Published in IET Communications
Received on 22nd May 2007
Revised on 27th July 2007
doi: 10.1049/iet-com:20070231

ISSN 1751-8628

Improving network anomaly detection via selective flow-based sampling

G. Androulidakis S. Papavassiliou

Network Management and Optimal Design Laboratory (NETMODE), School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Iroon Polytexneiou 9, Zografou 15780, Athens, Greece
E-mail: papavass@mail.ntua.gr

Abstract: Sampling has become an essential component of scalable Internet traffic monitoring and anomaly detection. A new flow-based sampling technique that focuses on the selection of small flows, which are usually the source of malicious traffic, is introduced and analysed. The proposed approach provides a flexible framework for preferential flow sampling that can effectively balance the tradeoff between the volume of the processed information and the anomaly detection accuracy. The performance evaluation of the impact of selective flow-based sampling on the anomaly detection process is achieved through the adoption and application of a sequential non-parametric change-point anomaly detection method on realistic data that have been collected from a real operational university campus network. The corresponding numerical results demonstrate that the proposed approach achieves to improve anomaly detection effectiveness and at the same time reduces the number of selected flows.

1 Introduction

As the Internet continues to develop rapidly in size and complexity, it has become increasingly clear that its evolution is closely tied to a detailed understanding of network traffic. Network traffic measurements are invaluable for a wide range of tasks such as network capacity planning, traffic engineering, fault diagnosis, anomaly detection or detection of network security violations [1, 2].

Because of the large number of flows on a high-capacity link and the corresponding difficulty in storing and processing all flow informations with a limited amount of resources, sampling has attracted a great deal of attention as a way to collect statistical information about flows [3, 4]. Sampling is the process of making partial observations of a system of interest and drawing conclusions about the full behaviour of the system from these limited observations. The observation problem is concerned with minimising information loss while reducing the volume of collected data in order to make this process feasible and scalable.

Among the fundamental issues regarding sampling is its accuracy. This question is particularly pertinent in the Internet environment, where its traffic is known to fluctuate dynamically and unpredictably. Inaccurate sampling can lead to wrong decisions by network operators. Another important question closely related to the accuracy issue is the efficiency of sampling: how many packets do one need to process in order to produce reliable results.

Most of the existing work on sampling has addressed the inversion of general traffic properties such as flow size distribution, average flow size or total number of flows. Although these approaches are interesting for general-purpose network management processes and traffic characterisation, it is only recently that some limited work has been done towards understanding the impact of sampling on individual flow properties. Towards this direction, intelligent sampling is required to obtain a reliable estimate of detailed information from only a subset of flow records. It exploits the fact that for specific-purpose applications, a large fraction of information is contained in a small fraction of flows. By preferentially sampling flows, we can

control the volume of statistics simultaneously by controlling the variance of statistical estimates derived from them.

Our research work is devoted to addressing the above issues for anomaly detection purposes. Network anomaly detection techniques [5–7] rely on the analysis of network traffic and the characterisation of the dynamic statistical properties of traffic normality, in order to accurately and timely detect network anomalies. Anomaly detection is based on the concept that perturbations of normal behaviour suggest the presence of anomalies, faults, attacks and so on. The problem of anomaly detection becomes more complex and challenging when network traffic data are sampled.

The main objective of this work is to design, develop and analyse sound and efficient flow-sampling techniques that have practical application in anomaly detection. To achieve this, based on experiences regarding the sources of the majority of attacks, we identify key elements that affect the accuracy of flow sampling and use mathematical models and statistical characteristics regarding the corresponding parameters of interest. Among the main contributions of this paper is the introduction of a new flow-based sampling technique that focuses on the selection of small flows, which are usually the source of malicious traffic [8, 9]. In order to evaluate the impact of the proposed flow-based sampling technique on anomaly detection effectiveness, an algorithm that exploits a commonly used sequential non-parametric change-point detection (CPD) method is adopted and utilised. It should be noted that our performance evaluation study is based on realistic data that have been collected from a real operational university campus network, which consists of more than 4000 hosts.

The remaining of this paper is organised as follows. In Section 2, we present some related research work and some observations that motivated our work. In Section 3, the proposed selective flow-sampling approach is described and analysed first, and then its application over a commonly used sequential non-parametric change-point anomaly detection method is presented. A detailed evaluation of the impact of the proposed selective flow-sampling approach on the overall performance of the anomaly detection process using real network traces is included in Section 4 and Section 5 concludes the paper.

2 Related work

The deployment of sampling techniques aims at the provisioning of information about a specific characteristic of the parent population at a lower cost than a full census would demand. The sampling techniques can be divided into two major categories: packet-based and flow-based samplings. In packet-based sampling, packets are selected using a deterministic or non-deterministic method. In flow-based sampling, packets are first classified into flows. A flow is defined as a set of packets that have in common the following packet header fields: source IP address, source port, destination IP address, destination port and protocol. In this case, sampling is performed in flows, which results in the selection of all packets that consists of a particular flow.

Application of packet sampling on network traffic measurements has been extensively studied in the literature [3, 10], mainly for traffic analysis, planning and management purposes. Researchers have proposed schemes that follow an adaptive packet sampling approach in order to achieve more accurate measurements of network traffic. Specifically, an adaptive packet sampling technique for flow-level traffic measurement with stratification approach, which provides unbiased estimation of flow size (in terms of packet and byte counts) for large flows, has been proposed in [11]. In [12], an adaptive packet-level sampling method on different traffic fluctuations and burst scales has been introduced. The method can dynamically adjust each packet sampling probability dependent on the magnitude of traffic fluctuation and can achieve better sampling accuracy contrary to static random sampling.

