International Journal of Cyber-Security and Digital Forensics (IJCSDF) 5(4): 166-174
The Society of Digital Information and Wireless Communications, 2016 (ISSN: 2305-0012)

Method for detecting a malicious domain by using only well-known information

MASAHIRO KUYAMA, YOSHIO KAKIZAKI, RYOICHI SASAKI
Tokyo Denki University
Tokyo, Japan
kuyama@isl.im.dendai.ac.jp

ABSTRACT

Damage caused by targeted attacks is a serious problem. It is not enough to prevent only the initial infections, because techniques for targeted attacks have become more sophisticated every year, especially attacks seeking to illegally acquire confidential information. In a targeted attack, the attacker wants to hide the C&C server so that it cannot be detected. Therefore, the C&C server may not be found by a web search engine. We pay attention to this lack of detection and the results of a web search engine. In this study, we propose a method for identifying the C&C server by using supervised machine learning and feature points obtained from WHOIS, DNS and search sites for domains of C&C servers and normal domains. Moreover, we conduct an experiment that applies real data, and we verify the usefulness of our method by cross-validation. The results indicated that we could obtain a high detection rate of about 99.3%.

KEYWORDS

Malware, C&C server, Neural network, SVM, Targeted attack

1 Introduction

The development of the Internet has contributed to the enrichment of society**.** In particular, the Internet allows connections to be made very quickly all over the world.
These connections have not only a good side but also a bad side, which is Internet-based crime. In such crimes, damage caused by targeted attacks aimed at a specific organization or company is a serious problem [1]. Many targeted attacks aim at illegal acquisition of confidential information, such as intellectual property. To achieve their objectives, attackers infect terminals with malware attached to e-mails and use the targeted attack to send information back to the command and control (C&C) server.
In Japan, many organizations, including a leading heavy-industry manufacturer, the House of Representatives, and the Japan Pension Service, have been subject to attacks and have suffered significant damage. Multi-layered countermeasures at the entry and exit points are required because it is very difficult to prevent attacks.
The sequence of targeted attacks consists of four steps, as shown in Figure 1.

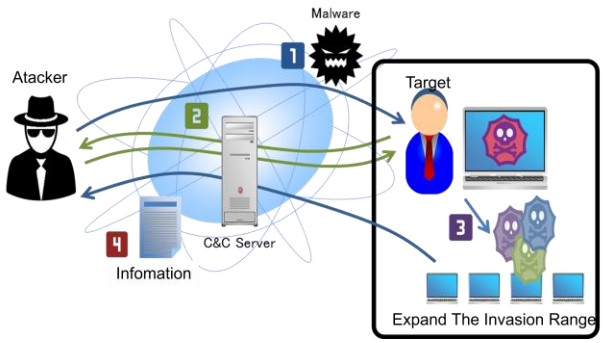

Figure 1 Sequence of targeted attacks

Step 1: A terminal such as a PC in a local area network (LAN) is infected with malware for the targeted attack.
Step 2: This terminal communicates with the C&C server. Then, more malware is downloaded to the terminal.
Step 3: The malware attempts to expand the invasion range to other PCs and servers in the LAN.

Step 4: Important information, confidential information and private information of the organization is transmitted to the C&C server owned by the attacker outside the LAN.

The malware used in the targeted attacks in step 1 is customized for each targeted organization and the malware is difficult for the anti-virus software to detect.
In targeted attacks, the C&C server is essential for the attack to succeed. Therefore, if we can detect the infection of a terminal in the LAN and the communication with the C&C server, we can guard against the expansion of damage. Also, the attacker wants to hide the C&C server so that it cannot be detected. In order to detect the anomalous communication, we must identify the C&C server in advance to detect the infection. New C&C servers are continuously made by attackers, and so their IP addresses are not typically on any blacklists. For this reason, we need to develop a method to find new C&C servers.
In this study, we extract the feature points from well-known information such as WHOIS, DNS, and the results of a web search engine for the C&C server domain, and we try to detect the C&C server by using a neural network.
In a presentation at DigitalSec2016, we showed that a method using WHOIS and DNS information could identify a C&C server and detect it with a 98.5% success rate [2]. In this paper, we report an improved detection rate by using the result of a web search engine.
The C&C server is hidden to avoid detection. This means that it is not typically found on search sites. Therefore, we assume that our search engine is effective to detect the C&C server.
We extract feature points according to their difficulty of spoofing: valid terms, expiration dates, and e-mail addresses from the WHOIS information, number of mail exchanger (MX) records, number of name server (NS) records from the DNS information, and results of the web search engine.

