VOL. 11, NO. 5, MARCH 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

FORMULATING GENERALIZE MALWARE ATTACK PATTERN USING FEATURES SELECTION

Robiah Yusof, Mohd Zaki Mas'ud, Siti Rahayu Selamat, Mohd Faizal Abdollah, Shahrin Sahib and Rudy Fadhlee Mohd Dollah
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia
E-Mail: robiah@utem.edu.my

ABSTRACT

Malicious software or malware activity is increasingly threatened the network security as the malicious codes can be easily obtained and can be used as a weapon to gain illegal objectives. Hence, network traffic gathered from a control experiment are explored and features selection method is used to identify the features involved in formulating the malware attack pattern. This paper proposes generalize malware attack pattern in two perspectives which is attacker and victim using traditional worm. This research shall facilitate the authorities in detecting the malware intrusion activities in cyber space while protecting the Critical National Information Infrastructure (CNII) in the country. These generalized malware attack pattern can be extended into research areas in alert correlation and computer forensic investigation.

Keywords: malware, attack pattern, features selection, network traffic.

INTRODUCTION

Malware especially worm is difficult to detect as it has the capability to change it behavior of infecting others system and cause a difficulties for an antivirus to notice them. The variants of worm are often created to defeat the security tools, for example a worm can mutate to a different variants, sometimes in only one hour [1]. Thus, it is difficult for a security tool to detect the threat. As a result, the study on internet attack or intrusion is very crucial, especially in developing an effective security tool to defend the internet user from the attack threat.

According to MyCERT [2] in 2014, the incident statistics for intrusion attempt are the third highest which is at 1302 after fraud at 4477 and spam at 3650. This phenomenon has created a relative common task for security researchers to collect data related to internet threats and to help the researchers to investigate the intrusion behavior or attack pattern in order to find the root cause and effect of an intrusion in victim and attacker perspectives based on the intrusion's anatomy described in [3]. This paper proposes the general malware attack pattern using features selection method for identifying the attacker and victim by analyzing the network traffic.

RELATED WORK

Malware has become a serious threat as discussed by [5], [12] and [13] to the economy and also national security in recent years. Malware or malicious software is software that is residing in a system and it is intended to cause harm to the system. In general, according to [1], the malware can be categorized into three main types which are Trojan horse, virus and worm. This research is focusing on worm.

For the purpose of this paper, the researchers have scope the malware to traditional worms which are Blaster.A, Lovesan. T and Sasser.B. These types of worm are selected due to its persistent in internet as claimed by [6] and [7]. The network traffic which consists of these three malware's activity are generated in a control environment which consists of four phases: Network Environment Setup (NES), Attack Activation (AA), Log Collection (LC) and Log Analysis (LA). The NES consists of four components of attack steps proposed by [4]. The steps involved are Attacker Goal Identification, Network Configuration, Privilege Profile and Trust Setting, and Vulnerability and Exploit Permission. These components are implemented in NES. Then, the attack is activated in Attack Activation phase. These three variants are installed and activated on the selected attacker machine and the experiment runs for one hour without any human interruption in order to obtain the attack logs. In LC phase, log generated in tcpdump files are collected and then analyze in LA phase. The objective of LA phase is to identify the attack by observing the specific attack pattern generated by the variants in the network traffic log. This analysis is an input towards the development of the proposed generalize malware attack pattern.

Attack pattern

Attack pattern as discussed by [8], [9] and [14], is identified as one of the important component to protect a system from any potential attack. It is also consider as a systematic description of the attack goals and attack strategies for defending against attack. Consequently, according to [9], an attack pattern is a method to cause an exploit against software used by attackers. The importance of attack pattern is that it can show a clear view on how the attack is performed and impact caused by the attack. In addition, the attack pattern can lead to the prevention of the system.

Features selection

The features selection as discussed by [10], has been widely used in machine learning for security applications to improve generalization and computational efficiency. Various features exist in the network traffic and all of these features originally are based on the IP Packet Header and TCP Packet Header. Selecting unnecessary

features may cause computational issues and decrease the accuracy of detection [15].

The features are group into four categories as mentioned by [16] which are basic feature, content feature, time based traffic feature and host based traffic feature. These features are further group into two category which is basic feature and derive feature. The basic feature is categorized under packet header feature and the derive feature is breakdown into two categories, time based traffic feature and connection based traffic feature. The analysis and finding for the selected features and attack pattern shall be further discussed in next section.

