Journal of Computer and Communications, 2015, 3, 1-8
Published Online September 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.39001

Two Pass Port Scan Detection Technique Based on Connection Pattern and Status on Sampled Data

Sunil Kumar, Kamlesh Dutta, Ankit Asati

Department of Computer Science and Engineering, National Institute of Technology, Hamirpur-177005, Himachal Pradesh, INDIA
Email: sunilkaushik27@gmail.com, kd@nith.ac.in, ankitasati.nith@gmail.com

Received 21 July 2015; accepted 29 August 2015; published 1 September 2015

Abstract

Anomaly detection is now very important in the network because the increasing use of the internet and security of a network or user is a main concern of any network administrator. As the use of the internet increases, so the chances of having a threat or attack in the network are also increasing day by day and traffic in the network is also increasing. It is very difficult to analyse all the traffic data in network for finding the anomaly in the network and sampling provides a way to analyse the anomalies in network with less traffic data. In this paper, we propose a port scan detection approach called CPST uses connection status and pattern of the connections to detect a particular source is scanner or benign host. We also show that this approach works efficiently under different sampling methods.

Keywords

Port Scan, TRW, TAPS, CPST, Packet Sampling, Flow Sampling

1. Introduction

Traffic analysis is essential for the network security, especially for intrusion detection system. Port scanning is one of the anomaly detection, which is generally carried out in the network for the security purpose. When an intruder or attacker wants to do any harmful activity in the network, then first he want to analyse the entire network, for example, which operating systems are using in network or what ports are open or accessible or which service is running on the particular host. So there is a need of intrusion detection techniques which identify the scanner in the early stage of network based on sampled as well as non-sample data and generate the alert to the

How to cite this paper: Kumar, S., Dutta, K. and Asati, A. (2015) Two Pass Port Scan Detection Technique Based on Connection Pattern and Status on Sampled Data. *Journal of Computer and Communications*, **3**, 1-8.
http://dx.doi.org/10.4236/jcc.2015.39001

network administrator.

In the present scenario, the network becomes larger and larger day by day and the link speed is also increasing. This results in the huge amount of traffic data in the network. It is very difficult to analyse that huge data due to limited resources like CPU, memory, etc. So to cope with the increasing link speed, sampled traffic data are used as an input for various anomaly detection or scan detection like "Denial of service attack" or "Port scan attack". However, sampling distorts traffic statistics such as mean rate and flow size distribution. But it is very useful for analysing the network traffic for detecting the attacks. Therefore, various sampling methods like packet sampling such as Cisco Net-flow [1] or flow sampling are often deployed in the routers and other devices. These techniques are also used for sampling offline data. The inputs on those devices are the original traffic and the output becomes the thinned traffic data for the detection of anomalies. Traditionally, an IP Flow is based on a set of five IP packet attributes.

IP Packet attributes used by Net-Flow:

- IP source address;
- IP destination address;
- Source port number;
- Destination port number;
- Protocol type.

In the literature, various port scan detection techniques have been developed like TRW [2], Snort [3] [4], TAPS [5], Snort Honeypot [6] etc. In this paper, a two pass port scan detection technique called CPST (Connection Pattern and Status Based Port Scan Detection Technique) is proposed, which is based on the concept of existing detection algorithms TRW [2] and TAPS [5].

The main idea behind CPST is that it is based on connection status as well as pattern to have a low degree of false positive and high degree of efficacy. We pursue the problem in the general framework of port scan detection through connection status as well as pattern of connection in the sampled data. In connection status approach a decision is made on the basis of the status of the connection, *i.e.* the connection is established or connection is failed. In connection pattern approach, a decision is made on the basis of the pattern of the connections, for example, calculate the ratio between the destination IP's and the destination Port's and then make a decision based on those values, that particular source is scanner or benign host.

