampled with a fixed probability independently of other packets being sampled. There is also a representative sampling method such as systematic sampling, in which packets are sampled from a packet stream at a fixed interval [9]. However, the results for random sampling is easily applicable for systematic sampling and the effect of systematic sampling is smaller than that of random sampling.

Here, we define the notation.

- P_t: Number of packets in the t-th measurement interval
- Pn_t: Number of normal packets in the t-th measurement interval
- Pa_t: Number of anomalous packets in the t-th measurement interval. ($P_t = Pn_t + Pa_t$.)
- m: Mean of Pn_t.
- σ^2: Variance of Pn_t.
- a: Mean of Pa_t. (We focus on to detect an anomaly that occurs in a measurement interval. Thus the variance of anomalous traffic in multiple measurement intervals is not considered.)
- p: Sampling rate.
- $P_t(p)$: Number of packets sampled in the t-th measurement interval with sampling probability p.
- $Pn_t(p)$: Number of packets of normal traffic in the measurement interval.
- $Pa_t(p)$: Number of packets of anomalous traffic sampled in the measurement interval with sampling probability p.
- $\sigma(p)^2$: Variance of $Pn_t(p)$.
- $m(p)$: Mean of $Pn_t(p)$.

In this paper, we focus on the volume-based anomaly and anomalous traffic is judged to occur if the traffic volume exceeds a threshold based on normal traffic statistics; that is:

$$P_t \quad > \quad m + c_{th}\sigma, \tag{1}$$

where c_{th} determines the degree of deviation from which we judge that traffic is anomalous.

With this definition, the false negative ratio (FNR) and false positive ratio (FPR) are expressed as follows (Fig. 1).

$$FNR \quad = \quad \Pr[Pn_t + Pa_t < m + c_{th}\sigma] \tag{2}$$

$$FPR \quad = \quad \Pr[Pn_t > m + c_{th}\sigma] \tag{3}$$

If the distribution function of normal traffic and the size of the anomalous traffic are given, then both FNR and FPR can be calculated using the parameter c_{th}. For example, if we can assume that the normal traffic has a Gaussian distribution[1] and that $c_{th} = 2.33$, then FPR will be 1.0% independently of the anomalous traffic. As for FNR, if we fix the anomalous traffic volume to be detected, such as $a = 6\sigma$, then FNR is also determined, such as 0.012%.

Next, we evaluate how FNR and FPR change when sampling is introduced. First, we define $FNR(p)$ and $FPR(p)$ for the sampled traffic as follows.

$$FNR(p) \quad := \quad \Pr[Pn_t(p) + Pa_t(p) < m(p) + c_{th}\sigma(p)] \tag{4}$$

$$FPR(p) \quad := \quad \Pr[Pn_t(p) > m(p) + c_{th}\sigma(p)] \tag{5}$$

[1]We give the evaluation on this assumption later.

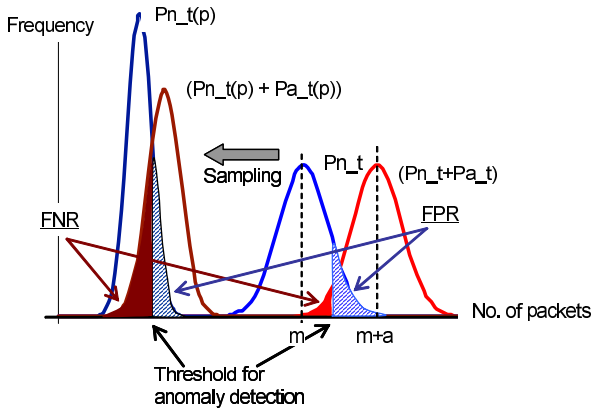

Fig. 1. Probability distribution functions of traffic for normal and anomalous periods. Sampled traffic has lower means, but its variances are not so much lower. Thus, FNR and/or FPR can increase.