Furthermore, Hohn and Veitch [4] compared packet sampling with flow sampling and showed that flow sampling performs better in recovering flow distributions, whereas in [13] smart sampling for selecting large flows over small ones has been proposed. However, all these studies and evaluation are targeted towards effective network traffic accounting and are not appropriate for anomaly detection.

The specific problem of studying and understanding the impact of sampling on the anomaly detection process is quite different and far more complicated than the corresponding ones regarding other network management processes. This is mainly due to the fact that anomaly detection may operate under abnormal conditions/attacks, while by its nature involves simultaneously several factors, such as normal traffic, abnormal traffic and various detection metrics, whose statistical characteristics and behaviour may be affected in quite diverse ways by the sampling process.

Recently, researchers started to focus on the impact of sampling on anomaly detection. Mai *et al.* [14], evaluated the impact of packet sampling on three portscan detection algorithms. Their results demonstrated that packet sampling introduces fundamental bias that degrades the detection effectiveness of these algorithms and dramatically

doi: 10.1049/iet-com:20070231

increases false positives. This work was extended in [15] to compare the impact of random packet sampling, random flow sampling, smart sampling [13] and sample-and-hold [16] on specific anomaly detection techniques. The corresponding results demonstrated that random flow sampling performs best, whereas smart sampling and sample-and-hold are not suitable for anomaly detection. Moreover, Brauckhoff *et al.* [17] studied the impact of random packet sampling on the blaster worm anomaly and showed that entropy-based metrics are less affected by sampling than volume-based metrics.

In our previous work [18], we evaluated the impact of three packet-sampling techniques (systematic, random n-out-of-N and uniform probabilistic random sampling) that have been proposed in the PSAMP IETF draft [19] on three widely used anomaly detection algorithms. Our results revealed that systematic sampling does not perform well under low sampling rates when the detection of the attack depends on certain packet characteristics (e.g. TCP ags). Furthermore, we showed that when flow-based metrics such as the number of source IP addresses or number of flows are used, the performance of the anomaly detection algorithm relies mainly on the sampling rate applied and is less dependent on the sampling technique used.

3 Selective sampling for anomaly detection

3.1 Selective sampling

Motivated by the concept of smart sampling [13] where flows are sampled with probability proportional to their size, a new flow-sampling approach – in the following, we refer to as selective sampling – that focuses on the selection of small flows instead is introduced. It should be emphasised that small flows are usually the source of many network attacks [distributed denial of service (DDoS), portscans, worm propagation] [8, 9] and have to be selected in order to maintain an efficient anomaly detection process.

According to the proposed scheme, the selection of an individual flow is based on the following expression

$$p(x) = \begin{cases} c & x \leq z \\ z/(n \cdot x) & x > z \end{cases} \quad (1)$$

where x is the flow size in packets, $0 < c \leq 1$, $n \geq 1$ and z a threshold (measured in packets). As we can observe from (1), flows that are smaller than z are sampled with a constant probability c, whereas flows that are larger in size than z are sampled with probability inversely proportional to their size. One of the main characteristics of our new sampling scheme is that it is more adaptive compared with other approaches (e.g. smart sampling) because it can further control and reduce the number of selected flows. With the appropriate value for parameter c, a significant proportion of small flows can be selected without decreasing the anomaly detection effectiveness. On the other hand, the selection of large flows can be further reduced by increasing the value of parameter n. For large values of n, flows with an enormous number of packets are expected to be missed in the selection process. This fact is very important because we reduce significantly the quantity of sampled packets that need to be processed by the anomaly detection algorithm.

More specifically, let the random variable x represent the flow size in packets and denote by $\ell(x)$ its probability mass function (pmf). The number of selected flows N_f will be given by the following expression

$$N_f = \sum_{x=1}^{N} p(x) \cdot \ell(x) \cdot S_f$$

where S_f is the total number of flows and N the maximum flow size.

Using (1), the above expression becomes

$$N_f = \sum_{x=1}^{z} c \cdot \ell(x) \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n \cdot x} \cdot \ell(x) \cdot S_f$$

The number of selected packets N_p will be given by the following expression

$$N_p = \sum_{x=1}^{N} p(x) \cdot \ell(x) \cdot x \cdot S_f$$

Using (1), we obtain

$$N_p = \sum_{x=1}^{z} c \cdot \ell(x) \cdot x \cdot S_f + \sum_{x=z+1}^{N} \frac{z}{n} \cdot \ell(x) \cdot S_f \quad (2)$$

3.2 Anomaly detection

In this section, we present a sequential non-parametric CPD method that represents a wide class of commonly used anomaly detection strategies. This method utilises one of the most popular algorithms used in the single-metric/single-link anomaly detection approach. This method is independent of the network topology and traffic characteristics and can be applied to monitor every type of network.

doi: 10.1049/iet-com:20070231

The objective of CPD is to determine whether the observed time series is statistically homogeneous and, if not, to find the point in time when the change happens [20, 21]. The attack detection algorithm that is described below belongs to the sequential category of CPD in which tests are done online with the data presented sequentially and the decisions are made on the fly.

Since non-parametric methods are not model-specific, they are more suitable for analysing systems like the Internet which is dynamic and complex. The non-parametric cumulative sum (CUSUM) algorithm is applied for the detection of attacks. The main idea of the non-parametric CUSUM algorithm is that the mean value of a random sequence $\{X_n\}$ is negative during normal operation and becomes positive when a change occurs.