Analysis and findings

In this research, three attack scenarios were generated for each Blaster.A, Sasser.B and Lovesan.T. The network traffic of these nine attack scenarios are further analyzed to enable the researcher to select the appropriate features to be included inside the attacker and victim pattern.

The general process of selecting malware features consist of three steps which are to identify features, select the malware features and generate malware attack pattern as depicted in Figure-1.

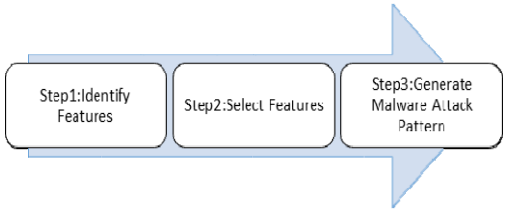

Figure-1. General process of selecting malware features.

In Figure-1, during Step1: identify features, several literature review on previous research were done and appropriate basic features in the tcpdump traffic are selected. In Step2, malware features is selected using Wireshark tools to analytically calculate the statistic of the packet captured. The features selected are then used to generate malware attack pattern in Step3. The step involved in the general process of selecting malware features in Figure-1 shall be further discussed in the next subsections.

Step 1: Identify features

The basic features which is also known as Packet Header Features is implemented in this research. Basic feature can be derived from packet header without inspecting the payload. The identified features in this basic feature are motivated by the research done by [17] which involving five tuples which are source IP address, destination IP address, source port, destination port and protocol. Information on IP address provides significant information to the identification of the attacker and victim [11]. These features are proposed to significantly identify the step being done by attacker and victim. The description of each features selected is depicted in Table-1.

Table-1. Description of features selected.

Features	Description
Source IP address	Determine the IP address of an attacker or victim
Destination IP address	Determine the IP address of a victim
Source port	Identify which port doing the attack.
Destination port	Identify which port being attacked
Protocol	Type of the protocol, e.g. tcp, udp, etc.

The identified features shall be further analyzed in Step2 to ensure that the five selected features can become an input to the proposed general malware attack pattern in Step3.

Step 2: Select features

In selecting features as identified in Step1, the percentage of feature is captured for all scenario in tcpdump traffic to support the features selection. In this step, only Blaster.A variant is covered. The same process were done for all scenario: Blaster.A (scenario 1-3), Sasser.B (scenario 4-6) and Lovesan.T (scenario 7-9). The captured features are selected using Wireshark tools to analytically calculate the statistic of the packet captured. Refer to Figure-2, the statistics displays that the highest percentage of IP addresses (192.168.2.2 and 192.168.2.10) was the attacker in this scenario. The rest of the IP addresses were victim.

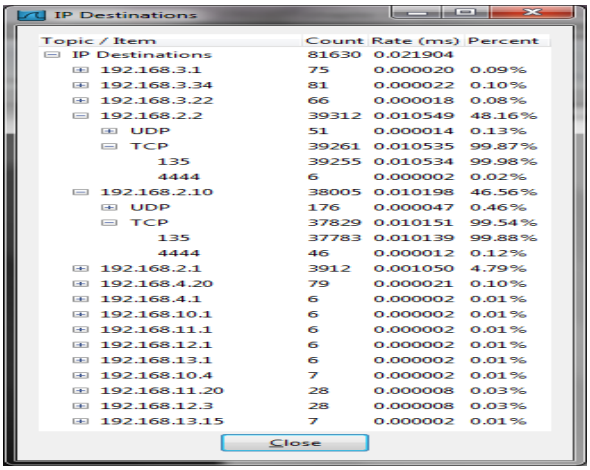

Figure-2. Sample of attacker features found on Scenario 1 of Blaster.A

The highest percentage in Figure-2, referred to IP address 192.168.2.2 with 48.16%, and 192.168.2.10 with 46.56%. These IP addresses used TCP and UDP protocol. TCP protocol use port 135 and port 4444 to do connection on victims. TCP protocol gets the highest percentage with 99.87% and 99.54 %, while UDP gets 0.13% and 0.46%.

The statistics in Figure-3, shows the lowest percentage was victim which referred to 192.168.11.20 and 192.168.12.3 with 0.03%.