By the same explanation as for Eq. (3), $FPR(p)$ is determined by only the choice of c_{th}. Thus, we focus on $FNR(p)$ hereafter and evaluate each term of Eq. (4). As for the right-hand side of Eq. (4), $m(p)$ is simply the scaled-down value of the packet rate for original traffic, mp. On the other hand, $\sigma(p)$ is not only the scaled-down variance of the original traffic, $\sigma^2 p^2$, but also includes the variance introduced by sampling. The latter term is obtained as $mp(1-p)$ as the variance of m Bernoulli trials with probability p. Thus, $\sigma(p)$ is obtained as[2]

$$\sigma(p) \quad = \quad \sqrt{\sigma^2 p^2 + mp(1-p)}. \tag{6}$$

Next, we evaluate the left side of Eq. (4). If we assume that Pn_t is distributed in a Gaussian manner, then $Pn_t \sim N(mp, \sigma^2 p^2 + mp(1-p))$. Pa_t is also the result of a Bernoulli trials and follows a binomial distribution, which can be approximated by a Gaussian distribution. Thus, with the same calculation as for Eq. (6), $Pa_t \sim N(ap, ap(1-p))$. Therefore, $Pn_t + Pa_t$ is the convolution of two Gaussian distributions, and

$$Pn_t + Pa_t \sim N((a+m)p, \sigma^2 p^2 + (a+m)p(1-p). \tag{7}$$

Therefore, $FNR(p)$ can be written using the cumulative distribution function for the Gaussian distribution of Eq. (7), $F_{a,n}(x)$ as

$$FNR(p) = F_{a,n}((a+m)p + c_{th}(\sigma^2 p^2 + (a+m)p(1-p)). \tag{8}$$

Here, for the given target FNR^*, the minimum sampling rate p^* is obtained by solving the following equation:

$$F_{a,n}((a+m)p^* + c_{th}(\sigma^2 {p^*}^2 + (a+m)p(1-p)) = FNR^*. \tag{9}$$

[2]Equation (7) is also mentioned in [7] as the variance of the number of sampled flows as an assumption. But it can be strictly derived and we show the derivation in the appendix.

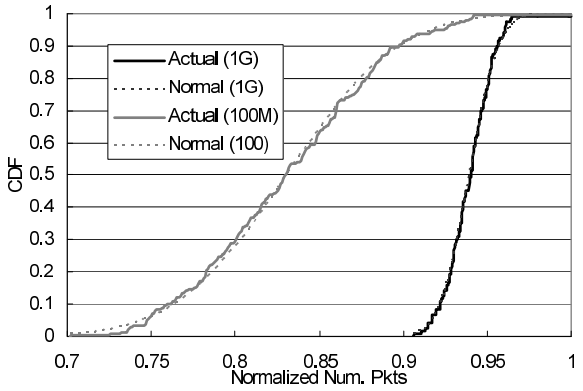

Fig. 2. Fitting the empirical distributions of normal traffic with Gaussian distributions. X axis is normalized with maximum number of packets.

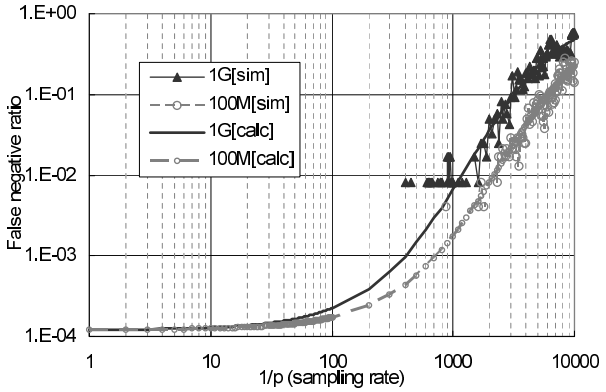

Fig. 3. Comparison of calculated and simulated false negative ratio

For example, if FNR^* is the same as FPR, which is determined by c_{th}, then by solving Eq. (9), we get

$$p^* = \frac{c_{th}^2}{a + 2m - 2\sqrt{am + m^2 + \sigma^2 c_{th}^2}} + 1. \quad (10)$$

B. Evaluation

We evaluated the results given in previous section by simulations using actual traffic data.

We used the following two sets of data:

- **1G**: Data measured on a 1-Gbps link with an average utilization ratio of 95%. The mean packet rate was 169 kpps and the standard deviation was 137,313 measured at 1 min. intervals. We used 2 hours of data during working hours, which can be regarded as being stationary.
- **100M**: Data measured on a 100-Mbps link with an average utilization ratio of 85%. The mean packet rate was 16 kpps and the standard deviation was 62,210 measured at 1 min. intervals. We used 4 hours of working-hour data, which can be regarded as being stationary. (Actually, we remove some apparent spikes to obtain stationary data, as same as in [6].)

