Mining Association Rules Consisting of Download Servers from Distributed Honeypot Observation

Masayuki Ohrui
Hiroaki Kikuchi
Department of Information Science and Technology,
Graduate School of Engineering, Tokai University
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
yama@cs.dm.u-tokai.ac.jp
kikn@cs.dm.u-tokai.ac.jp

Masato Terada
Hitachi Incident Response Team (HIRT),
Hitachi, Ltd.
890 Kashimada, Kawasaki, Kanagawa, 212-8567, Japan
masato.terada.rd@hitachi.com

***Abstract*—This paper aims to find interested association rules, known as data mining technique, out of the dataset of downloading logs by focusing on the coordinated activity among downloading servers. The result of the analysis shows the association rules of the downloading servers and that of the malwares.**

***Keywords*-Association Rules; Malware; Botnets; Coordinated Attack; Sequential Pattern;**

I. INTRODUCTION

The botnet has a feature that cooperative attacks for multiple servers making a victim infected by a set of malwares [1]. For example, Table I shows sequential infections observed the Cyber Clean Center (CCC) DATAset 2009, the captured packets data by 94 honeypots [2] in which a host is infected by three malwares, `PE_VIRUT.AV`, `TROJ_BUZUS.AGB` and `WORM_SWTYMLAI.CD` as scheduled in the same way. Although the these servers are assigned different IP addresses, it turns out to be a correlation in the malware infections. In this paper, we call the multiple infections made by several servers *the botnets coordinated attacks*.

The discovery of the coordinated attacks is, however, very difficult because it is necessary to investigate huge amount of captured data for looking for common sequential patterns. For example, the CCC DATAset 2009 of downloading logs [2][3] contains more than 30,000 packets even in top 4 IP addresses shown in Figure 1.

There are many difficulties in identifying coordinated attacks. (1) The number of infections depends on IP address of honeypot. (2) The number of infections is not stable. There is a period when no malware is observed at all, before/after many infections happen. (3) The complexity of generating is high. Since 1,335 kinds of malware will be observed in one year [2], there are possible ${}_{1,335}C_3 = 395,654,395$ combination of three kinds of distinct malware chosen out of total number of slots of 365 day $\times$ 94 honeypots $\times$ 24 hour $\times$ 3 slots/hour $= 2,470,320$, it is not feasible to examine [1]. Therefore, a coordinated attacks is hard to be discovered.

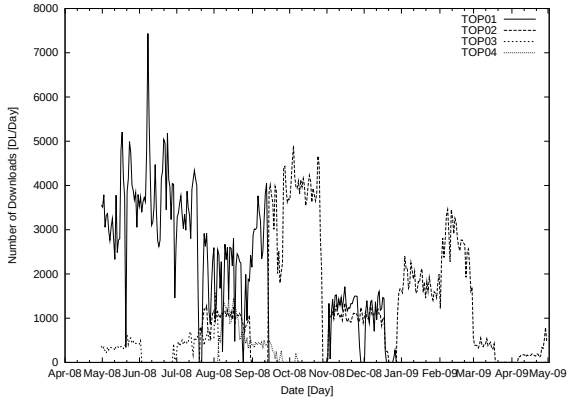


Figure 1. Number of downloadings observed in one year

Then, in this paper, we try to introduce an association analysis, Apriori, well-known the data mining technique from large-scale data, in order to extract association rules of coordinated attacks.

Apriori algorithm was proposed by Agrawal et al. [4]. By pruning useless combinations with the minimum support and confidence, *the association rules* can be efficiently extracted. In this paper, we apply Apriori Program [5] to the alert log, in order for extracting the most frequently attack patterns in a year.

The paper is organized as follows. Firstly, the algorithm of the association analysis is explained. An example the association analysis is presented, and how to detect the association rules is shown. Secondly, we explain experimental data and experiment objectives. Thirdly, we show the experimental result. The attack patterns has combinations of malwares, and coordination between download servers. We verify dependency of honeypots and observation time,

[1]The definition of the slot is described later by Section III-A.