Thus, we can consider $\{X_n\}$ as a stationary random process which, under the normal conditions, the mean is $E(X_n) = k$. A parameter α is chosen to be an upper bound of k, that is, $\alpha > k$, and another random process $\{Z_n\}$ is defined so that $Z_n = X_n - \alpha$, which has a negative mean during normal operation. The purpose of introducing α is to offset the possible positive mean in $\{X_n\}$ caused by small network anomalies so that the test statistic y_n, which is described below, will be reset to zero frequently and will not accumulate with time.

When an attack takes place, Z_n will suddenly increase and become a large positive number. Suppose, during an attack, the increase in the mean of Z_n can be lower bounded by h. The change detection is based on the observation of $h \gg k$.

More specifically, let $y_n = (y_{n-1} + Z_n)^+$

$$y_0 = 0$$

where x^+ is equal to x if $x > 0$ and 0 otherwise. The decision function can be described as follows

$$d_N(y_n) = 0 \quad \text{if } y_n \le N$$
$$d_N(y_n) = 1 \quad \text{if } y_n > N$$

where $d_N(y_n)$ is the decision at time n: '0' stands for normal operation and '1' for attack (a change occurs) and N represents the attack threshold. This anomaly detection method has been used with different types of metrics, such as the SYN/FIN packet ratio [20] or the percentage of new source IP addresses in a time bin [21] in order to detect DoS attacks.

3.3 Application of selective sampling over CPD with SYN/FIN ratio

In this section, we describe the application and impact of the proposed selective sampling over a CPD-based anomaly detection approach [20] that uses as a metric (variable X_n) the difference between SYN and FIN packets divided by the number of FIN packets. The SYN packets are the ones that have the TCP SYN flag set, whereas the number of FIN packets that we use is actually the sum of packets that have the FIN or RST flag set. SYN packets denote the beginning of a TCP connection, whereas FIN packets indicate the normal termination of a TCP connection and RST packets indicate abnormal termination. Under normal conditions, the appearance of an SYN packet will be followed either by an FIN packet or an RST packet when the connection terminates. Thus, X_n has a positive mean value near zero in normal condition, whereas Z_n has a negative value. When an SYN attack occurs, X_n becomes a large positive number and as a result Z_n becomes positive. This leads to the increase in y_n which indicates an attack if its value exceeds a certain threshold.

The detection efficiency of the CPD method is reflected by the slope of the curve of y_n and is defined by the corresponding degree of detection, expressed in relation (3). In order to achieve better degree of detection, the value of X_n, which is the difference between SYN and FIN packets divided by the number of FIN packets, must be maximised. Therefore the following expression that gives the degree of detection must be maximised

$$\text{Degree of detection} = \frac{N_{\text{SYN}} - N_{\text{FIN}}}{N_{\text{FIN}}} \tag{3}$$

where N_{SYN} is the total number of selected SYN packets in a time bin and N_{FIN} the total number of selected FIN packets in the same time bin.

The total number of selected SYN packets, N_{SYN}, is given by the following expression

$$N_{\text{SYN}} = \sum_{x=1}^{N} \left(p(x) \cdot \ell(x) \cdot S_{\text{f}} \cdot \sum_{y=1}^{x} (y \cdot f(y|x)) \right)$$

where S_{f} is the total number of flows, N the maximum flow size, $\ell(x)$ the pmf of random variable x, which represents the flow size in packets, $p(x)$ the probability of selecting a flow of size x and is given by (1) and $f(y|x)$ the conditional probability of a flow having y SYN packets given that the flow size is equal to x.

doi: 10.1049/iet-com:20070231

Using (1), we obtain

$$N_{\mathrm{SYN}} = \sum_{x=1}^{z}\left(c \cdot \ell(x) \cdot S_{\mathrm{f}} \cdot \sum_{y=1}^{x}\left(y \cdot f(y|x)\right)\right) + \sum_{x=z+1}^{N}\left(\frac{z}{n \cdot x} \cdot \ell(x) \cdot S_{\mathrm{f}} \cdot \sum_{y=1}^{x}\left(y \cdot f(y|x)\right)\right) \quad (4)$$

The total number of selected FIN packets, N_{FIN}, is given by the following expression

$$N_{\mathrm{FIN}} = \sum_{x=1}^{N}\left(p(x) \cdot \ell(x) \cdot S_{\mathrm{f}} \cdot \sum_{y=1}^{x}\left(y \cdot g(y|x)\right)\right)$$

where $g(y|x)$ is the conditional probability of a flow having y FIN packets given that the flow size is equal to x. Using (1), we obtain

$$N_{\mathrm{FIN}} = \sum_{x=1}^{z}\left(c \cdot \ell(x) \cdot S_{\mathrm{f}} \cdot \sum_{y=1}^{x}\left(y \cdot g(y|x)\right)\right) + \sum_{x=z+1}^{N}\left(\frac{z}{n \cdot x} \cdot \ell(x) \cdot S_{\mathrm{f}} \cdot \sum_{y=1}^{x}\left(y \cdot g(y|x)\right)\right) \quad (5)$$

Substituting (4) and (5) in (3), we obtain an expression that states the degree of detection with respect to parameters z, c and n. By selecting the appropriate values for the parameters z, c and n, we can maximise (3), which will result in a better degree of attack detection. The impact of these parameters on the attack detection accuracy is studied in detail in the next section.