First, we evaluate the assumption of Gaussian distribution for normal traffic. Fig. 2 shows the results of fitting the empirical distributions ([Actual]) of two normal traffic with Gaussian distributions ([Normal]). It can be seen that Gaussian assumption can be applied for those high-volume traffic as reported as [10]. It is worthful to mention that the discussion given in this paper can be easily extended to heavy-tailed distributions while some equations are not explicitly solved but numerical calculation is required.

We then evaluated FNRs for various sampling rates. To verify the detectability Eq. (4), we inserted the synthesized anomalous traffic whose size was 6σ into the observed traffic and checked whether the total traffic exceeded the threshold: $m(p)+c_{th}\sigma(p)$, where $c_{th} = 2.33$, so that FPR will be to 1 % (Hereafter, we fix $c_{th} = 2.33$ unless specifically mentioned otherwise). By changing the bin into which the anomalous traffic was inserted, we counted the number of misdetections and calculated the ratio, FNR.

The simulated ([sim] in the figure) and calculated ([calc] in the figure) FNRs are compared in Fig. 3. Because the data was 2- and 4-hour data and the total numbers of bins were 120 and 240, respectively, we could not accurately evaluate the FNRs below about 1 %. However, for the range above 1%, the FNRs calculated using Eq. (4) accurately predict those obtained through simulation. We also observed that when the sampling rates decreased to $1/100$, FNR started to degrade and when the rates were under $1/1000$ for 1G data and $1/2000$ for 100M data, FNR exceeded 1%. Thus even if anomaly whose size is 6σ, it can be difficult to detect it with small sampling rate, while it can be easily detected by eyeballing the time series data when sampling is not applied,

The reason that the FNR of 1G data was more sensitive to the sampling rate can be explained using Eq. (7) as follows. Detectability mainly depends on the ratio of the standard deviation of the normal traffic to the size of the anomalous traffic. The standard deviation of sampled normal traffic consists of two terms as shown in Eq. (7). As the sampling rate decreased, the second term on the right side of Eq. (7), which is the mean rate of normal traffic, increased relative to the first term. Thus, the smaller the ratio of the traffic's variance to its mean, which is known as the index of dispersion for count (IDC) [11], the greater the sensitivity of FNR to low sampling rate. Because 1G data has smaller IDC than that of 100M data, FNR can be relatively degraded with larger sampling rate.

We also ran the simulation by changing the anomaly size for a fixed sampling rate, $p = 1/1000$, and found the relationship between the anomaly size and FNR (Fig. 4). It can be seen that to achieve an FNR of under 1% with a sampling rate of 1000, the anomaly size should be larger than 5 or 6 times σ, the standard deviation of normal traffic.

Then, we evaluated the relationship between anomaly size and sampling rate to achieve FNR=1%. The minimum anomaly that can be detected with FNR=1% for a given sampling rate is shown in Fig. 5. Like Fig.4, the anomaly size was normalized by the standard deviation of normal traffic.