Table I
SAMPLE OF COORDINATED ATTACKS OF MALWARE

Time	Source IP address	Dst Port	Protocol	MW
0:02:11	124.86.***.111	47556	TCP	PE_VIRUT.AV
0:03:48	67.215.*.206	80	TCP	TROJ_BUZUS.AGB
0:03:48	72.10.***.195	80	TCP	WORM_SWTYMLAI.CD
0:36:46	124.86.**.109	33258	TCP	PE_VIRUT.AV
0:36:52	72.10.***.195	80	TCP	WORM_SWTYMLAI.CD
0:36:52	67.215.*.206	80	TCP	TROJ_BUZUS.AGB
0:46:56	124.86.**.109	33258	TCP	PE_VIRUT.AV
0:48:52	67.215.*.206	80	TCP	TROJ_BUZUS.AGB
0:48:52	72.10.***.195	80	TCP	WORM_SWTYMLAI.CD

Table II
HIGH RANK 10 OF SOURCE IP ADDRESS FROM PACKETS CAPTURED DATA

Rank	Source IP address	Infections	Average	MW	Honeypot ID
1	72.10.***.74	462246	3884.4	119	91
2	72.10.***.195	399562	8324.2	48	92
3	85.114.***.2	33283	1147.7	29	82
4	85.114.***.207	32202	870.3	37	78
5	67.215.*.206	26780	3825.7	7	59
6	211.95.**.6	19641	198.4	99	85
7	72.10.***.26	14951	287.5	52	82
8	92.48.**.63	11699	117.0	100	69
9	67.18.***.250	10060	76.8	131	68
10	72.8.***.164	5099	127.5	40	81

respectively. Finally, we conclude this paper.

II. BUILDING BLOCKS

Apriori is an algorithm of data mining for extracting association rules of the form

$$X(antecedent) \Rightarrow Y(consequent)$$

from a given set.

A support is a probability of set of an association rule ($X \Rightarrow Y$) to be shown out of all transactions N, which is defined as

$$\textit{Supp}(X \Rightarrow Y) = \frac{|X \cap Y|}{N}$$

A confidence is a probability of the rule is satisfied, namely, a chance of Y is true if X is true. The definition is given by

$$\textit{Conf}(X \Rightarrow Y) = \frac{|X \cap Y|}{X}$$

Apriori is a well-known algorithm for association rule discovery due to Agrawal et al. [4]. It allows to efficiently discover all useful association rules by excluding the rules those support and confidence are smaller than giving minimum support and confidence. With the minimum support, we can squeeze many useless rules to be examined.

For example, the association rule $B, C \Rightarrow E$ in Table III has support and confidence as

$$\textit{Supp}(B, C \Rightarrow E) = 2/4 = 0.5,$$

$$\textit{Conf}(B, C \Rightarrow E) = 2/2 = 1.$$

Table III
EXAMPLE OF TRANSACTION

TID	A	B	C	D	E
1	1		1	1	
2		1	1		1
3	1	1	1		1
4		1			1

Thus, the rule of $B, C \Rightarrow E$ is support 50% and confidence 100%. In other word, this rule shows with probability of 50%, and $B, C \Rightarrow E$ appears with probability of 100% when B and C appear.

III. EXAMINATION METHOD

A. Experimental Data

In order to verify efficiency of our proposed method, we apply the Apriori algorithm to the experimental data, CCC DATAset 2009. More than 90 independent honeypots have observed malware traffic at the Japanese tier-1 backbone under coordination of the Cyber Clean Center (CCC). CCC DATAset 2009 consists of the access log of attack for a year during May 1, 2008 until April 30, 2009. 94 honeypots are periodically rebooted every 20 minutes. We call the time interval time slot in this paper. Observation in 2 days gives 145 time slots. A transaction consists of malware names that are downloaded in a time slot. Similarly, malware downloading logs is divided in terms of time slots. In out experiment, we use vulnerable Windows XP as honeypot.

B. Experiment Objectives

Our experiments are shown as follows.

1) Association rules of malware names extracted from the captured packets data
2) Association rules of downloading servers extracted from the captured packets data
3) Dependency on location of honeypots in extracting rules from malware downloading logs
4) Dependency of observation time in extracting rules from malware downloading logs

The malware names are identified by commercial antivirus signature. In the 1st and 2nd experiments, we investigate relativity strength of the coordinated attacks from two viewpoints of malware name and IP address. We compare the result of manual analysis in Table I with one in the automated method. In the 3rd experiment, we aim to extract general association rule in the sense that a common patterns are observed on a different honeypots. The purpose of 4th experiment is to examine whether association rules varies in long period.

IV. Experimental Results

A. Association Rules of Malwares

Malicious coordinated servers send same kind of malware to a single target host. Table IV shows an instance of sequences of malware observed in a each time slot, indicating 58 infected slots out of 145 slots. The most frequent infection observes 11 distinct names of malware at a single slot.

