A Study on Association Rule Mining of Darknet Big Data

Tao BAN, Masashi ETO, Shanqing GUO, Daisuke INOUE, Koji NAKAO, Runhe HUANG

***Abstract*—Global darknet monitoring provides an effective way to observe cyber-attacks that are significantly threatening network security and management. In this paper, we present a study on characterization of cyberattacks in the big stream data collected in a large scale distributed darknet using association rule learning. The experiment shows that association rule learning in the darknet stream data can support strategic cyberattack countermeasure in the following ways. First, statistics computed from malware-specific rules can lead to better understanding of the global trend of cyberattacks in the Internet. Second, strong association rules can lead to further insights into the nature of the attacking tools and hence expedite the diagnosis. Then, the discovery of emerging new attacks may lead to early detection and prompt prevention of pandemic incidents, preventing damage to the IT infrastructure and extensive financial loss. Finally, exploring the knowledge in the frequent attacking patterns can enable accurate prediction of future attacks from analyzed hosts, which could improve the performance of honeypot systems to collect more pertinent malware information using limited system and network resources.**

I. Introduction

In the past few years, there have been emerging cyber attacks discovered and reported, posing significant threats to the confidentiality, integrity, and availability of the data stored and communicated using the Internet. Distributed Denial of Service Attacks (DDoS), such as Distributed Reflective DoS (DRDoS) attacks using Network Time Protocol (NTP) or Domain Name Service (DNS) protocol [1], [2], have showed growing attack bandwidth and increased degree of difficulty to mitigate. Computer viruses, such as CryptoLocker [3] and Gameover Zeus [4], show a clear trend of profit driven and increased code obfuscation. Botnets, such as BredoLab and Conficker, are now featured by large scale infection using sophisticated, deliberate, and well directed exploits [5], [6], [7].

To address the concerns raised by these threats, there is a pressing need for the development of global early warning systems which could provide detailed forensic information on new threats in a timely manner. While the computation and communication costs for monitoring a densely populated network of global scale may render the task impossible, the monitoring of unused address space called darknet, usually provides a good trade-off between the monitoring cost and global knowledge acquisition. A darknet, also known as network telescope, blackhole monitors, Sinkholes, or background radiation monitors, is a portion of routed, allocated IP space that contains no advertised services [8], [9]. Because of the absence of legitimate hosts on the darknet, any traffic observed on a darknet is by its presence aberrant: it is either caused by malicious intent or mis-configuration. Assorted works have deployed darknets in existing networks to help identify the types and sources of malicious traffic present on the larger network of which they form a part, where darknets are used to host flow collectors, backscatter detectors, packet sniffers and so on [10], [11]. Considerable improvement in detection rate and cutdown in false positive rate are reported in related works, which in turn increases the awareness of malicious or mistaken activities and eases the mitigation.

The study presented in this paper is driven by the necessity to gain further understanding of the nature of malware-infected hosts, to identify their behavioral regularities, and to predict their future activities based on historical information. We describe a study on behavior analysis of the attacking hosts monitored on a darknet employing association rule learning [12], [13]. The association rules discovered from the darknet traffic are used to characterize the regularities among the scanning behavior of attacking hosts. To the best of our knowledge, this is the first application of association rule learning to darknet traffic, with a large number of association rules of particular interest discovered and explained.

The series of experiments produced the following findings: (1) Network devices infected by the same type of malware tend to probe the Internet in a predefined way, resulting in strong association rules discovered from the darknet traffic; (2) There are strongly correlated destination ports probed by the attacking host, which can be used to identify the scanning activity of a specific malware program. (3) There exists a large group of devices which tend to probe a large range if IP spaces, and their next targets are partially predictable based on the past behavior. We believe that based on these discoveries, timely and effective countermeasures can be devised to counterattack the emerging threats on the Internet.

This paper is organized as follows. Section II introduces the terminology, the concepts, and related work of association rule learning. Section III describes how to apply association rule learning to analyze the darknet traffic. Section IV reports the experiment results. Finally, Section V presents our conclusions.

II. Association Rule Learning

This section briefs association rule learning, a commonly applied technique for discovering interesting relationships hidden in a database.

A. Frequent Pattern Mining

The problem of association rule learning was originally proposed in the context of market basket data in order to find frequent groups of items that are purchased together [12], [14]. Following the original definition in [12], the problem of association rule learning is defined as follows.