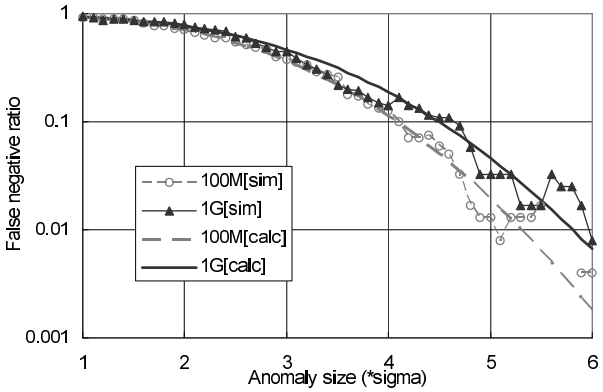

Fig. 4. False negative ratio vs. anomaly size for sampling rate $1/1000$

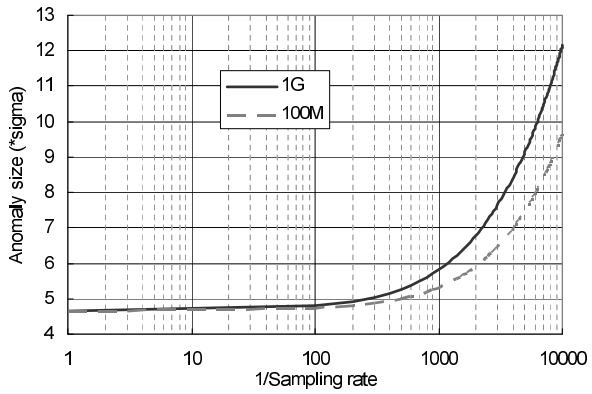

Fig. 5. Anomaly size that can be detected with given sampling rate.

With a sampling rate of 1, the minimum anomaly size was $4.66(= 2 \times 2.33)\sigma$. The same as for the result shown in Fig. 3, when the sampling rate decreased under $1/100$, the minimum anomaly size started to increase, and when the sampling rate was $1/10000$, the minimum anomaly sizes were 10σ for 100M data and 12σ for 1G data.

III. Monitoring granularity

A. Mean and variance of classified traffic and its detectability

In the previous section, we considered only a single time series of traffic data (packet volume) and detected an anomaly as a deviation of the time series. In many types of anomalous traffic, the values of some packet header fields are concentrated in specific ranges. For example, in DDoS attacks, the destination IP addresses are, and in scans, the source IP addresses are. Thus, by counting the number of packets with each particular field value and making a multiple time series, we should be able to detect anomalies that we failed to find when we treated the entire traffic deviation. However, counting traffic for each source/destination IP address requires a large number of counters and is impossible in backbone monitoring. Thus, it usual to aggregate some ranges of the field values and measure traffic for each aggregated value (e.g., IP address prefixes). Here, there is a trade-off between detectability and the aggregation level or the monitoring granularity. In this section, we evaluate the optimal monitoring granularity in terms of the mean rate of counters, with given anomaly size to be detected and sampling rate. Through the evaluation, we derive some rules, such as "to detect an anomaly of 10 kpps with a sampling rate $1/10000$, we must measure the traffic with granularity of x pps", which is useful for network operators.

In determining the granularity in terms of mean packet rate, we need to have the relationship between the mean and standard deviation of normal traffic because the detectability equation, Eq. (4), is based on both the mean and standard deviation. Here, we assume that the standard deviation of a counter σ_m can be written with mean rate m as

$$\sigma_m = cm^{\phi}, \tag{11}$$

where where $0 \leq \phi \leq 1$ and c is a constant.

As for ϕ, If the time series of non-aggregated traffic (e.g., time-series for each destination IP address) are mutually independent, then the variance of the aggregated traffic is the sum of the individual traffic variances and $\phi = 0.5$. On the other hand, if time series are highly correlated, then phi can be nearly one. There have been some studies of the relationship for actual Internet traffic. Gunnar et al. analyzed the relationship shown in Eq. (11) for each cell of the real traffic matrix and found that $\phi = 0.7 \sim 0.8$ [12]. When various aggregate levels for web traffic were measured, $\phi = 0.5$ was found to fit the actual data [13]. To evaluate the relationship with the two datasets that we used, we divided the traffic according to its destination IP address with various prefix lengths and calculated its mean and standard deviation. A scatter plot of the mean and standard deviation of each address prefix is shown in Fig. 6. We also plotted the standard deviations that follow the relationship in Eq. 11, where c was determined so that when m was equal to the total traffic, σ_m coincided with the standard deviation of the total traffic. Most of the points fall into the region between $\phi = 0.5$ and $\phi = 1.0$. Here, because as the standard deviation is large, detectability is low, so we used $\phi = 0.5$ for an evaluation that is mostly on the safe side.

We first evaluated how the spatial granularity affects the detectability for a given anomaly size and sampling ratio. We compared the calculated FNR obtained by substituting σ_m in Eq. (11) into Eq.(4) with the simulated FNRs for each divided traffic time series using the same method as in the previous section. The results are shown in Fig. 7. While there are some discrepancies between the simulation results and the calculated results for 1G data, which are expected because of the negatively correlated traffic, most of the plots coincide with each other.