4 Performance evaluation and discussions

In this section, the effectiveness of the selective flow-sampling technique designed in Section 3.1 combined with the anomaly detection method described in Section 3.2, under different attack/anomalies scenarios, is studied and evaluated. The results and corresponding observations presented in this section are based on real network data that have been collected from an operational campus network. More specifically, we monitored the link between the National Technical University of Athens (NTUA) and the Greek Research and Technology Network (GRNET), which connects the university campus with the Internet. This link has an average traffic of 70–80 Mbit/s and 20 000 packets/s, containing a rich network traffic mix carrying standard network services such as web, mail, ftp and p2p application traffic. In the following evaluation, a DDoS attack (TCP SYN attack) against a single host inside the NTUA campus generated by a real DoS attack tool is studied in detail.

In our experiments, in order to gain some insight regarding the effectiveness of the selective sampling technique, we first compare it against random flow sampling. In random flow sampling, each flow is selected independently with the same probability p. The percentage of sampled flows was chosen as the common criterion for the comparison. Furthermore, in order to better evaluate and understand the operational characteristics of selective sampling, we study the tradeoff between the achievable degree of detection and the number of selected packets during the process of selective sampling with respect to its various design parameters.

4.1 Comparative results

To better demonstrate the results and corresponding observations, we study two scenarios for the CPD method. In the first scenario, we compare selective sampling with the random flow sampling experimenting with attacks that correspond to 2% of the background traffic (in packets), whereas in the second scenario, we further decrease the attack rate to correspond to 1% of the background traffic. These scenarios are examined under different sampling rates.

Fig. 1 presents the flow size distribution (in packets) of the attack flows in the case of 2% attack rate. As we can observe most of the attack flows have 1 or 2 packets.

In Fig. 1, we choose some characteristic values for the parameters of selective sampling. More specifically, we select values for z in order to distinct between small and large flows. In our case, where the attack spreads

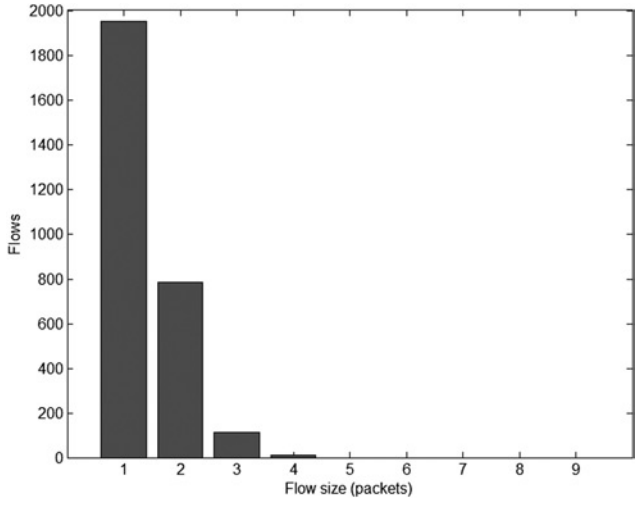

Figure 1 *Flow size distribution of the attack flows for 2% attack rate*

doi: 10.1049/iet-com:20070231

mainly among flows that have 1 or 2 packets, we choose $z = 1$ and 2. In the following, with the appropriate value for parameter c, we define the proportion of small flows that are sampled (in our experiments, we use $c = 1.0$ and 0.2), and with parameter n, we adjust the selection of large flows. Greater values of n result in the selection of smaller number of large flows and consequently fewer packets. For the random flow sampling algorithm, we select the appropriate probability p which results in the same percentage of sampled flows. Table 1 presents the parameters for both sampling techniques and the corresponding percentage of sampled packets.

It should be noted that the percentage of selected packets in the case of selective sampling is significantly smaller than the corresponding for the case of random flow sampling. The number of selected packets is of major importance, because the anomaly detection algorithm inspects every sampled packet in order to examine the TCP-flags header field.

In general, given the fact that small flows are usually the source of many network attacks, the network operator can select the appropriate values for z, c and n to detect a large set of attacks that consist of small flows. Because of the fact that selective sampling targets small flows, the number of sampled packets is very small compared with the random sampling case in which flows are selected with the same probability, independent of their size. This is clearly demonstrated by the results presented in Table 1; in selective sampling, only 4% of the total number of packets were sampled, whereas in the corresponding random flow sampling case, the percentage of sampled packets was ~45%.

The results depicted in Fig. 2 correspond to 2% attack rate at the sampling rate of 45% with respect to the number of flows. As we can observe, the curve that corresponds to random flow sampling resembles the curve of the unsampled case. On the other hand, the selective sampling outperforms both these cases (unsampled case and random flow sampling) since the slope of the corresponding curve

Table 1 Parameters for each sampling technique and the corresponding packet percentage

Flows, (%)	Random flow sampling		Selective sampling			
	p	Packets, (%)	z	c	n	Packets, (%)
45.12	0.45	45.67	1	1.0	3	4.25
12.05	0.12	12.53	2	0.2	6	2.05

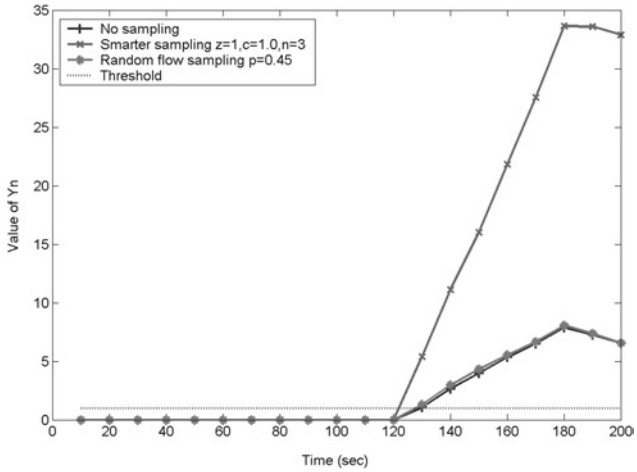

Figure 2 *Attack detection for 45% flow sampling (2% attack rate)*

for the selective sampling has increased significantly. As we mentioned earlier, the slope represents the degree of detection. In this scenario, it is obvious that the detection effectiveness is significantly improved in the case of selective sampling. This is attributed to the fact that fewer FIN packets were sampled (which are normally present on large flows), and at the same time, all the SYN attack packets have been captured.