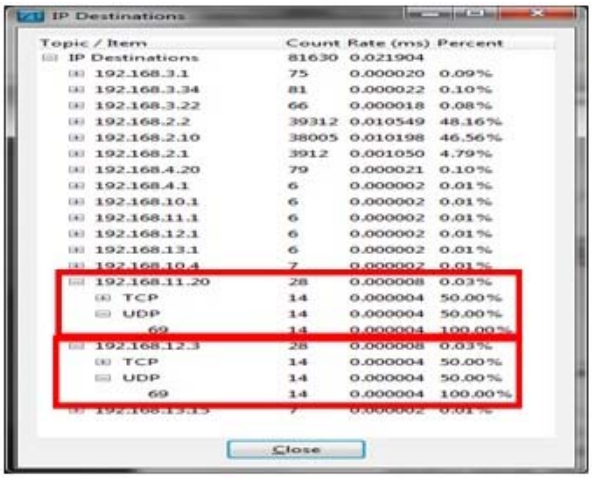

Figure-3. Sample of victim features found on Scenario 1 of Blaster.A

Both IP addresses has shown 50% on UDP protocol and 100% on port 69. Similar process were done on Scenario 2 until Scenario 9. The summary of the feature's percentage gathered in Step2 is depicted in Table-2.

Table-2. Summary of feature's percentage for all scenario.

Features	Scenario 1	Scenario2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
UDP protocol	0.13 %	0.58 %	0.58 %	None	None	None	27.57%	18.28%	37.62%
UDP port 69	100 %	100 %	100 %	None	None	None	None	None	None
TCP protocol	99.87 %	99.42 %	99.42 %	99.71%	99.65%	99.68%	72.43%	81.72%	62.38%
TCP port 135/445	99.98 %	99.52 %	99.52 %	99.93%	99.96%	99.98%	93.45%	98.00%	91.63%
TCP port 4444/9996	0.02 %	0.48 %	0.48%	0.01%	0.01%	0.01%	2.24%	2.00%	4.06%
IP address	48.16%	90.63%	90.63%	46.34%	41.64%	83.44%	44.25%	76.42%	30.14%

The summary of feature's percentage in Table-2 shows that the protocol and port (source port and destination port) is valid to be selected since the percentage of UDP and TCP protocol are nearly 100% occurrence. IP address are also valid since it will enable us to identify the perspective which is either attacker or victim.

Step 3: Generate malware attack pattern

In step3, these selected features are then used to construct the attack pattern. To suit this research which involves raw traffic data, the attack pattern is mapped to the modified basic worm attack model as motivated by [18]. It consists of three activity namely scan, exploit and impact/effect. The three malware's attack pattern obtain in this analysis are then used as the primary guideline in developing the proposed general malware attack pattern. The attack pattern of Blaster.A, Lovesan.T and Sasser.B on attacker and victim perspective are further discussed in the next subsections.

- **Blaster. A attack pattern**

In perspective of attacker for Blaster.A, the attacker establishes connection from any random port to victim on port 135 with TCP protocol. Then it establishes connection on port 4444 with TCP protocol. Next, port 69 is opened using UDP protocol to enable it to transfer the file to victim. The summary of the Blaster.A's attacker attack pattern is depicted in Figure-4.

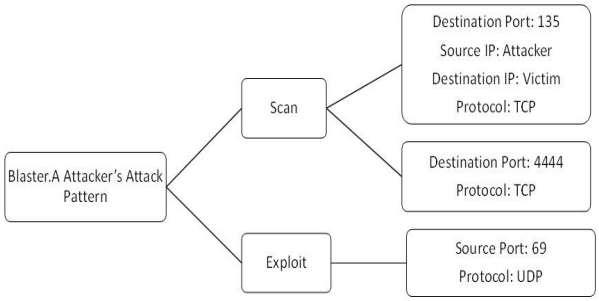

Figure-4. Blaster.A's attacker attack pattern.

In Figure-4, the activity involves in attacker are only scan and exploit. Meanwhile, in perspective of victim, port 135, port 4444 are opened, and then it requests to open port 69 on attacker to read the file of msblast.exe with 13 blocks of data packets. The summary of the victim attack pattern is shown in Figure-5.

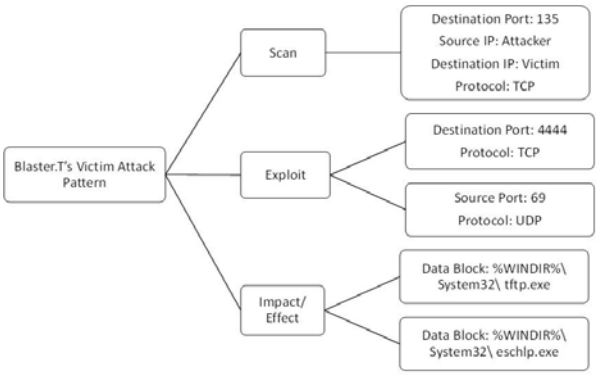

Figure-5. Blaster.A's victim attack pattern.