Next, we investigated which granularity is optimal in the sense the granularity is maximum (number of time series is minimum) while the $FNR(p)$ is above the threshold for a given sampling rate and anomaly size.

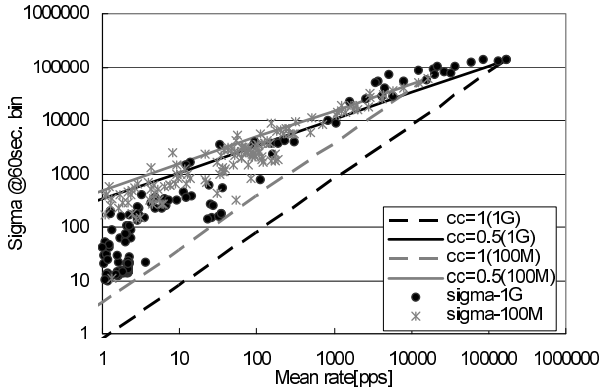

Fig. 6. Scatter plot of mean and sigma of number of packets

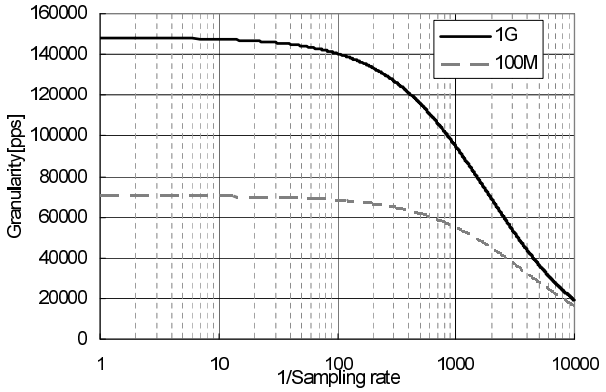

Fig. 8. Maximum granularity to detect 10 Kpps anomaly for each sampling rate

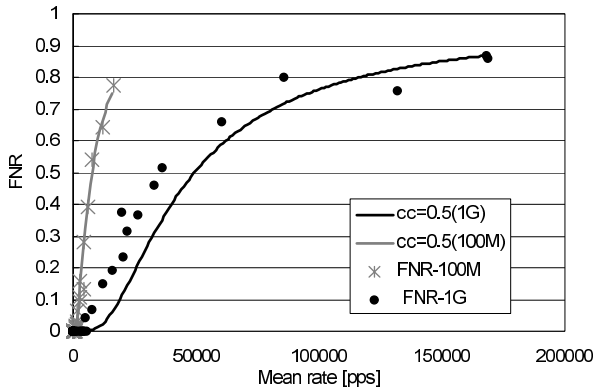

Fig. 7. Comparison of calculated and simulated FNR for each partitioned traffic

Prefix length	Granularity (max/avg/min) [Kpps]	FNR (max/avg/min)[%]
28bit (hash)	21.3/10.5/7.5	2.5/1.3/0.0
28bit	86.3/10.6/0.00	60.8/5.0/0.0
29bit (hash)	30.9/21.1/17.2	3.3/1.5/0.0
29bit	106.1/21.2/0.27	58.3/12.4/0.0

TABLE I
FNRS WITH PARTITIONED TRAFFIC WITH HASHED IP ADDRESSES

Here, for the given target FNR^*, the maximum monitoring granularity m^* is obtained as the solution for the following equation in terms of m, in a similar manner to finding the optimal sampling rate p^*.

$$F_{a,n}((a+m)p+c_{th}((cm^{\phi}))^2p^2+(a+m)p(1-p))=FNR^* \quad (12)$$

Here, by setting FNR^* to be the same as FPR, we get the explicit solution as

$$m^*=\frac{(a/c_{th})^2+2a+c_{th}^2-2a/p-2c^2/p+c^2/p^2}{4(1/p-1+c^2)}. \quad (13)$$

The optimal granularity in terms of the mean packet rate for various sampling rates is shown in Fig. 8. We set the target FNR^* as 1%, anomaly size as 10 kpps, $\phi = 0.5$, and two values of c corresponding to 1G and 100M data. It can be seen that the monitored granularity of the two parameters differed for a high sampling rate because the major factor in the sampling rate is the standard deviation of the original traffic, that is, the first term of Eq. (7). On the other hand, when the sampling rate was very low, the second term become dominant and the two optimal granularity approached the same value, about 20 kpps (19.3 Kpps for 1G data and 16.5 Kpps for 100M data) with sampling rate $1/10000$.