In Fig. 3, we present the corresponding results for a sampling rate of 12% with respect to the number of sampled flows. In this case, the slope of the curve is decreased, but we still achieve a better degree of attack detection in the selective sampling case compared with the unsampled case, even though a small value of $c = 0.2$ was used for the selection probability of small flows.

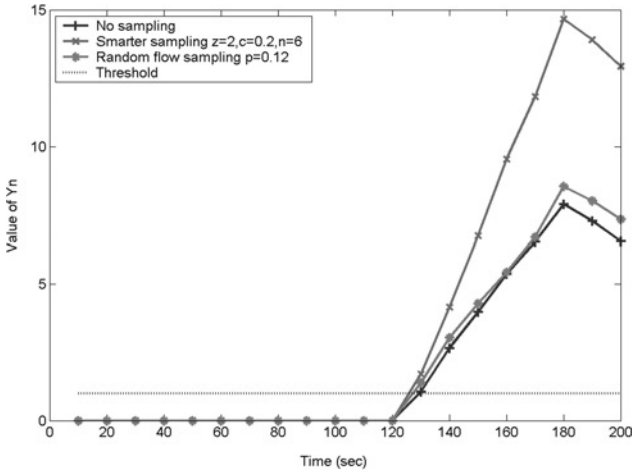

Figure 3 *Attack detection for 12% flow sampling (2% attack rate)*

doi: 10.1049/iet-com:20070231

In the following, we reduced the attack rate to correspond to 1% of the background traffic. Fig. 4 shows the flow size distribution (in packets) of the attack flows. As we can observe most of the attack flows have 1 packet. For this attack rate, we applied random flow sampling and selective sampling with the same sampling parameters as in the previous case of 2% attack rate. The corresponding results are depicted in Figs. 5 and 6.

More specifically, in Fig. 5, we present the corresponding results for the sampling rate of 45% with respect to the number of sampled flows. It is obvious that in the selective sampling case, the degree of attack detection is still larger than the corresponding of the unsampled case. This is due to the fact that a significant portion of the attack flows has been sampled. On the contrary, we have a smaller degree of detection for the random flow sampling case. More specifically, we observe that the attack is not detected in time bin of 130 s for the random flow sampling case.