978-1-4799-1959-8/15/$31.00

Let $\mathcal{D} = \{T_1, T_2, \ldots, T_N\}$ be a set of N transactions called the *database*. Let $I = \{i_1, i_2, \ldots, i_M\}$ be the universal set of M all items present in the database. Each transaction in $\mathcal{D}$ has a unique transaction ID and contains a subset of the items in I. The *support* $\mathrm{supp}(X)$ of a set of item (for short itemset) X is defined as the number/proportion of transactions in the database which contain the itemset.

Frequent pattern mining is to determine all patterns $P \subset I$ that are present in at least a fraction S of the transactions. The fraction S is referred to as the *minimum support*. It can be expressed either as an absolute number, or as a fraction of the total number of transactions in the database.

An *association rule* is defined as an implication of the form

$$X \to Y, \text{ for } \mathrm{X}, \mathrm{Y} \subseteq \mathrm{I}, \mathrm{X} \cap \mathrm{Y} = \emptyset. \tag{1}$$

The itemsets X and Y are called *antecedent* and *consequent* of the rule respectively. The *confidence* of a rule is presented by the conditional probability, $P(Y|X)$, i.e.,

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X). \tag{2}$$

To select interesting rules from the set of all possible rules, rules that satisfy both a minimum support threshold, S, and a minimum confidence threshold, C, are called *strong*.

In general, association rule learning can be done in the following two steps:

1) *Frequent pattern mining:* Each of the itemsets will satisfy the minimum support, i.e., occurs at least as frequently as S.
2) *Strong association rule generation:* By definition, rules created from the frequent itemsets with guaranteed minimum support must satisfy the minimum confidence constraint.

B. Frequent Pattern Mining using FP-tree

The first step in association rule learning involves searching in a power set of all possible combinations of items, whereas the size of this set grows exponentially in the number of items n in I. The key to an efficient search algorithm is the so-called *Apriori property*:

All nonempty subsets of a frequent itemset must also be frequent. Thus for an infrequent itemset, all its supersets must also be infrequent.

One of the currently fastest and most popular algorithms for frequent itemset mining is the Frequent Pattern growth (FP-growth) algorithm [14]. It is based on a prefix tree representation of the given database. By using a prefix tree data structure – the so-called FP-tree – FP-growth can save considerable amounts of memory for storing the transactions.

The basic idea of the FP-growth algorithm can be described as a recursive elimination scheme as follows.

1) In the first pass, derive the set of frequent items and their support counts. Delete all items from the transactions which do not satisfy the minimum support constraint. All frequent items are stored in a header table in descending order of their frequency.
2) In the second pass, build an FP-tree by inserting instances into a tree with a root node labeled as "null". To speed up the processing of the FP-tree, items in each transaction are sorted in the same order as in the header table. All nodes referring to the same item are indexed by a list so that all transactions containing the item can be accessed and counted by traversing this list. The header elements to the list are associated with the corresponding items in the header table.
3) Recursive mining of the FP-tree can grows large itemsets directly, without generating candidate items and testing them against the entire database. Start from the bottom of the header table, build the conditional item base for the length-1-pattern, which consists of a set of prefix paths in the FP-tree co-occurring with the suffix item. Then, a conditional FP-tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet the minimum support threshold, and processing continues on the remaining header items of the original FP-tree.
4) Once the recursive process has completed, all large itemsets satisfying the minimum support constraint are found, and association rule creation begins.

C. Association Rule Generation from Frequent Itemsets

Association rules can be generated based on the frequent itemsets in the following steps.

1) For each frequent itemset l, generate all nonempty subset of l.
2) For every nonempty subset s of l, output the rule "$s \to (l - s)$" if its confidence is higher than minimum confidence threshold C.

Since the rules are generated from frequent itemsets, all association rules created in such a way automatically satisfy the minimum support.

D. Other Related Work

Frequent pattern mining have numerous applications to major data mining problems such as customer transaction analysis, web log mining, software bug analysis, and chemical and biological applications.

III. Application to Attacking Host Behavior Characterization

Traffic data captured on a darknet contain valuable forensic information of programming techniques that are exploited to scan the Internet. In this section, we describe the application of association rule learning to characterize the behavior of attacking hosts observed in the darknet.