B. Classifying packets to achieve target granularity

In this subsection, we discuss how to realize the granularity calculated with Eqn.(13), e.g. how to classify packet so that counters for those classified packets will be the given granularity. Here, we describe two classifications, based on IP address and based on DDoS type.

As for the former classification, we must consider the effect of high spatial locality in the Internet traffic [14]. That is the concentration of large number of packets to a small number of IP address space. Thus, only by classifying the aggregated IP address range, there will be large bias among the counters as shown in Fig. 7. In that case, increasing the number of counters may not improve the detectability. By hashing the IP addresses, it is expected that the effect of locality can be avoided and the number of packets among the counters will be balanced.

For example, to achieve the granularity of 19.3 Kpps for 1G data to detect 10 Kpps anomaly with $1/10000$ sampling as shown in the previous subsection, we should divide into about nine counters of whole 170 Kpps traffic. That is achieved by using first 28 or 29 bits of hashed IP addresses. Table I shows the results. It can be seen that by using hashed IP address, the objective FNR is almost achieved, while with non-hashed IP addresses, FNR can be 60% even with classified traffic.

Next, we evaluate the classification based on DDoS type. We classified ICMP, UDP, and TCP SYN packets, each of which is used for typical DDoS attacks [15]. We found that it consists of 12%, 8%, and 2% of total packets, respectively for both 1G and 100M data, except that in 1G data, only 0.3% of

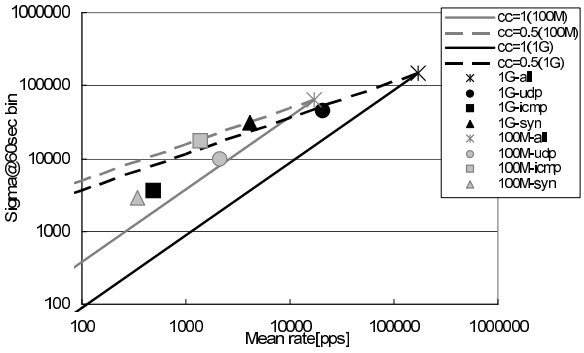

Fig. 9. Scatter plot of mean and sigma of number of packets for classified packets

icmp packets, where filtering is suspected. As shown in Fig. 9, with this classification, standard deviations almost fall into the region of $0.5 \leq \phi \leq 1$ in Eqn. (11), so we can apply results provided in the previous section. For example, the mean packet rate of UDP traffic in 1Gdata is 20.7 Kpps, which is almost the same as the maximum granularity to detect 10 Kpps anomaly with sampling rate $1/10000$ and $FNR = 1$ %. Through the simulation for UDP packets with the above condition, we can achieve $FNR = 2.5$ %, which is nearly to the target FNR

IV. Conclusion

In this paper, we showed how sampling decreases the detectability of anomalous traffic. We derived equations for FNR and FPR when the random packet sampling method is used. Using the equations, we derived the minimum sampling rate to achieve the target values of FNR and FPR. The results were evaluated through simulations using actual traffic data. Especially, we showed that when sampling rate decrease under a value such as 1/100, the FNR rapidly degrade and the degradation depends on IDC of the normal traffic. Then we found that by changing the measurement granularity, we could detect anomalies even with sampled traffic. In addition, how to realize the optimal granularity to detect anomalies with given volume to be detected. For example, by counting only UDP traffic, FNR can be improved from 48 % to 2.5 % for detecting UDP Flooding attack.

Acknowledgment

We would like to thank Kenjiro Cho, who to admit the use of MAWI data [16]. This study was supported in part by the Ministry of Internal Affairs and Communications of Japan.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proc. 2nd ACM SIGCOMM Workshop in Internet Measurement (IMW2002)*, Marseille, France, October 2002.

[2] B. Claise, "Cisco Systems NetFlow Services Export Version 9," RFC 3954, October 2004.