We apply the Apriori algorithm to dataset of malware shown in Table IV and successfully discover significant association rules of malware in Table V. This result shows all association rules with support more than 10% and confidence more than 80%. A support is a percentage of the slots that the association rule appears for 145 slots. A confidence is defined as a conditional probability of malware of consequent of the rule to be observed given malware of antecedent of the rule.

We are interested in whether the rule of coordinated attacks pattern `PE_VIRUT.AV` $\Rightarrow$ `TROJ_BUZUS.AGB`, `WORM_SWTYMLAI.CD` marked as Table V is automatically detected or not. Unfortunately, Table V does not contain the exactly same rules to the rule, but has many similar association rules to No.5 `PE_VIRUT.AV`, `TROJ_BUZUS.AGB` $\Rightarrow$ `WORM_SWTYMLAI.CD` and rule No.6 `PE_VIRUT.AV`, `WORM_SWTYMLAI.CD` $\Rightarrow$ `TROJ_BUZUS.AGB`. From the observation of Table V, we found significant correlation between `TROJ_BUZUS.AGB` and `WORM_SWTYMLAI.CD` in Rule No.1, 2, 3.

B. Association Rules of Downloading Servers (IP address)

We investigate the correlation of downloading servers using the Apriori algorithm. Table VI shows the association rules of IP address of downloading servers. We specify the minimum support of 10% and confidence of 50%.

The input of out analysis is data that consists of IP address from which the list of malware in Table IV are downloaded. Note that the malware name does not mean one-to-one corresponding IP Address, e.g. `PE_VIRUT.AV` is downloaded from 16 in rule distinct IP addresses.

Note that some IP addresses are assigned for particular malware. For example, addresses 114.145.**.166 and 122.18.***.123, used 12 and 21 times in rule No.1 and 2, respectively, are mainly used for `PE_VIRUT.AV`.

While, rules No.3 and 4 are dedicated to downloading `TROJ_BUZUS.AGB` and `WORM_SWTYMLAI.CD`. However, we fail to detect the first downloading server for `PE_VIRUT.AV` in Table I.

C. Dependency on Honeypot

We are interested in whether the extracted association rules depends are honeypots, or not. We show the numbers of honeypots that have observed the top 10 association rules of malware in Table VII.

We investigate 94 honeypot IDs in March 13, 2009. For example, 36 out of 94 honeypots have observed common rule No.1, suppressing the difference of support and confidence.

Figure 2 shows the number of association rules by number of distinct honeypots that observe the rule, denoted by k. The vertical axis shows number of distinct association rules, $N(k)$, where k honeypots detect the rule. We note that using the most common association rules using `TROJ_BUZUS.AGB` and `WORM_SWTYMLAI.CD` are typically observed by single honeypot. On the other-hand, the rule consisting of `TROJ_BUZUS.AGB` and `WORM_SWTYMLAI.CD` has been widely observed by more than 1/3 honeypots. The widely observed hosts can be considered as ones being used for coordinated attacks. We also note that only specific malware is used to coordinate attack.

D. Lifecycle of Association Rule of Malware

A duration of coordinated attacks is not so long. Table VIII shows the number of association rule observed in Honey003 with the minimum confidence more than 80%. Association rules are observed 365 days, Table VIII shows that `PE_VIRUT.AV` is observed throughout year.

The top 3 frequent malware names, `PE_VIRUT.AV (PE)`, `TROJ_BUZUS.AGB (TROJ)`, `WORM_SWTYMLAI.CD (WORM)`, are detected as correlated in the association rules as follows

1) `PE_VIRUT.AV WORM_SWTYMLAI.CD` $\Rightarrow$ `TSPY_KOLABC.CH`,
2) `TROJ_BUZUS.AGB` $\Rightarrow$ `WORM_SWTYMLAI.CD`,
3) `TSPY_KOLABC.CH` $\Rightarrow$ `WORM_SWTYMLAI.CD`.