The remaining of this paper is organized as follows. In Section 2, the port scan detection is widely explained. In Section 3, we describe existing two sampling algorithms (TRW and TAPS) used in the network. In Section 4, we provide the detail of CPST approach with mathematical analysis. In Section 5, we compare the performance of CPST with TRW and TAPS under two sampling techniques and Section 6 concludes the paper.

2. Port Scan Detection

In [7], Lee, Roedel and Silenok presented a comprehensive study of different scanning attacks with their general characteristics. They classified the scanning attacks into three categories: "vertical", "horizontal" and "mixture of horizontal and vertical". First type refers to a scanner looking for open ports on a single target destination IP (scanner scans various ports for a single IP and finds the possibility of some open ports for that IP); while the second one refers to a scanner looking for one specific open port on several target machines. For example, scanner scans http port 80 for every IP present in the network. The "mixture" scan type refers to a scanner checking fixed range of ports on a specific set of destination machines. The most basic detection mechanism just maintains a counter of number of contacted destination ports and IPs with a given source IP.

A large number of port scan detection techniques have been proposed in the literature. These techniques are broadly categorized into two categories, namely "single source scan detection" and "distributed scan detection". These techniques are further divided into sub categories like threshold based, algorithmic based, soft computing based or rule based etc. [8]. TRW and TAPS are two main two port scan detection techniques.

2.1. Threshold Random Walk (TRW)

The main idea behind TRW [2] method is that scanners will fail for more connections than a benign host, thus classifying a host as a scanner when it makes too many consecutive failed connections. One of the main characteristic of the scanner is that they are more likely to choose those remote hosts which do not exist or do not have the requested service activated. This algorithm performs probabilities reasoning and sequential hypothesis test-

ing to observe the connection status of each source. According to TRW algorithm, if a given remote source tries to connect with a local host l, the connection attempt can a successful (marked 0) or a failure (marked 1). Then, the system can decide whether the remote host is a scanner or benign host based on sequence of connection attempts and test of sequential hypothesis. This algorithm requires very few packets (only four or five) to reach a conclusion and does not require any training of the system beforehand. It focuses only on TCP traffic for detection of port scan attack. With these results, the sources or hosts which are benign host come under the hypothesis $H0$ and the sources which are scanner come under the hypothesis $H1$.

For a given source r let Y_i be a random variable that represents the outcome of the first connection attempt by r to the ith distinct local host, where

$$Y_i = \begin{cases} 0 & \text{if connection attempt is successful} \\ 1 & \text{if connection attempt is unsuccessful} \end{cases}$$

With these two hypotheses, four outcomes are possible when a decision is made as shown in **Table 1**.

2.2. Time Based Access Pattern Sequential Hypothesis Testing (TAPS)

TAPS [5] is based on the observation that scanners visit many more destination IPs vs. ports (or the reverse) than benign host. It utilizes the access pattern of each source for hypotheses testing. This technique is based on the concept of horizontal and vertical scanning *i.e.* either the scanner access a particular port number on a multiple destination machine (so that IP/Port $\gg k$) or a scanner wants to access a list of various port number on to a single destination machine (so that Port/IP $\gg k$). All the hosts which have IP/Port $\gg 1$ or vice versa are considered as a scanner. TAPS does not depend on any specific property of the packet as TRW (looks for single SYN-packet flows). TAPS is connectionless-oriented (works with both UDP and TCP) whereas TRW works only with TCP scanners.

In [9], Mai, Sridharan, Chuah, Zang and Ye analyzed the impact of packet sampling on TRW and TAPS. The simulation results demonstrate that flow size becomes lower in the presence of sampling which results in more false positive rates in TRW as compared to TAPS. TAPS exhibits lower false positive rates.

In [10], Mai, Chuah, Sridharan, Ye and Zang tested several sampling methods (*Packet Sampling*, *Flow Sampling*, *Sample-and-Hold* and *Smart Sampling*) against port scan detection techniques TRW and TAPS. The experiment results demonstrate that TRW is less resilient to sampling as compared to TAPS. TAPS exhibits lower false positive ratio and TRW has a better success ratio. They concluded the paper with the assessment that flow sampling performed better for both port scan techniques while the other sampling methods produce very poor results.