A. Objective

Discovery of behavior regularities of the attacking host may complement existing malware countermeasures in the following aspects. First, discovery of prevalent attack patterns may lead to further insights into the mechanism of the attack and thus enables countermeasure for the attack. For example, if the attacks to a number of services are shown to be highly correlated, efficient firewall rules and IDS signatures can be created by taking account of the fact. Second, the emergence of new attack patterns/graphs may be the symptom of pandemic incidents whose early detection and take-down could lead to prevention of heavy loss. The detection of new attack patterns can be achieved buy finding the abrupt change points on the time series of number of specific frequent patterns. Last but not least, such information can be used to improve the performance of monitoring systems so that more pertinent malware information can be collected using limited system and network resources. For example, darknet sensors can be updated to a responsive systems to obtain more information if prediction of the attacked IP addresses can be done based on the frequent patterns.

B. Characteristics of the Darknet Traffic

The results reported in this paper are based on long term observation over a group of darknet sensors hosted in the NICTER project [15], [16]. The sensors are installed in a variety of network environments. Darknet monitoring relies on the fact that most kinds of malware engage a exploitation procedure in search for the next potential victims. Thus the more IP addresses encompassed by the darknet, the more essential information the darknet could gather. Of course, due to the limited scale of a darknet compared with the IPV4 space and its passive nature, only partial information of the attacks is available on the darknet.

As depicted in Fig. 1, there are two typical settings for the sensors. For sensor type I, all IP addresses within the network range are allocated to the sensor, therefore packets that are directed to a consecutive IP subspace enveloped by the sensor's IP range will be fully captured. For sensor type II, the IP address space is sparsely populated with darknet addresses interpreted by active IP addresses. Such a variant is also known as a greynet [17]. Generally, type I has increased chance to capture continuous attack information because of the continuity of the monitored IP space. On the other hand, a greynet can implement better secrecy policy to prevent information leak of the network setting. In addition, greynet can capture information from malware clients which reside in the local network and tend to target only neighboring IP addresses. The detail information of the sensors are listed in Table I.

The attacks towards the darknet are captured in the form of network packets. In general, a packet consists of two kinds of data: control information and user data (also known as payload). Due to the passive nature of the darknet, expect for some special cases, e.g., a mis-configured server which connects to a presumed printer in the darknet space, there will

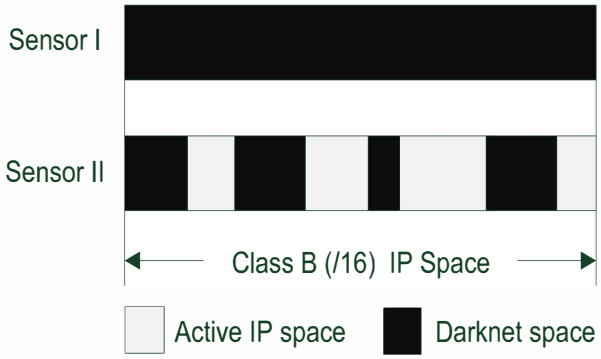

Fig. 1. Formulation of darknet sensors

be little persisting connections toward a darknet, especially toward a single IP address. Most of the observed hosts send only a couple of connection initializing packets to the darknet, i.e., TCP packets with SYN flag on. This renders the darknet connections more fragmented, lack of application level information. Therefore our analysis focus on the control information in packet headers, including time stamp of the communication, source and destination IP addresses, source and destination ports, used IP protocols, and many other pieces of information about the communication.

C. Experiment Design

The regularities in the scanning packets can be exploitable at multiple levels. First, they can be at packet level, e.g., scans towards a certain vulnerability usually are confined to a specific destination port. Second, they can be at target-host level, e.g., an attack can be featured by the combination and order of destination ports on a targeted host. Then, they can be at network-level, e.g., the precoded rules to select the next targets. Moreover, they can be at meta-level, e.g., the strength and frequency of the scan could be exploited for characterization purpose. All of the above information can be helpful to determine the type and source of the attacking host in order to acquire information of emerging network attacks or gain the global situation of malware contamination.

In this section, we present two case studies of association rule learning on darknet traffic analysis.

1) Correlation between Destination Ports: The first experiment tries to explore the correlation among the destination ports probed by the attacking hosts.

Network ports, which provide essential identifying information for open services, are the entry points to any networked device. The port number, identified by a 16-bit number, together with a device's IP address, completes the destination address for a communication session. The open ports on a device are usually probed by malware to determine available services before exploitation of known vulnerability on the service.