Table IV
SEQUENCES OF MALWARES OBSERVED IN A TIME SLOT

Time Slot	Sequence of Malware			
0	PE_VIRUT.AV	TROJ_BUZUS.AGB	WORM_SWTYMLAI.CD	
2	WORM_ALLAPLE.IK	PE_VIRUT.AV	WORM_SWTYMLAI.CD	TROJ_BUZUS.AGB
3	PE_VIRUT.AV	TROJ_BUZUS.AGB	WORM_SWTYMLAI.CD	PE_VIRUT.AV
14	BKDR_POEBOT.GN	TROJ_BUZUS.AGB	WORM_SWTYMLAI.CD	
15	BKDR_MYBOT.AH	PE_VIRUT.AV		
⋮				
141	PE_BOBAX.AK	WORM_SWTYMLAI.CD	WORM_AUTORUN.CZU	WORM_IRCBOT.CHZ

Table V
ASSOCIATION RULES OF MALWARE INFECTIONS

Rule.	Antecedent		Consequent	Supp	Conf
1	TROJ_BUZUS.AGB	⇒	WORM_SWTYMLAI.CD	41.4	100
2	WORM_SWTYMLAI.CD	⇒	TROJ_BUZUS.AGB	46.6	88.9
3	TROJ_BUZUS.AGB, BKDR_POEBOT.GN	⇒	WORM_SWTYMLAI.CD	10.3	100
4	WORM_SWTYMLAI.CD, BKDR_POEBOT.GN	⇒	TROJ_BUZUS.AGB	10.3	100
5	PE_VIRUT.AV, TROJ_BUZUS.AGB	⇒	WORM_SWTYMLAI.CD	29.3	100
6	PE_VIRUT.AV, WORM_SWTYMLAI.CD	⇒	TROJ_BUZUS.AGB	29.3	100
*	PE_VIRUT.AV ⇒ WORM_SWTYMLAI.CD		TROJ_BUZUS.AGB	N/A	N/A

Table VI
ASSOCIATION RULES OF DOWNLOADING SERVERS TO THE PACKETS CAPTURED DATA

No.	Antecedent		Consequent	Supp	Conf	Corresponding Malwares	Corresponding Rank
1	114.145.**.166	⇒	122.18.***.123	12.1	85.7	PE ⇒ PE	
2	122.18.***.123	⇒	114.145.**.166	15.5	66.7	PE ⇒ PE	
3	67.215.*.206	⇒	72.10.***.195	46.6	100	TROJ ⇒ WORM	TOP5 ⇒ TOP2
4	72.10.***.195	⇒	67.215.*.206	46.6	100	WORM ⇒ TROJ	TOP2 ⇒ TOP5

Table VII
NUMBER OF HONEYPOTS THAT HAVE THE ASSOCIATION RULES OF MALWARE IN MARCH 13, 2009

No.	Antecedent		Consequent	Honeypots
1	TROJ_BUZUS.AGB	⇒	WORM_SWTYMLAI.CD	36
2	WORM_SWTYMLAI.CD	⇒	TROJ_BUZUS.AGB	36
3	TROJ_BUZUS.AGB, BKDR_VANBOT.AHH	⇒	WORM_SWTYMLAI.CD	12
4	WORM_SWTYMLAI.CD, BKDR_VANBOT.AHH	⇒	TROJ_BUZUS.AGB	12
5	TROJ_DLOADR.CBK	⇒	UNKNOWN	8
6	TROJ_BUZUS.AGB, PE_VIRUT.AV	⇒	WORM_SWTYMLAI.CD	7
7	WORM_SWTYMLAI.CD, PE_VIRUT.AV	⇒	TROJ_BUZUS.AGB	7
8	PE_VIRUT.AV, TROJ_BUZUS.AGB	⇒	WORM_SWTYMLAI.CD	6
9	TROJ_AGENT.ANDF	⇒	UNKNOWN	6
10	PE_VIRUT.AV, WORM_SWTYMLAI.CD	⇒	TROJ_BUZUS.AGB	6

We see the first Rule is related with WORM_SWTYMLAI.CD, which is the most frequent malware over a long period. We note that Rule 1 and 3 contain TSPY_KOLABC.CH, which was not found in 2-days analysis in Section IV-A.