3. Sampling Techniques

In this section, two sampling techniques are described: random packet sampling and random flow sampling.

3.1. Random Packet Sampling

Packet sampling techniques are currently being standardized by the Packet Sampling (PSAMP) Working Group of the Internet Engineering Task Forces (IETF) [11]. In packet sampling technique, each packet is considered with probability p. In this method n samples are selected out of N packets, hence it is sometimes called *n*-out-of-*N* sampling. For this sampling each packet has an equal chance of being drawn. One way for simple random sample is to randomly generate n different numbers in the range of 1 to *N* and then choose all packets with a packet position equal to one of these *n* numbers. This procedure is repeated for every *N* packet [12]. The

Table 1. Possible outcomes of TRW algorithm under two hypothesis.

Sr. No.	Original source	Algorithm outcomes	Decision
1	Scanner (under $H1$)	Under $H1$	True Detection
2	Scanner (Under $H1$)	Under $H0$	False Negative
3	Benign Host (Under $H0$)	Under $H1$	False Positive
4	Benign Host (Under $H0$)	Under $H0$	Normal

packet sampling technique is mainly of two types (1) Systematic Packet Sampling and (2) Random Packet Sampling. Systematic packet sampling involves the selection of packets into a systematic method or according to a deterministic function. In Random packet sampling the selection of packets is generated according to a random process.

3.2. Random Flow Sampling

A flow in RTFM [13] model can be loosely defined as the set of packets that have in common values of certain fields found in the headers of packets. The fields used to aggregate traffic typically specify addresses at various levels of the protocol stack (e.g. IP addresses, IP protocol, and TCP/UDP port numbers). The flow is also defined as a unidirectional set of packets that arrive at the router on same sub-interface, have the same source and destination IP address, have the same source and destination port, same protocol and the same type of service bytes in IP header. This technique usually implements hashing table of flow ID which consist IP address, port number and the protocol type. The flow is then selected if the resulted value in below than a specified value q [14].

4. CPST (Connection Pattern and Status Based Port Scan Detection Technique)

One of the main characteristics of the scanner is that maximum time they do not make a successful connection with the server or destination or they do not complete three way handshaking. The second characteristic of the scanner is the ratio between the destination host ip vs. destination host port for a particular source is always greater than a particular value k. So, if the ratio of destination ip/port or destination port/ip is greater than this value k, then the particular source is treated as a scanner, and if its value is less than k then it is declared as benign host.

The novel feature of our proposedtwo pass port scan detection technique—CPST is that it uses both connections status and connection pattern approaches for the detection of scanners. In connection status approach, a decision is made on the basis of the status of the connection, *i.e.* the connection is established or connection is failed. In connection pattern approach, a decision is made on the basis of the pattern of the connections, for example, calculate the ratio between the destination IP's and the destination Port's and then make a decision based on those values, that particular source is scanner or benign host. In CPST, two levels of detection are performed. In the first level, the scanner is detected on the basis of pattern of destination ip/port or vice versa for a particular host.

In the second level, connection status is checked and a decision in made for a source in accordance to connection status. CPST is based on the sequential and pattern inference testing. Sequential inference testing observes connection status of each source IP in a flow to check whether the connection is fail or successful. For particular, IP if connection is fail then there are more chances of having a scanner or if the connection is successful then there are more chances of having benign host. Similarly pattern inference testing observes the connection pattern of each source to check whether it is scanner or benign host (see **Figure 1**).

Let us suppose that when a remote source or a local source r makes a connection attempt to a local destination, then an event E is generated. The result of that event is either a "success" or a "failure", depending on the connection status of the particular source. Now there are two possibilities of connection of a particular source to a destination host, either the source tries a connection attempt to an inactive host or to an inactive service or it tries a connection attempt to an active host or active service. Now if the host is a scanner then it will try to connect with different ports on a same destination IP or same port on different destination IP addresses.