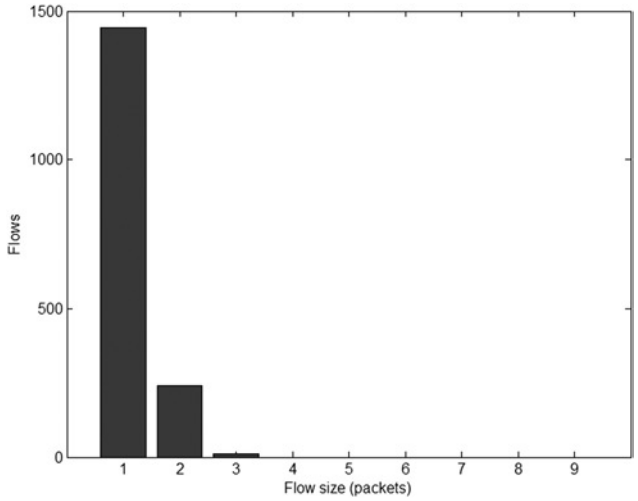

Figure 4 *Flow size distribution of the attack flows for 1% attack rate*

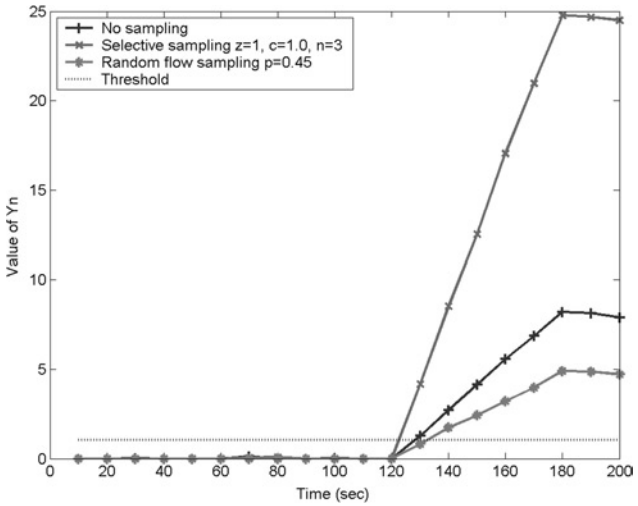

Figure 5 *Attack detection for 45% flow sampling (1% attack rate)*

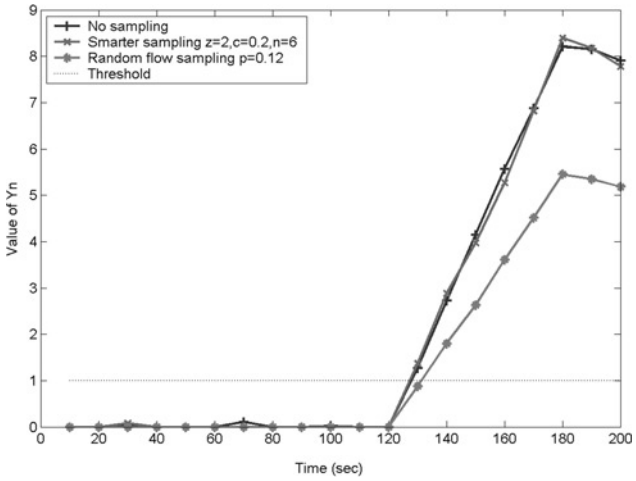

Figure 6 *Attack detection for 12% flow sampling (1% attack rate)*

Fig. 6 presents the corresponding results for 12% flow sampling using 1% attack rate. The slope of the curve for the case of selective sampling has decreased and matches the unsampled case. This is attributed to the fact that a smaller number of SYN packets is generated during the attack, thus causing the difference between the sum of SYN and FIN packets to decrease. However, the algorithm fails to detect the attack during the first time bin for the case of random flow sampling.

4.2 Design tradeoffs for selective sampling

In the following, we study the tradeoff between the degree of detection [as it was defined in (3)] and the number of selected packets during the process of selective sampling. This study is based on the scenario of SYN attack rate corresponding to 1% of the background traffic. On the basis of the fact that the major percentage of attack flows has a single packet (Fig. 4), we chose parameter $z = 1$. Using (3)–(5), we calculate the degree of detection for several values of parameters c and n.

Table 2 presents the degree of detection for different values of parameter c within the range $0.1 \leq c \leq 1.0$ and step 0.1, and for different values of parameter n ranging from 1 to 1000. As we mentioned earlier in Section 3, parameter c defines the constant probability of the selection of small flows (flows that have equal or less than z packets), whereas

doi: 10.1049/iet-com:20070231

Table 2 Degree of detection with respect to parameters c and n for selective sampling

c	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 1000$
0.1	0.72	1.57	2.36	3.41	4.89	5.77	6.94
0.2	0.96	2.36	3.41	4.55	5.77	6.36	7.02
0.3	1.18	2.95	4.08	5.15	6.15	6.59	7.05
0.4	1.38	3.41	4.55	5.52	6.36	6.71	7.06
0.5	1.57	3.78	4.89	5.77	6.50	6.78	7.07
0.6	1.75	4.08	5.15	5.95	6.59	6.83	7.07
0.7	1.92	4.33	5.35	6.09	6.66	6.87	7.08
0.8	2.07	4.55	5.52	6.20	6.71	6.90	7.08
0.9	2.22	4.73	5.66	6.29	6.75	6.92	7.08
1.0	2.36	4.89	5.77	6.36	6.78	6.94	7.08

Table 3 Number of selected packets with respect to parameters c and n for selective sampling

c	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 1000$
0.1	9536	2471	1588	1147	882	793	714
0.2	10 241	3176	2293	1852	1587	1499	1419
0.3	10 946	3881	2998	2557	2292	2204	2124
0.4	11 651	4587	3703	3262	2997	2909	2829
0.5	12 356	5292	4409	3967	3702	3614	3534
0.6	13 062	5997	5114	4672	4407	4319	4239
0.7	13 767	6702	5819	5377	5112	5024	4945
0.8	14 472	7407	6524	6082	5817	5729	5650
0.9	15 177	8112	7229	6787	6523	6434	6355
1.0	15 882	8817	7934	7493	7228	7139	7060

parameter n characterise the selection of large flows. The selection of large flows is reduced by increasing the value of parameter n. Specifically, in Tables 3 and 4, we present the absolute number and the corresponding percentage of selected packets (with reference to the unsampled case) in a time bin of 10 s with respect to parameters c and n.

In the lower left corner of Tables 3 and 4, the number and percentage of selected packets, respectively, for the values $c = 1$ and $n = 1$ for selective sampling are depicted. These values correspond to the simple case of selective sampling that resembles the reverse of smart sampling [13]. In the upper left corner, we observe the case where only the 10% of small flows are selected, which results in a decrease of selected packets. In the lower right corner of the table, the case in which the majority of large flows has been discarded is depicted, and at the same time all small flows have been selected. Finally, in the upper right corner, we can observe the case where $c = 0.1$ and $n = 1000$. This corresponds to the case where we have selected the minimum number of packets. As it is illustrated in Table 4, the proposed algorithm provides the flexibility of reducing the percentage of selected packets from 8.14% to 0.37% by appropriately defining the corresponding values for parameters c and n.

doi: 10.1049/iet-com:20070231

Table 4 Percentage of selected packets with respect to parameters c and n for selective sampling

c	$n = 1$, (%)	$n = 5$, (%)	$n = 10$, (%)	$n = 20$, (%)	$n = 50$, (%)	$n = 100$, (%)	$n = 1000$, (%)
0.1	4.89	1.27	0.81	0.59	0.45	0.41	0.37
0.2	5.25	1.63	1.18	0.95	0.81	0.77	0.73
0.3	5.61	1.99	1.54	1.31	1.18	1.13	1.09
0.4	5.97	2.35	1.90	1.67	1.54	1.49	1.45
0.5	6.34	2.71	2.26	2.03	1.90	1.85	1.81
0.6	6.70	3.08	2.62	2.40	2.26	2.21	2.17
0.7	7.06	3.44	2.98	2.76	2.62	2.58	2.54
0.8	7.42	3.80	3.35	3.12	2.98	2.94	2.90
0.9	7.78	4.16	3.71	3.48	3.35	3.30	3.26
1.0	8.14	4.52	4.07	3.84	3.71	3.66	3.62