2 Related Work

Studies for specifying the C&C server are classified into the following two types.

(1) Studies that focus on communication packets between the bot PC and the C&C server

Jang et al. [3] and Lu et al. [4] proposed methods to detect a C&C server by analyzing the payload of communication packets between the bot PC and the C&C server. Ikuse et al. [5] proposed methods to detect a C&C server in order to identify the falsification of communication data by performing a malware analysis that applies taint analysis technology.
These methods have a high detection rate, because the data body is used for verification and excludes header information, such as the destination address and the source address, which might have a specification change, such as the port number or the proprietary protocol of the transport layer.
However, we have to observe real-time communication. Moreover, inadequacies in the response to unverified issues such as zero-day attacks must be solved.

(2) Studies that focus on domain information for the C&C server

Tsai et al. [6] reported a detection method based on a data mining technique called RIPPER, which uses a combination of information obtained from the domain information and the external repositories of the C&C server. Felegyhazi et al. [7] proposed a method for estimating the identity of an unknown malignant domain from WHOIS and DNS information. Ma et al. [8] proposed a technique using machine learning and DNS, WHOIS, and geographic information for the URL. Invernizzi et al. [9] reported a method for estimating the identity of an unknown malignant domain by using a search engine to

obtain information such as the content of a WHOIS known malignant website.
Although the detection failure rate is high, this method does have an adequate accuracy rate for detecting C&C servers.

In our previous study, we proposed a detection technique that had a 96.5% detection rate in 2009 [10].
Our method used a valid term and reverse lookup of C&C domain information from DNS and WHOIS information. Therefore, acquisition of the data required for analysis was easy. In addition, the method was highly unlikely to be affected by malware because it did not need direct access to the C&C server.
We have been continuing our investigation of the detection rate, which has decreased over time [11]. As shown in Table 1, the detection rate for the data of 2009 was 96.5%. However, the detection rate fell to 85.0% in 2010 and 76.2% in 2011.

Table 1 Detection rates of our method over time

Method year	Detection rate by year (%)				
	2009	**2010**	**2011**	**2013**	**2014**
2009	96.5	85.0	76.5	-	-
2011	-	-	95.2	42.5	-
2013	-	-	-	80.3	80.8
2014	-	-	-	-	96.7

These results revealed that the 2009 parameters were not suitable for 2010 and 2011[11].
We updated the discriminative model by using recent data to optimize the detection method in each period [12]. Although our results improved, it was necessary to frequently update the discriminative model.
In our 2014 update, we revised our method to use quantification theory and machine learning. The result was still not high enough, although the detection rate improved to 96.7% in 2014 [11].
We improved the detection rate to 98.5% in 2016 [2].Table 2 shows the changes in the characteristics used in the ongoing investigation.

Table 2 Changes in the characteristics used

Features using		Model year				
		'09	**'11**	**'13**	**'14**	**'16**
DNS	Reverse resolution	✓	✓	✓		
	TTL				✓	
	minimum	✓	✓		✓	
	A records		✓	✓		
	MX records					✓
	NS records				✓	✓
	CNAME records			✓		
	TXT records				✓	
WHOIS	Valid terms	✓	✓	✓	✓	✓
	e-mail addresses					✓
Total		3	4	4	5	4

We used WHOIS and DNS information for detection in the previous studies. Once we added new features for a search site, the detection rate went up.

3 Methods

Our proposed method focuses on the domain of the C&C server.
The method uses well-known information such as WHOIS, DNS, and the result of web search engine for the domain of the C&C server. Our method can identify C&C servers by extracting the feature points for machine learning.
To classify a domain as malignant (C&C) or benign (normal), we use machine learning to construct a training model in advance.