In CPST, sequential hypothesis technique is used. As per the metric of access pattern for a scanner:

$$\text{DEST-IP}/\text{DEST-PORT} \gg 1 \quad \text{or} \quad \text{DEST-PORT}/\text{DEST-IP} \gg 1$$

The indicator random variable is defined as follows:

$$Y_i = \begin{cases} 0 & \text{if DEST-IP}/\text{DEST-PORT} \le k \text{ and DEST-PORT}/\text{DEST-IP} \le k \left(\text{unsuccessful event}\right) \\ & \text{and if event } i \text{ is successful} \\ 1 & \text{if DEST-IP}/\text{DEST-PORT} > k \text{ or DEST-PORT}/\text{DEST-IP} > k \left(\text{successful event}\right) \\ & \text{or if event } i \text{ is not successful} \end{cases}$$

There are possibilities of four events associated with the random variable Y_i and their probabilities [2]:

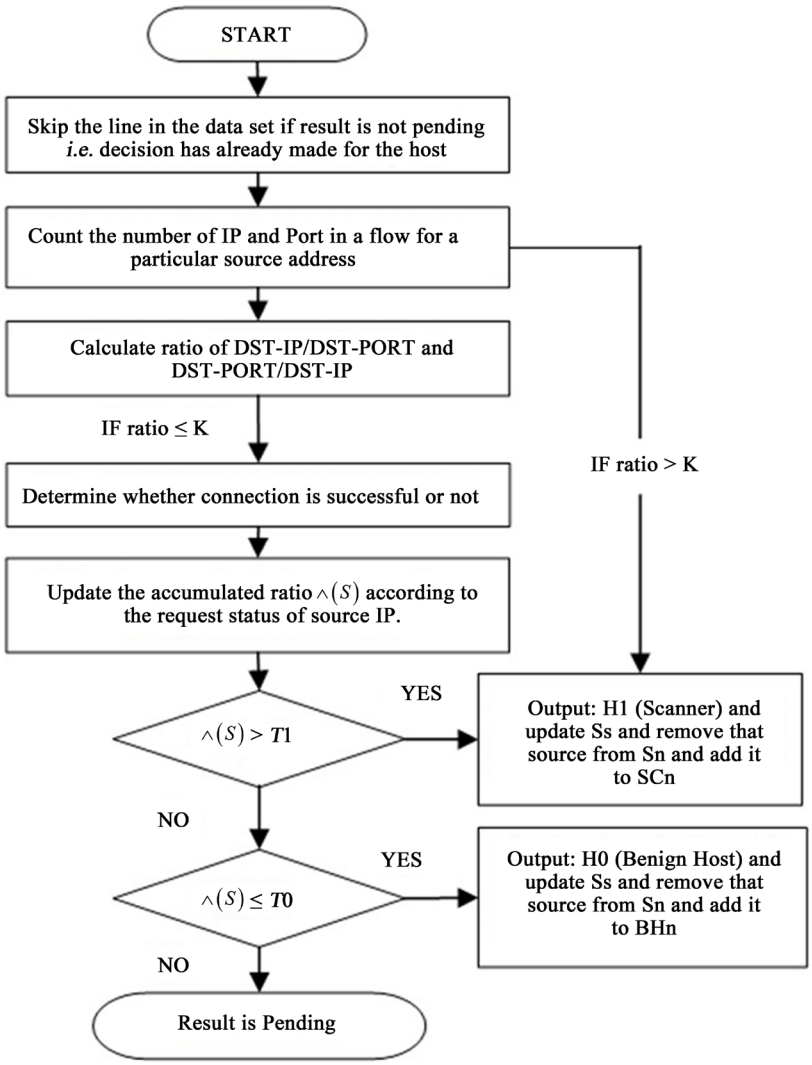

Figure 1. Flow graph of CPST algorithm.