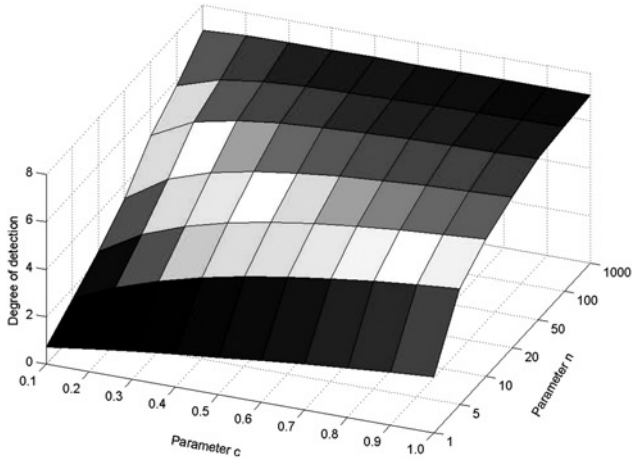

Figure 7 *Degree of detection based on parameters c and n for selective sampling*

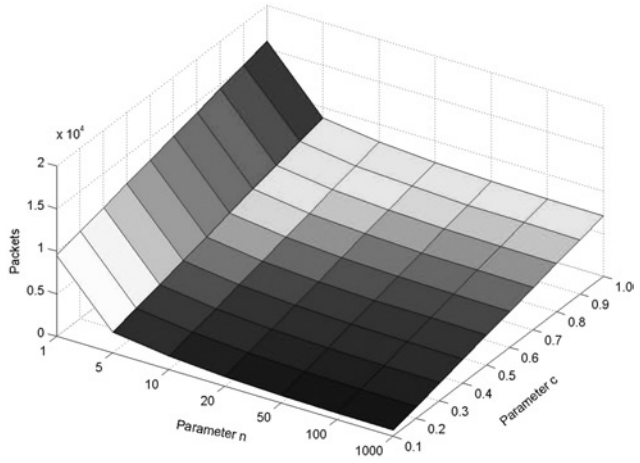

Figure 8 *Number of selected packets based on parameters c and n for selective sampling*

In Fig. 7, we graphically present the degree of detection for the above values of parameters c and n. In this figure, the degree of detection increases as parameter c increases, whereas parameter n remains constant, due to the fact that the majority of the attack flows in our scenario has only one packet. Thus, for large values of c, the absolute number of SYN packets is larger, which results in a greater difference between the sum of SYN and FIN packets causing the increase of the degree of detection. In contrast, the degree of detection decreases with the reduction in parameter n for a constant value of parameter c. This is attributed to the fact that for small values of n, we tend to select more large flows. Large flows are likely to contain one SYN and one FIN packet. The selection of these packets causes the ratio of the difference between SYN and FIN packets divided by the number of FIN packets to decrease, as the denominator increases.

As mentioned before, our goal is to achieve a high degree of detection while maintaining a small number of processed packets. The number of selected packets is given by (2). In Fig. 8, we present the number of selected packets with respect to parameters c and n. As we can observe, the number of selected packets decreases significantly for large values of the parameter n and small values of parameter c. The decrease in selected packets is significant from the case of $n = 1$ to 5, as a considerable amount of large flows is discarded. Combining the results of Figs. 7 and 8, we observe that for the case under consideration optimal values for parameters n and c would be 1000 and 0.7. Specifically, from Fig. 7 (which corresponds to Table 2), we note that the best degree of detection

doi: 10.1049/iet-com:20070231

is given for values $n = 1000$ and $c = 0.7$, 0.8, 0.9 and 1.0. Fig. 8 (which corresponds to Tables 3 and 4) depicts the number of selected packets for each pair of parameters c and n. Therefore the values $c = 0.7$ and $n = 1000$ provide the best degree of detection with the minimum number of selected packets.

5 Concluding remarks

In this paper, we considered the problem of improving network anomaly detection effectiveness and efficiency through the introduction and application of selective flow-based sampling. Among the key motivations of our approach is the exploitation of the fact that for specific-purpose applications (such as anomaly detection) a large fraction of information is contained in a small fraction of flows. Therefore based on the observation that a wide range of attacks usually originate from flows with a small number of packets (e.g. DDoS attacks, portscans, worm propagation and so on), we introduced and analysed a new flow-based sampling technique that focuses on the preferential selection of small flows.

Furthermore, we evaluated the impact of this sampling technique under different scenarios in a highly DDoS attack using a sequential non-parametric change-point anomaly detection method. Our experiments and results demonstrated that even with small attack and sampling rates, the detection effectiveness is significantly improved and in some scenarios outperforms even the unsampled case, and at the same time, the number of packets that need to be processed is reduced.

Moreover, the insight gained from this study can also be used for the design of improved anomaly detection algorithms. For instance, to achieve a better degree of detection in the unsampled or the random flow sampling case, we could consider a weight for each SYN packet taking into account the flow size that the particular SYN packet belongs to. In the current CPD algorithm, the size of the flow that belongs to an SYN packet is not considered. However, as we mentioned earlier in this paper, an SYN packet that belongs to a small flow (e.g. a single-packet flow) is more likely to be part of an attack. Thus, if the CPD algorithm used a weighted sum for the SYN packets, the attack detection would be more effective, especially in the case of small attacks.