[3] P. Phaal, S. Panchen, and N. McKee, "InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks," RFC 3176, September 2001.

[4] P. Barford and D. Plonka, "Characteristics of Network Traffic Flow Anomalies,"in *Proc. 1st ACM SIGCOMM Workshop in Internet Measurement (IMW2001)*, San Francisco, USA, November 2001.

[5] A. Lakhina, M. Crovella and C. Diot, " Characterization of Network-Wide Anomalies in Traffic Flows," in *Proc. 4th ACM SIGCOMM Conference on Internet Measurement (IMC 2004)*, Taormina, Italy, October 2004.

[6] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina, "Impact of Packet Sampling on Anomaly Detection Metrics," in *Proc. 6th ACM SIGCOMM Conference on Internet Measurement (IMC 2006)*, Rio de Janeiro, Brazil, October 2006.

[7] J. Mai. C.-N. Chuah, A. Sridharan, T. Ye, and H. Zang, "Is Sampled Data Sufficient for Anomaly Detection?" in *Proc. 6th ACM SIGCOMM Conference on Internet Measurement (IMC 2006)*, Rio de Janeiro, Brazil, October 2006.

[8] Ryoichi Kawahara, Tatsuya Mori, Noriaki Kamiyama, Shigeaki Harada, and Shoichiro Asano, "A study on detecting network anomalies using sampled flow statistics," in *Proc. IEEE SAINT 2007 Workshop on Internet Measurement Technology and its Applications to Building Next Generation Internet*, Hiroshima, Japan, January 2007.

[9] T. Zseby, M. Molina, N. Duffield, S. Niccolini, and F. Raspall, "Sampling and filtering techniques for IP packet selection," Internet-Draft, draft-ietf-psamp-sample-tech-07.txt, July 2005.

[10] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "Internet Traffic Tends Toward Poisson and Independent as the Load Increases," in Nonlinear Estimation and Classification, Springer Verlag, Dec. 2002.

[11] R. Gustella, "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 2, pp. 203–211, February 1991.

[12] A. Gunnar, M. Johansson, and T. Telkamp, "Traffic Matrix Estimation on a Large IP Backbone: a Comparison on Real Data," in *Proc. 4th ACM SIGCOMM Conference on Internet Measurement (IMC 2004)*, Taormina, Italy, October 2004.

[13] R. Morris and D. Lin, "Variance of aggregated web traffic," in *Proc. IEEE INFOCOM 2000*, Tel Aviv, March 2000. n

[14] K. Ishibashi, M. Aida and M. Imase, "Characteristics of Temporal and Spatial Locality of Internet Access Patterns," in *Proc. IFIP Networking 2002 Workshop on Web Engineering,* Pisa, Italy, May 2002.

[15] J. Mirkovic and P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," *ACM SIGCOMM Computer Communication Review*, vol.34, no.2, April 2004.

[16] MAWI Working Group Traffic Archive, available at http://tracer.csl.sony.co.jp/mawi/.

Appendix

Let $f(n) := \Pr[P_t = n]$ and $f_p(n) := \Pr[Pn_t(p) = n]$. Then,

$$\begin{aligned}
& \sigma^2(p) \\
= & \sum n^2 f_p(n) - [m(p)]^2 \\
= & \sum_n n^2 \sum_m \Pr[Pn_t(p) = n | Pn_t = m] f(m) - (mp)^2 \\
= & \sum_n n^2 \sum_{m=n}^{\infty} {}_mC_n p^n (1-p)^{m-n} f(m) - (mp)^2 \\
= & \sum_{m=1}^{\infty} \sum_{n=1}^{m} n^2 [{}_mC_n p^n (1-p)^{m-n}] f(m) - (mp)^2 \\
= & \sum_{m=1}^{\infty} f(m) \sum_{n=1}^{m} n^2 [{}_mC_n p^n (1-p)^{m-n}] - (mp)^2 \\
= & \sum_{m=1}^{\infty} f(m)[mp(1-p) + (mp)^2] - (mp)^2 \\
= & p(1-p)m + p^2[\sigma^2 - m^2] - (mp)^2 \\
= & p(1-p)m + p^2\sigma^2 \qquad (14)
\end{aligned}$$