$$\Pr\left[Y_i = 0 \middle| H0\right] = \theta_0, \quad \Pr\left[Y_i = 1 \middle| H0\right] = 1 - \theta_0$$
$$\Pr\left[Y_i = 0 \middle| H1\right] = \theta_1, \quad \Pr\left[Y_i = 1 \middle| H1\right] = 1 - \theta_1$$

where, $H0$ is the set of benign hosts and $H1$ is the set of scanners. The observation that a connection attempt is more likely to be a success from a benign source than a malicious one implies the condition:

$$\theta_0 > \theta_1$$

Whenever an event occurs, the sequential hypothesis testing updates the likelihood ratio (*flow*: *srcip*) is defined similarly to the TRWSYN and TAPS cases as follows:

$$\wedge(S) = \prod_{i=1}^{n} \frac{\Pr\left[Y_i \middle| H1\right]}{\Pr\left[Y_i \middle| H0\right]} \tag{1}$$

Y_i can take the value 1 or 0 depending upon the above mentioned conditions.

If $\wedge(S) > T1$ where $T1 \leq \dfrac{PD}{PF}$

*H*1 (*Scanner*) and update Ss (*Source*) and removed that source from Sn (*List of sources under test*) and add it in to SCn (*List of scanners*)

else If $\wedge(S) < T0$ where $T0 \geq \frac{1-PD}{1-PF}$

*H*0 (*Benign Host*) and update Ss (*Source*) and removed that source from Sn (*List of sources under test*) and add it in to BHn (*List of Benign Host*)

esle

Continue with more observations (results pending).

PF is probability of false positives and *PD* is the probability of detection for port scan detection [2].

5. Performance Evaluation

The performances of existing techniques are evaluated mainly on the basis of the detection rate and false positive rate metrics. The performance of CPST is evaluated and analysed with existing algorithms (TRW and TAPS) on the basis of these metrics.

The detection or success rate is defined as the ratio of total number of detected scanners in a dataset to the total number of scanners as shown in Equation (2). In the ideal case the detection rate is equal to 1.

$$\text{Detection Ratio} = \frac{\text{Total number of scanners detected}}{\text{Total number of scanners}} \tag{2}$$

The false positive rate is defined as the ratio of total number of false scanners detected to the total number of scanners present in dataset as shown in Equation (3). In other words, if a benign host is considered as a scanner then the result is called false positive. In the ideal case the false positive rate is equal to 0.

$$\text{False Positive Ratio} = \frac{\text{Total number of false scanners detected}}{\text{Total number of scanners}} \tag{3}$$

DARPA dataset [15] is used under sampled and non sampled data for evaluating the performance of scan detection algorithm CPST.

Figure 2 shows the effect of sampling on the success ratio for TRW, TAPS and CPST algorithms. It can be observed from the figure that in case of without sampling the success ratio is its maximum value. When the sampling interval increases, success ratio decreases, but rate of decreasing of success ratio of CPST is lower as compared to TRW and TAPS. In **Figure 2**, it is clearly shown that in term of success rate, the algorithm gives better performance for flow sampling as compared to packet sampling. In flow sampling, separate flow is created for every source, so that it is easy to identify the scanner and the benign host.

Figure 3 shows the effect of sampling on the false positive ratio for TRW, TAPS and CPST algorithms. It can be observed from the figure that at initial condition when there is no sampling, the false positive ratio is low but