Discovered strong association rules with regards to the destination ports could provide useful information in the

TABLE I
ONE-DAY STATISTICS OF DARKNET SENSORS ON JUNE 11, 2014.

Sensor ID	A	B	C	D	E	F
Type	I	II	I	II	II	II
Size	/16	/16	/18	/18	/18	/17
#Pkt/IP	161.51	190.18	193.86	281.77	414.97	406.66
#Host/IP	85.55	118.48	118.97	161.07	175.40	230.87
#Ports	65536	63227	65224	30728	29651	46678
Port 1	23	8	8	445	445	445
Port 2	8	23	23	23	23	23
Port 3	29735	3389	29735	8	8	8
Port 4	29991	80	3389	3389	3389	3389
Port 5	30247	29735	29991	21060	30759	30759
Port 6	30503	8080	80	60557	80	80

Type: I for darknet sensor and II for greynet sensor. Size: size of the sensors in CIDR (Classless Inter-Domain Routing) notation. #Pkt/IP: average number of received packets for each darknet IP address. #Host/IP: average number of unique attacking hosts observed on each darknet IP address. #Ports: number of probed ports on the darknet sensor. Port 1 to 6: top six probed destination ports, ranked by the number of unique hosts observed on the port.

following aspects. First, because different malware programs usually exploit different combinations of vulnerable ports, the destination ports may provide deterministic information to identify the specific malware or offer hints to the intent of the attacker. Therefore, frequent pattern mining can be an efficient approach to automated malware signature extraction. Second, frequently probed port sets can reveal the most vulnerable services and therefore provide valuable clues for malware diagnosis.

2) Correlation between Probed Darknet Sensors: The darknet sensors described in Table I are installed in a variation of network environments. The second experiment is designed to explore the correlation between these sensors. By finding the strong relationships between the sensors, we can tell which sensors can capture the same scan information, so that traffic of the highly correlated sensors can be summed up for better analytical performance. Moreover, if some sensors has a strong probability of co-occurrence, system can be more reasonably adjusted to improve the monitoring performance. For example, a dynamic honeypot system can be can be reconfigured in real-time to accommodate a predictable probe.

IV. EXPERIMENTS

A. Correlation on Destination Ports

To discover the correlation among destination ports, the mining problem on destination ports is formulated by defining the set of unique port numbers probed by an attacking IP in one day as the transactions in the database. Note that due to the dynamic IP address assignment mechanisms commonly used in modern networks, e.g., IP addresses assigned via the DHCP protocol, the packets from a single IP address in a long run may contain communication information of multiple attacking hosts. On the other hand, because the scale of the darknet under discussion only occupies a comparatively small portion of the IPV4 space, the chance for multiple independent hosts scanning the same darknet is small enough to be ignored. Therefore, it is reasonable to take all the packets launched by an IP during a 24-hour period as from a single host. Naturally, due to the DHCP problem, a host can contribute more than one transaction to the database, which will not render significant problem to this study.

TABLE II
FREQUENT ITEMSETS RELATED TO DESTINATION PORT 80, FROM 1-DAY TRAFFIC OF SENSOR A.

ID	DPort 1	DPort 2	DPort 3	Occur.
1	80			2932
2	80	8		747
3	80	443		786
4	80	13		715
5	80	8	443	741
6	80	8	13	713
7	80	13	443	712
8	80	8	13	443 711

The network services on involved ports are as follows. 8: service unassigned; 13: the daytime protocol; 80: hypertext transfer protocol (HTTP); 443: hypertext transfer protocol over TLS/SSL (HTTPS).

Table II shows the frequent itemsets learned from the 1-day traffic of sensor A, with minimum support set to be 700. Eight frequent itemsets which are related to the well known port 80 are selected from a pool of 610 frequent itemsets. Because of the popularity of port 80 used for hosting web service, many attacks tend to probe this port. As shown in the table, 2932 hosts had attacked port 80 on the day. Many ports are probed together with port 80, among which are ports 8, 13, and 443. In the table, all the frequent itemsets that are highly related to these 4 ports are shown, with the number of their occurrences shown on the last column. Obviously, ports 8, 13, and 443 have a strong correlation, i.e., they tend to be probed at the same time. This is confirmed by the association rules shown in Table III, which are generated from the frequent patterns in Table II. In the table, despite of the high number of co-occurrence between ports 80 and 13, the association rule $P80 \rightarrow P13$ only has a confidence of 24.3%, failing to meet

the minimum confidence requirement 80%. On the contrary, the association rule $P13 \rightarrow P80$ has a strong confidence of 94.7%. Therefore, probes to port 13 can be considered as the causal factor of the probes to port 80, e.g., if a packet directed to port 13 is observed from a host, then port 80 will be also probed with a large chance.