In Figure-5, the activity involves in victim attack pattern are scan, exploit and impact/effect.

- **Lovesan. T attack pattern**

In scanning activities, which is usually used port 135 TCP, the attacker will do connection with sent SYN to establish connection on victim using TCP protocol. Meanwhile in exploiting activities, attacker use port 135,

4444, 69 and 3xxx to exploit the system-level command shell on its victims. It will send the eschlp.exe file to the victim using TFTP protocol. The summary of the Lovesan.T's attacker attack pattern is illustrated in Figure-6.

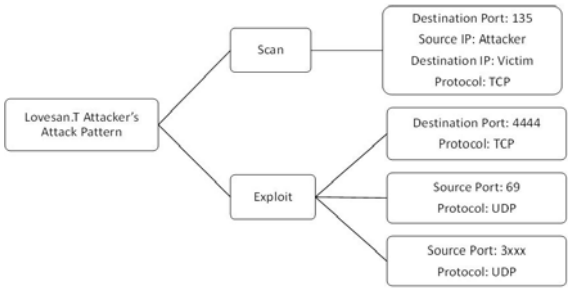

Figure-6. Lovesan.T's attacker attack pattern.

In Figure- 6, the activity involves in attacker are only scan and exploit, which are similar to Blaster.A's attacker attack pattern. On victim perspective, the worm does the scanning and exploiting activities. In scanning activities, it usually used port 135 and TCP protocol. Meanwhile, the worm will infiltrate the vulnerable port 135, 4444, 69 and 3xxx to exploit the victim. The victim had been infected will receive the data eschlp.exe and tftp.exe from attacker. This data had been sent using TFTP protocol on port 69. The summary of Lovesan.T's victim pattern is depicted in Figure-7.

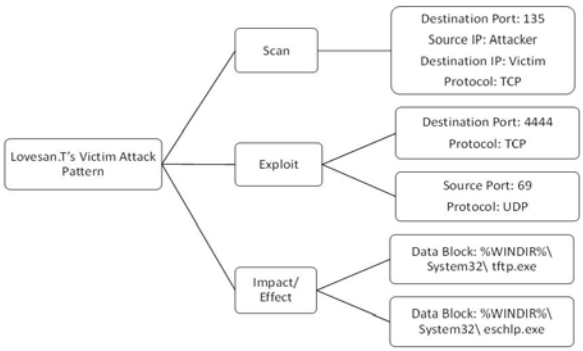

Figure-7. Lovesan.T's victim attack pattern.

In Figure-7, the activity involves in Lovesan.T's victim attack pattern are similar to Blaster.A's victim attack pattern which are scan, exploit and impact/effect.

- **Sasser.B attack pattern**

In attacker's perspective, attacker will scan the targeted machines using destination port 445 (microsoft-ds) using TCP protocol. Once the connection is established between attacker and victim, attacker will sends SMB packets using SMB protocol and the victim will response back. After that, attacker will use destination port 9996 (palace-5) to begin exploiting the victim and use source port 5554 (sgi-esphttp). Figure-8 illustrated the Sasser.B's attacker's attack pattern.

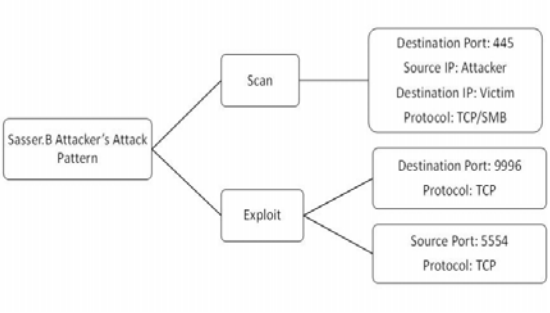

Figure-8. Sasser.B's attacker attack pattern.

In Figure-8, the activity involves in attacker are also similar to Blaster.A and Lovesan.T attacker pattern which are only scan and exploit. For victim's perspective, victim will communicate with attacker using port 445 (microsoft-ds) as a source port. The victims will response to the SMB packets that attacker request. Next, by using port 9996 (palace-5) attacker will instruct the victim to open port 5554 (sgi-esphttp) to transfer the attack. The traces that can be found in victim's traffic is *_up.exe. Figure-9 illustrated the victim's attack pattern.