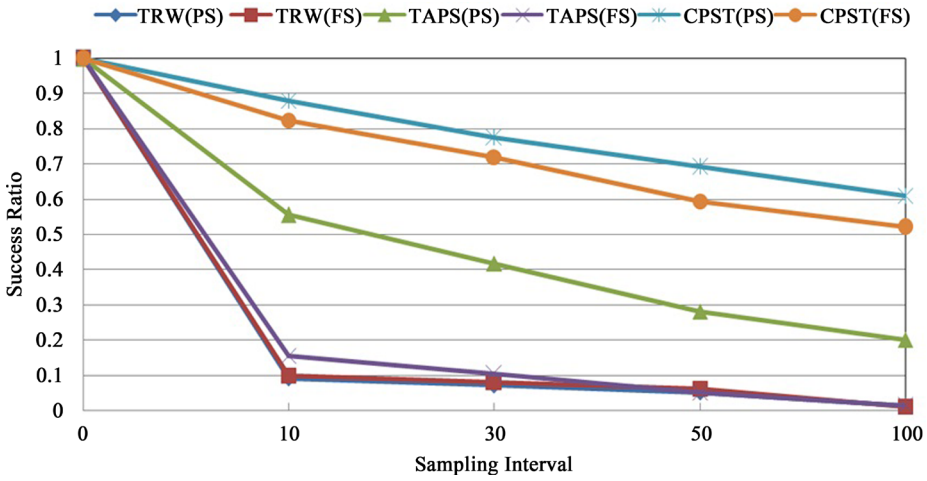

Figure 2. Success ratio vs. sampling interval for CPST, TRW and TAPS algorithms.

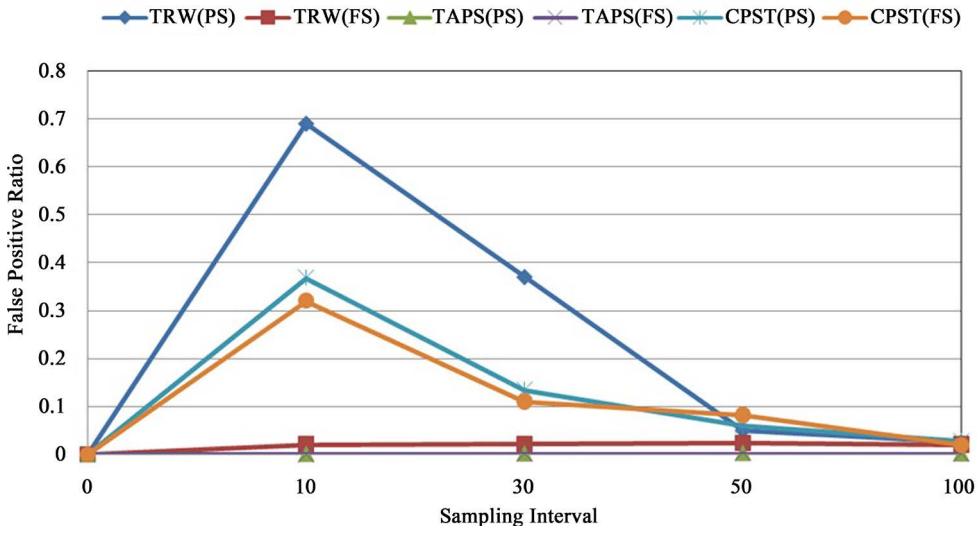

Figure 3. False positive ratio vs. sampling interval for CPST, TRW and TAPS algorithms.

not at its minimum value. It increases a while for low sampling rate, but when sampling rate increases the ratio monotonically decreases and it reaches nearly to zero for the higher sampling interval. But CPST exhibits the lower false positive rate as compared to TRW with packet sampling and slightly more as compared to TRW with flow sampling, and lower false positive rate as compared to TAPS with both sampling (packet sampling and flow sampling). In **Figure 3**, it is also clear that CPST algorithm performs better in flow sampling as compared to packet sampling and false positive rate in all the sampling approaches for higher sampling interval reaches near to zero.