TABLE III
ASSOCIATION RULES CREATED FROM FREQUENT ITEMS IN TABLE II

ID	Rule	Support	Confidence
1	$80 \rightarrow 8$	747	27.5%
2	$8 \rightarrow 80$	747	4.7%
3	$80 \rightarrow 13$	715	24.3%
4	$13 \rightarrow 80$	715	94.7%
5	$80, 443 \rightarrow 8$	741	94.3%
6	$8, 443 \rightarrow 80$	741	95.5%
7	$8, 80 \rightarrow 443$	741	99.2%
8	$13, 443 \rightarrow 80$	712	95.3%
9	$80, 443 \rightarrow 13$	712	90.6%
10	$13, 80 \rightarrow 443$	712	99.6%
11	$8, 13 \rightarrow 80$	713	95.2%
12	$8, 80 \rightarrow 13$	713	95.4%
13	$13, 80 \rightarrow 8$	713	99.7%
14	$13, 8, 443 \rightarrow 80$	711	95.4%
15	$8, 80, 443 \rightarrow 13$	711	96.0%
16	$13, 80, 443 \rightarrow 8$	711	99.9%
17	$8, 13, 80 \rightarrow 443$	711	99.7%

The first three rules which do not satisfy the minimum confidence $C = 80\%$ are not considered as strong association rules.

Take the rules 5 to 7 of in Table III as another example. These three rules illustrate the correlation between ports 8, 80, and 443. If two of the ports are probed, the chance for the other port to be probed is over 94%. Because of the high correlation of these three ports, they can be treated as the signature of the scanning behavior.

TABLE IV
FREQUENT ITEMSETS RELATED TO DESTINATION PORT 80, FROM 1-DAY TRAFFIC OF SENSOR A.

ID	DPort 1	DPort 2	DPort 3	Occur.
1	23			20141
2	210			20047
3	23	210		19778
4	23	210	1526	1414
5	23	210	3351	1334
6	23	210	8010	1145

The network services on ports are as follows. 8: an unassigned port; 13: the daytime protocol; 80: hypertext transfer protocol (HTTP); 443: hypertext transfer protocol over TLS/SSL (HTTPS).

Table V shows another group of highly correlated probed ports including ports 23, 210, 1526, 3351, and 8010. Ports 23 and 210 yield very high support counts: there are more than 20,000 hosts that are probing both of them on the day. The high correlation between these two ports can be illustrated by the high confidence of the two rules shown in line 3 and line 4 in the table, both of which have more than 98% confidence. A further breakdown of the scans show that some side probes are also involved in the scans, including ports like 1526, 3351, and 8010. Although the support counts of these transactions

TABLE V
ASSOCIATION RULES CREATED FROM FREQUENT ITEMS IN TABLE IV-A

ID	Rule	Support	Confidence
1	$210 \rightarrow 23$	20047	98.66%
2	$23 \rightarrow 210$	20141	98.20%
3	$23, 1526 \rightarrow 210$	1150	99.57%
4	$210, 1526 \rightarrow 23$	1422	99.44%
5	$210, 8010 \rightarrow 23$	1150	99.57%
6	$23, 8010 \rightarrow 210$	1156	99.05%
7	$210, 3351 \rightarrow 23$	1343	99.33%
8	$23, 3351 \rightarrow 210$	1341	99.48%

The network services on the ports are as follows. 23: telnet protocol, which supports unencrypted text communications; 210: ANSI Z39.50 service, which is an application layer communications protocol for searching and retrieving information from a database over a TCP/IP computer network; 8010: Extensible Messaging and Presence Protocol (XMPP) based file transfer service; 1526 and 3351: service assigned.

decrease as more ports are involved, the confidence of the association rules are improved by including more information on the left hand side of the rules.