Figure 3 shows the distribution of activities of the top 3 association rules except for UNKNOWN, defined by

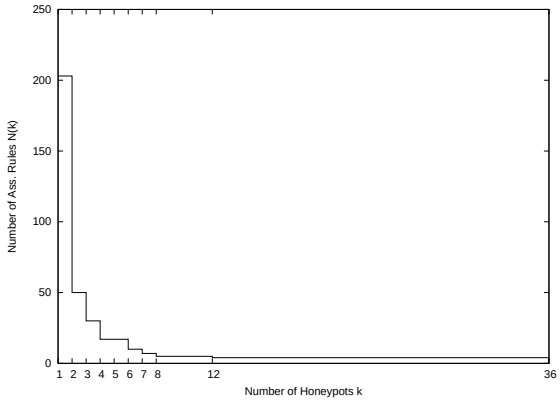
1) BKDR_VANBOT.HI ⇒ BKDR_SDBOT.BU,
2) BKDR_POEBOT.AHP ⇒ TROJ_QHOST.WT,
3) TSPY_KOLABC.CH ⇒ WORM_SWTYMLAI.CD.

The period of the above rules is short. The reason of short period is that the short period of coordinated attacks is hard to be detected. Moreover, coordinated pattern is constantly renewed every time new malware is developed.

V. CONCLUSIONS

In this paper, we proposed an automated method to detect the association rule of malware for coordinated attacks. We showed that our proposed method can extract all coordinate attacks correctly. The result of our experiment shows strong correlation between PE_VIRUT.AV, TROJ_BUZUS.AGB and WORM_SWTYMLAI.CD.

The widely observed rules are likely to be coordinated attacks. As a result of our observation in a long term, the



tb]

Figure 2. Number of association rules by number of honeypot observing rules

Table VIII
NUMBER OF RULES CONTAINING MALWARE PE, TROJ, WORM

	PE	TROJ	WORM
2008/05	31	0	0
2008/06	76	0	0
2008/07	111	0	0
2008/08	5	0	0
2008/09	8	0	0
2008/10	44	0	0
2008/11	27	0	0
2008/12	35	0	0
2009/01	135	0	0
2009/02	125	0	226
2009/03	79	53	74
2009/04	30	0	0

duration of coordinated attacks is very short, mostly within a month. It is hard to detect significant association rules because PE_VIRUT.AV is distributed by multiple servers.

Our future works include considering the relation of time of the malwares. We plan to apply yet another algorithm of data mining in [6].

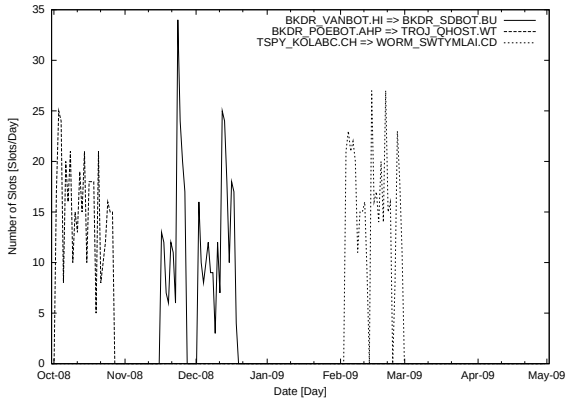


Figure 3. Distribution of activities of top 3 rules, excepting label UNKNOWN

ACKNOWLEDGEMENT

We thank Mr. Masashi Fujiwara, Mr. Hiroshi Nakakoji and Mr. Tetsuro Kito at Hitachi Ltd., Mr. Shunji Mastuo at Tokai University, and Mr. Nur Rohman Rosyid at King Mongkut's Institute of Technology Ladkrabang for their useful suggestions.

REFERENCES

[1] K. Kuwabara, H. Kikuchi, M. Terada and M. Fujiwara, "Heuristics for Detecting Types of Infections from Captured Packets", IPSJ Malware Workshop (MWS2009), pp. 397-402, 2009.

[2] M. Hatada, Y. Nakatsuru, M. Terada and Y. Shinoda, "Dataset for anti-malware research and research achievements shared at the workshop", IPSJ Malware Workshop (MWS2009), pp. 1-8, 2009.

[3] T. Kobori, H. Kikuchi and M. Terada, "Interrelation between Interactive and Non-interaction Sensors", IPSJ Malware Workshop (MWS2008), Vol. 3, pp. 67-74, 2008.

[4] R. Agrawal R, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of ACM SIGMOD-93, pp. 207-216, 1993.

[5] Christian Borgelt, Apriori - Association Rule Induction, http://www.borgelt.net/apriori.html

[6] N. R. Rosyid, M. Ohrui, H. Kikuchi and P. Sooraksa, M. Terada, "Frequent Sequential Attack Patterns of Malware in Botnets", IPSJ Technical Report Computer Security Group (CSEC48), Vol.2010-CSEC-48, No.37, 2010.