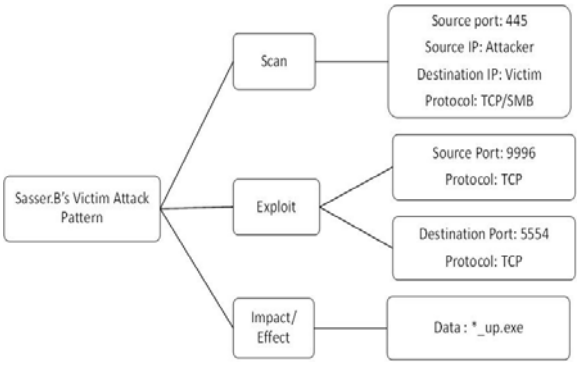

Figure-9. Sasser.B's victim attack pattern.

In Figure-9, the activity involves in Sasser.B's victim attack pattern are similar to Blaster.A and Lovesan.T's victim attack pattern which are scan, exploit and impact/effect In this analysis, the researchers have identified the features in the attacker and the victim pattern. These findings are further used to construct the proposed general malware attack pattern.

PROPOSED GENERAL MALWARE ATTACK PATTERN

This research proposed the general malware attack pattern based on victim and attacker which will be described in following section. The three malware attack pattern for attacker and victim discussed previously is further analysed. The finding of the analysis were summarized in Table-3 and Table-4.

Table-3. Summary on general's malware attacker attack pattern (features found=√, features not found=x).

Attack Steps	Features	Blaster.A	Lovesan.T	Sasser.B	General malware Attacker Attack Pattern
Scan	Destination Port	√	√	√	Destination Port
	Source IP	√	√	√	Source IP
	Destination IP	√	√	√	Destination IP
	Protocol	√	√	√	Protocol
Exploit	Source Port	√	√	√	Source Port
	Protocol	√	√	√	Protocol
	Destination Port	x	√	√	Destination Port
	Protocol	x	√	√	Protocol

Based on Table-3, in attack steps, during scan activity, data related to this activity can be found in Destination Port, Source IP Address, Destination IP Address and Protocol. Meanwhile in exploit activity, three features related to these activities which are Source Port, Destination Port and Protocol. These features are represented in diagram as depicted in Figure-10.

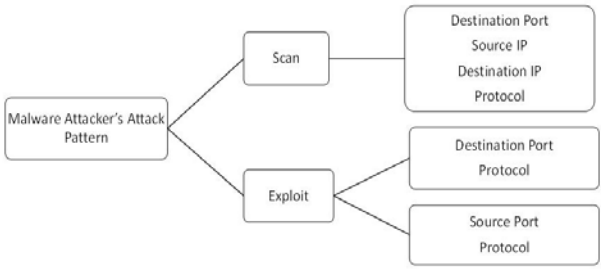

Figure-10. General Malware's attacker attack pattern.

Referring to Table-4, during scan activity, similar features found in attacker pattern except for Source Port. Subsequently, in exploit activity, all features found are also similar to attacker pattern. The major difference between attacker and victim is the impact/effect activity where data block is identified only in victim attack pattern.

Table-4. Summary on general's malware victim attack pattern (features found=√, features not found=x).

Attack Steps	Features	Blaster.A	Lovesan.T	Sasser.B	General malware Victim Attack Pattern
Scan	Destination Port	v	v	x	Destination Port
	Source IP	v	v	v	Source IP
	Destination IP	v	v	v	Destination IP
	Protocol	v	v	v	Protocol
	Source Port	x	x	v	Source Port
Exploit	Source Port	v	v	v	Source Port
	Protocol	v	v	v	Protocol
	Destination Port	x	v	v	Destination Port
Impact/Effect	Data Block	v	v	v	Data Block

The features found in Table-4 are then represented in diagram as depicted in Figure-11.

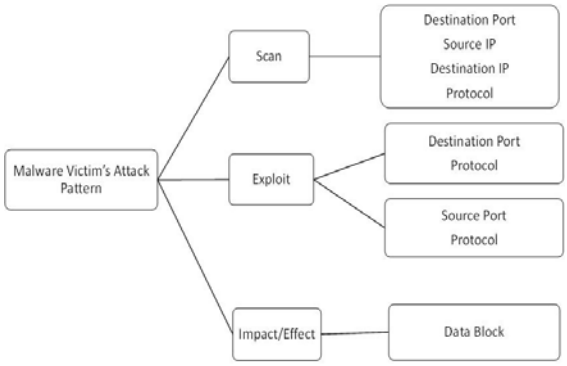

Figure-11. General Malware's victim attack pattern

In conclusion, both general malware's attacker and victim attack pattern as illustrated in Figure-10 and Figure-11 are proposed in this research. This finding could assist other researcher to identify the true victim or true attacker in an incident.