The strong association rules discovered in the above experiments indicate that the strongly correlated destination ports could be the identifying signature of malware. However, to prove this, information from other data sources are needed to give precise information of the malware programs performing the probes. In fact, the above findings are confirmed to be associated with the Carna botnet [18]. The Carna botnet was created by intruding more than 420,000 embedded devices that were accessible online with default credentials. After the intrusions, a small binary are uploaded to those devices to conduct an Internet-wide scan of the IPv4 address space. The owners of the Carna botnet claimed that the botnet was created for research purpose and they published a detailed description of how they operated, along with 9TB of raw logs of the scanning activity. According to previous work in [19], probes to ports 8, 80, and 433, and probes to ports 23 and 210 are reported as the signatures of the network scans performed by different fractions of the botnet.

TABLE VI
ATTACKING HOSTS OBSERVED CROSS THE SENSORS.

ID	A	B	C	D	E	F
A	**506805**					
B	36798	**90512**				
C	44870	26205	**159907**			
D	13385	9905	10810	**63693**		
E	14099	10649	11690	27832	**62003**	
F	20149	14138	15461	16257	16563	**57703**

The number of hosts observed on a sensor is shown on the diagonal in bold font.

B. Correlation on Probed Darknet Sensors

Many hosts tend to probe a large range of IP addresses. Despite that the first octet of the sensors listed in Table I are different from each other, there are still many hosts observed cross the sensors. This could be roughly depicted by the

result of correlation analysis in Table VI. Take the last line in the table as an example, we can read as flows. Sensor F has been probed by 57,703 attacking hosts on the day. Among these hosts, 20,149 have probed sensor A, 14,138 have probed sensor B, 15,461 have probed sensor C, 16, 257 have probed sensor D, and 16,563 have probed sensor E. Although the majority of the hosts probe only one sensor, there are a considerable group of hosts tend to probe two or more sensors. The simple correlations shown in Table VI could only evaluate the relationship between two sensors, they suggest the necessity for further analysis of the correlation between multiple sensors.

To discover the correlation among multiple darknet sensors, the transaction database is created as follows. Each transaction in the database is defined as the set of unique darknet sensor IDs probed by a host on the day. Table VII lists all the frequent itemsets which satisfy the minimum support of 10,000. Most of these frequent itemsets involve more than two darknet sensors. Especially, the frequent pattern on the last line spreads among sensors A, B, C, and F, with a high support count of 12,258. Compare this result with that in Table VI, we can see that higher order correlation could be better described by the frequent itemsets.

TABLE VII
FREQUENT ITEMSETS ON SIMULTANEOUSLY PROBED DARKNET SENSORS, FROM 1-DAY TRAFFIC OF SENSORS A TO F.

ID	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Occur.
1	E	F			16563
2	D	F			16257
3	D	E			27832
4	D	E	F		11486
5	B	F			14138
6	B	E			10649
7	C	B			26205
8	C	B	F		12353
9	A	B	F		13775
10	A	B	E		10408
11	A	C	F		14833
12	A	C	E		11242
13	A	C	D		10366
14	A	C	B		24826
15	A	C	B	F	12258

Table VIII contains all the association rules created from the frequent itemsets in Table VII, which satisfy the minimum confidence of 80%. Based the high confidence, it is possible to predict the next targeted darknet sensor when all the sensors on the left hand side of the rules are all probed. For example, based on rule 2, if sensors B, and F are both probed, then it is very likely (with probability 97.43%) that sensor C will also be probed. Therefore, we can either appoint a predefined remedy to protect sensor C if the purpose is to reduce the risk or prepare a honeypot system to collect more information about the probe if the purpose is to collect information.

V. CONCLUSION

In this paper we presented a study on application of association rule learning on the darknet traffic. Frequent itemsets with respects to probed destination ports and probed darknet sensors are reported in the experiments. Some of the significant association rules discovered in the experiment are proved to correspond to special attacks from known botnets. We believe strategic countermeasure to related attacks can be enabled based on these discoveries.