6. Conclusion

In this paper, we present a two pass port scan detection technique called CPST which uses the fundamental concepts of connection status and pattern of connections for detecting the scanner or malicious host in the network. CPST is an effective technique. We compare the performance of this technique using DARPA data setting for packet sampling and flow sampling with existing TRW and TAPS scan detection techniques. The results show that CPST has better success and false positive ratio. It gives better detection ratio under high sampling rate as compared to the existing scan detection techniques, but CPST exhibits the lower false positive rate as compared to TRW with packet sampling and slightly more as compared to TRW with flow sampling and TAPS with both sampling (packet sampling and flow sampling). The proposed scheme exploits the access pattern and status of a particular source in a network flow. The success rate of the proposed scheme is about 61 % and the false positive rate is less than 2 % with higher sampling interval.

References

[1] Claise, B. (2004) Cisco Systems Net Flow Services Export Version, RFC 3954 (Informational). http://www.ietf.org/rfc/rfc3954.txt

[2] Jung, J., Paxson, V., Berger, A.W. and Balakrishnan, H. (2004) Fast Ports Can Detection Using Sequential Hypothesis Testing. *Proceeding of the IEEE Symposium on Security and Privacy*, Oakland, 9-12 May 2004, 221-225.

[3] Roesch, M. (1999) Snort-Lightweight Intrusion Detection for Networks. *Proceedings of* 13*th USENIX Conference on System Administration*, USENIX Association, Seattle, 7-12 November 1999, 229-238.

[4] Snort. http://www.snort.org

[5] Sridharan, A., Ye, T. and Bhattacharyya, S. (2006) Connectionless Port Scan Detection on the Backbone. 25*th IEEE International Performance*, *Computing, and Communications Conference* (*IPCCC* 2006), Mesa, 10-12 April 2006, 10-19. http://dx.doi.org/10.1109/.2006.1629454

[6] Spitzner, L. (2001) The Value of Honeypots, Part One: Definitions and Values of Honeypots. Security Focus. http://www.securityfocus.com/infocus/1492

[7] Lee, C.B., Roedel, C. and Silenok, E. (2003) Detection and Characterization of Port Scan Attacks. Department of Computer Science and Engineering, University of California, San Diego.

[8] Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2011) Surveying Port Scans and Their Detection Methodologies. *The Computer Journal*, **54**, 1565-1581. http://dx.doi.org/10.1093/comjnl/bxr035

[9] Mai, J., Sridharan, A., Chuah, C.N., Zang, S.M.H. and Ye, T. (2006) Impact of Packet Sampling on Ports Can Detection. *IEEE Journal on Selected Areas in Communications*, **24**, 2285-2298. http://dx.doi.org/10.1109/JSAC.2006.884027

[10] Mai, J., Chuah, C.N., Sridharan, A., Ye, T. and Zang, H. (2006) Is Sampled Data Sufficient for Anomaly Detection? *Proceedings of the* 6*th ACM SIGCOMM Conference on Internet Measurement*, Rio de Janeiro, New York, 165-176. http://dx.doi.org/10.1145/1177080.1177102

[11] IETF Packet Sampling (PSAMP) Working Group. http://www.ietf.org/html.charters/psamp-charter.html

[12] Zseby, T., Molina, M., Duffield, N., Niccolini, S. and Raspall, F. (2009) Sampling and Filtering Techniques for IP Packet Selection (RFC 5475). http://www.rfc-editor.org/rfc/rfc5475.txt

[13] Brownlee, N. (1997) Traffic Flow Measurement: Experiences with Ne Tra Met (RFC2123). http://tools.ietf.org/html/rfc2123

[14] Duffield, N. (2004) Sampling for Passive Internet Measurement: A Review. *Statistical Science*, **19**, 472-498. http://dx.doi.org/10.1214/088342304000000206

[15] Lippmann, R.P., Fried, D.J., Graf, I., Haines, J.W., Kendall, K.R., McClung, D., Weber, D., Webster, S.E., Wyschogrod, D., Cunningham, R.K. and Zissman, M.A. (2000) Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation. *Proceedings of the IEEE DARPA Information Survivability Conference and Exposition*, **2**, 12-26.