Finally, the selective sampling technique described in this paper, can be incorporated with IP traceback techniques [22, 23] that focus on tracing the attack flows from the target back to the real sources. More specifically, deterministic packet marking (DPM) [24] which marks every packet at the ingress router of a network could benefit from selective sampling if the packet marking process is applied after the sampling process. A considerable amount of packets that is usually not part of the attack does not need to be marked, thus making the DPM scheme more efficient.

6 References

[1] DERI L., SUIN S.: 'Effective traffic measurement using ntop', *IEEE Commun. Mag.*, 2000, **38**, (5), pp. 138–143

[2] MCGREGOR T., BRAUN H.W., BROWN J.: 'The NLANR network analysis infrastructure', *IEEE Commun. Mag.*, 2000, **38**, (5), pp. 122–128

[3] DUFFIELD N., LUND C., THORUP M.: 'Estimating flow distributions from sampled flow statistics', *IEEE/ACM Trans. Netw.*, 2005, **13**, (5), pp. 933–946

[4] HOHN N., VEITCH D.: 'Inverting sampled traffic', *IEEE/ACM Trans. Netw.*, 2006, **14**, (1), pp. 68–80

[5] BARFORD P., KLINE J., PLONKA D., *ET AL.*: 'A signal analysis of network traffic anomalies'. Proc. 2nd ACM SIGCOMM Workshop on Internet Measurement, Marseille, France, November 2002, pp. 71–82

[6] YE N., EMRAN S., CHEN Q., *ET AL.*: 'Multivariate statistical analysis of audit trails for host-based intrusion detection', *IEEE Trans. Comput*, 2002, **51**, (7), pp. 810–820

[7] LEE W., XIANG D.: 'Information-theoretic measures for anomaly detection'. Proc. IEEE Symp. Security and Privacy, Oakland, CA, USA, May 2001, pp. 130–143

[8] BARFORD P., PLONKA D.: 'Characteristics of network traffic flow anomalies'. Proc. 1st ACM SIGCOMM Internet Measurement Workshop, San Francisco, CA, USA, November 2001, pp. 69–74

[9] SRIDHARAN A., YE T., BHATTACHARYYA S.: 'Connectionless port scan detection on the backbone'. Malware Workshop (in conjunction with IEEE IPCCC 2006), Phoenix, AZ, USA, April 2006

[10] ESTAN C., KEYS K., MOORE D., *ET AL.*: 'Building a better netflow'. Proc. ACM SIGCOMM'04, Portland, OR, USA, August 2004, pp. 245–256

[11] CHOI B.Y., PARK J., ZHANG Z.L.: 'Adaptive packet sampling for accurate and scalable flow measurement'. IEEE Global Telecommunications Conf. (GLOBECOM'04), Dallas, TX, USA, November 2004, pp. 1448–1452

[12] XU L.B., WU G.X., LI J.F.: 'Packet-level adaptive sampling on multi-fluctuation scale traffic'. Proc. Int. Conf. Communications, Circuits and Systems, Hong-Kong, China, May 2005, pp. 604–608

[13] DUFFIELD N.G., LUND C.: 'Predicting resource usage and estimation accuracy in an IP flow measurement collection infrastructure'. ACM SIGCOMM Internet Measurement Conf., Miami, FL, USA, October 2003, pp. 179–191

[14] MAI J., SRIDHARAN A., CHUAH C.N., *ET AL.*: 'Impact of packet sampling on portscan detection', *IEEE J. Sel. Areas Commun.*, 2006, **24**, (12), pp. 2285–2298

[15] MAI J., SRIDHARAN A., CHUAH C.N., *ET AL.*: 'Is sampled data sufficient for anomaly detection?'. Internet Measurement Conf., Rio de Janeiro, Brazil, October 2006, pp. 165–176

[16] ESTAN C., VARGHESE G.: 'New directions in traffic measurement and accounting'. Proc. SIGCOMM'02, Pittsburgh, PN, USA, August 2002, pp. 323–336

[17] BRAUCKHOFF D., TELLENBACH B., WAGNER A., *ET AL.*: 'Impact of packet sampling on anomaly detection metrics'. Internet Measurement Conf., Rio de Janeiro, Brazil, October 2006, pp. 159–164

[18] ANDROULIDAKIS G., CHATZIGIANNAKIS V., PAPAVASSILIOU S., *ET AL.*: 'Understanding and evaluating the impact of sampling on anomaly detection techniques'. IEEE Military Communications Conf., Washington, DC, USA, October 2006

[19] PACKET SAMPLING (PSAMP) IETF WORKING GROUP CHARTER: available at: http://www.ietf.org/html.charters/psamp-charter.html

[20] WANG H., ZHANG D., SHIN K.G.: 'Change-point monitoring for the detection of DoS attacks', *IEEE Trans. Dependable Secur. Comput.*, 2004, **1**, (4), pp. 193–208

[21] PENG T., LECKIE C., RAMAMOHANARAO K.: 'Proactively detecting distributed denial of service attacks using source IP address monitoring'. Proc. 3rd Int. IFIP-TC6 Networking Conf., Athens, Greece, May 2004

[22] BELENKY A., ANSARI N.: 'On IP traceback', *IEEE Commun. Mag.*, 2003, **41**, (7), pp. 142–153

[23] GAO Z., ANSARI N.: 'A practical and robust inter-domain marking scheme for IP traceback', *Comput. Netw.*, 2007, **51**, (3), pp. 732–750

[24] BELENKY A., ANSARI N.: 'On deterministic packet marking', *Comput. Netw.*, 2007, **51**, (10), pp. 2677–2700

doi: 10.1049/iet-com:20070231