TABLE VIII
ASSOCIATION RULES CREATED FROM FREQUENT ITEMSETS IN TABLE IV-B

ID	Rule	Support	Confidence
1	$B, F \rightarrow C$	14138	87.37%
2	$B, F \rightarrow A$	14138	97.43%
3	$B, E \rightarrow A$	10649	97.73%
4	$C, F \rightarrow A$	15461	95.94%
5	$C, E \rightarrow A$	11690	96.17%
6	$C, D \rightarrow A$	10810	95.89%
7	$A, C, F \rightarrow B$	14833	82.64%
9	$A, B, F \rightarrow C$	13775	88.99%
10	$C, B, F \rightarrow A$	12353	99.23%
11	$C, B \rightarrow A$	26205	94.74%

For future work, we will explore the ways to perform real-time prediction of host behavior, to facilitate prompt response and prevention against the cyberattacks.

REFERENCES

[1] G. Loukas and G. Öke, "Protection against denial of service attacks: A survey," *The Computer Journal*, p. bxp078, 2009.

[2] R. van Rijswijk-Deij, A. Sperotto, and A. Pras, "Dnssec and its potential for ddos attacks: A comprehensive measurement study," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, ser. IMC '14. New York, NY, USA: ACM, 2014, pp. 449–460. [Online]. Available: http://doi.acm.org/10.1145/2663716.2663731

[3] L. Kelion, "Cryptolocker ransomware has 'infected about 250,000 pcs," *BBC*, 12/2013 2013.

[4] D. Andriesse, C. Rossow, B. Stone-Gross, D. Plohmann, and H. Bos, "Highly resilient peer-to-peer botnets are here: An analysis of gameover zeus." in *MALWARE*. IEEE, 2013, pp. 116–123.

[5] D. Graaf, A. Shosha, and P. Gladyshev, "Bredolab: Shopping in the cybercrime underworld," in *4th International Conference on Digital Forensics & Cyber Crime*, Lafayette, Indiana, USA, 10/2012 2012. [Online]. Available: http://hdl.handle.net/10344/2896

[6] Y. Nadji, M. Antonakakis, R. Perdisci, D. Dagon, and W. Lee, "Beheading hydras: Performing effective botnet takedowns," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 121–132. [Online]. Available: http://doi.acm.org/10.1145/2508859.2516749

[7] R. Shirazi, "Botnet takedown initiatives: A taxonomy and performance model," *Technology Innovation Management Review*, vol. 5, pp. 15–20, 01/2015 2015.

[8] M. Bailey, E. cooke, F. Jahanian, N. Provos, K. Rosaen, and D. Watson, "Data reduction for the scalable automated analysis of distributed darknet traffic," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, ser. IMC '05. USENIX Association, 2005, pp. 239–252.

[9] T. Ban, L. Zhu, J. Shimamura, S. Pang, D. Inoue, and K. Nakao, "Behavior analysis of long-term cyber attacks in the darknet," in *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part V*, 2012, pp. 620–628.

[10] U. Harder, M. W. Johnson, J. T. Bradley, and W. J. Knottenbelt, "Observing internet worm and virus attacks with a small network telescope," *Electronic Notes in Theoretical Computer Science*, vol. 151, no. 3, pp. 47–59, 2006.

[11] K. Benson, A. Dainotti, K. Claffy, and E. Aben, "Gaining insight into as-level outages through analysis of internet background radiation," in *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on*. IEEE, 2013, pp. 447–452.

[12] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993.
[13] C. Borgelt, "Frequent item set mining," *Data Mining Knowledge Discovery*, vol. 2, no. 6, pp. 437–456, 2012.
[14] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, 2000.
[15] K. Nakao, K. Yoshioka, D. Inoue, and M. Eto, "A novel concept of network incident analysis based on multi-layer ovservation of malware activities," in *The 2nd Joint Workshop on Information Security (JWIS07)*, 2007, pp. 267–279.
[16] D. Inoue, K. Yoshioka, M. Eto, M. Yamagata, E. Nishino, J. Takeuchi, K. Ohkouchi, and K. Nakao, "An incident analysis system nicter and its analysis engines based on data mining techniques," in *15th International Conference on Neuro- Information Processing of the Asia Pasific Neural Netowrk Assembly (ICONIP 2008)*, 2008.
[17] W. Harrop and G. J. Armitage, "Defining and evaluating greynets (sparse darknets)," in *LCN'05*, 2005.
[18] C. Stocker and J. Horchert, "Mapping the internet: A hacker's secret internet census," *Spiegel Online*, 22/3 2013.
[19] E. Le Malécot and D. Inoue, "The carna botnet through the lens of a network telescope," in *Foundations and Practice of Security*. Springer, 2014, pp. 426–441.