CONCLUSION AND FUTURE WORKS

In this research, the network traffic generated by Blaster. A, Sasser. B and Lovesan. T are further analyzed to identify the features to be selected. The attack pattern related to attacker and victim for each variant is constructed using the five selected features which then become an input to the proposed general malware's attacker and victim attack pattern. Both attack pattern can be further extended used in alert correlation and computer forensic investigation.

ACKNOWLEDGEMENTS

We would like to thank to Universiti Teknikal Malaysia Melaka for the Short Grant funding (PJP/2013/FTMK(9D)/S01167) in this research project.

REFERENCES

[1] Karresand, M. 2003. A proposed taxonomy of software weapons. FOI-Swedish Defence Research Agency.

[2] MyCERT 2014 Summary Report. Retrieved 16 April 2015 from http://www.mycert.org.my/statistics/2014.php

[3] McHugh. J. 2001. Intrusion and intrusion detection. International Journal of Information Security. pp. 14-35.

[4] Zhang, S., Li, J., Chen, X, Fan, L. 2008. Building network attack graph for alert causal correlation. Elsevier. pp. 188 – 196.

[5] Alkaabi, A., Mohay, G. M., McCullagh, A. J., & Chantler, A. N. 2010. Dealing with the Problem of Cybercrime. Proceedings of the 2nd International ICST

Conference on Digital Forensics & Cyber Crime. pp. 1-18

[6] Bailey. M., Cooke. E., Jahanian. F., Watson. D. Nazario. J. 2005. The Blaster Worm: Then and Now, IEEE Computer Society.

[7] Crandall, J.R., Ensafi, R., Forrest, S., Ladau, J. Shebaro, B. 2008. The Ecology of Malware. ACM.

[8] P. Moore, A., J. Ellison, R., C. Linger, R. 2001. Attack Modeling for Information Security and Survivability. (No. CMU/SEI-2001-TN-001): Pittsburgh, Pennsylvania: Software Engineering Institute, Carnegie Mellon University.

[9] Barnum, S., Sethi, A. 2006. Introduction to Attack Patterns. [Electronic Version]. Retrieved 18 April 2012.

[10] Xiao. H., Biggio. B., Brown. G., Fumera. G., Eckert. C., Roli, F. 2015. Is Feature Selection Secure against Training Data Poisoning? Proceedings of the 32nd International Conference on Machine Learning, Lille,France. JMLR: W&CP. Vol. 37.

[11] Gavrilis, D & Dermatas. 2005. E. Real Time Detection of Distributed Denial of Service Attack using RBF Network and Statistical Feature. International Journal of Computer Network. Vol. 48. pp 235-245

[12] Mathur. K, Hiranwal. S. 2013. A survey on techniques in detection and analyzing malware executables. International Journal of Advanced Research in Computer Science and Software Engineering. Vol 3. pp. 422-428.

[13] Uppal. D., Mehra. V., Verma. V. 2014. Basic Survey on Malware Analysis, Tools and Techniques. International Journal on Computational Sciences & Applications. Vol 4(1). pp. 103-112.

[14] Hoglund. G., McGraw. G. 2004. Exploiting Software: How to Break Code. Boston, Massachussets: Addison-Wesley/Pearson.

[15] Sung. A.H., Mukkamala. S. 2005. The Feature Selection and Intrusion Detection Problems. Advances in Computer Science – ASIAN 2004. Higher-Level Decision Making. Lecture Notes in Computer Science. Vol. 3321. pp. 468-482.

[16] Kayacik. H.G., Heywood. A. N. Z, M.I. Heywood. 2005. Selecting Features for Intrusion Detection: A Feature Relevance on KDD 99 Intrusion Datasets. Proceeding of the 3rd Annual Conference on Privacy and Trust. St. Andrew, NB, Canada.

[17] Avinash. S., Ye. T., Bhattacharrya. S. 2007. Connectionless Portscan Detection on the Backbone. Proceeding of Malware Workshop Conjunction with IPCCC.

[18] Liu. Z., Wang. C, Chen. S. 2008. Correlating Multi-Step Attack and Constructing Attack Scenarios Based on Attack Pattern Modeling. IEEE Computer Society. pp. 214-219.