3.1 Detection method

First, we prepare benign domains and malignant domains. Then, feature points are extracted from the WHOIS, DNS, and search information for each domain.
The extracted features are used in machine learning to construct a training model (Figure 2). The training model determines whether an accessed domain is malignant or benign.

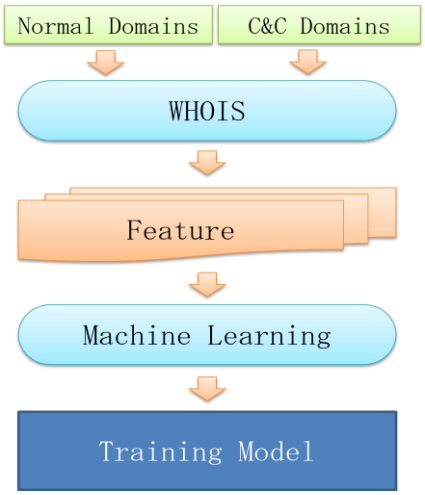

Figure 2 Flow of detection method

3.2 Preparing domains

We prepare two types of domains: normal domains and C&C domains.
We choose the normal domains from "the top 500 sites on the web" of Alexa [13], because the top sites have highly secure domains, which best represent normal domains.
The C&C servers are extracted by analyzing Emdivi, PlugX, and PoisonIvy, which are major malwares for targeted attacks [14][15].
We obtain 163 malwares by using VirusTotal [16]. Table 3 shows the breakdown of the collected malwares.

Table 3 Collected malwares from VirusTotal

Malware type	Samples
Emdivi	50
PlugX	63
PoisonIvy	50

The collected malwares are deeply analyzed by using the Sandbox analyzer called LastLine [17]. LastLine extracted 54 domain destinations as the analysis results.

3.3 Features of WHOIS

WHOIS is a service that provides management and information for the registration of a domain. Technical specifications and operational rules of WHOIS are established in RFC812 [18] and RFC3912 [19].
We can obtain the following information from WHOIS.

a) Registered domain name
b) Registrar name
c) DNS server name for the registered domain
d) Valid term for the domain
e) Expiration date for the domain
f) Domain name registrant contact
g) Person in charge for technical contact
h) Contact for registration personnel
i) Contact point for the registrant

It is difficult to tamper with the information from a) to e). The valid period d) for normal servers is long, but that for C&C servers is short, because C&C domains are canceled if their purpose is achieved [7][8][9]. From this viewpoint, we calculate the valid term by subtracting the date in d) from the date in e). Figure 3 shows the valid period for C&C domains and normal domains.

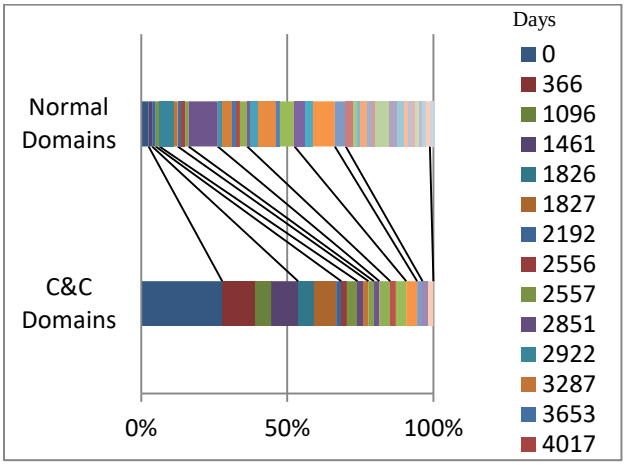

Figure 3 Valid terms for domains

As can be seen in Figure 3, the valid terms for C&C domains are shorter than those for normal domains.
Next, we obtain the following information for each contact person from f) to i).

a) ID
b) Name
c) Organization name
d) Address
e) Postal code
f) Phone number
g) Country
h) Fax number
i) E-mail address

All of the above information can be easily falsified.
Especially, the registration information for most C&C domains is false, because attackers often use WHOIS registration agency services to hide. However, the probability of a true e-mail address is high even if the other information is false, because the e-mail address is required for contact.
Thus, we pay attention to e-mail addresses obtained from WHOIS. First, we extract e-mail addresses from WHOIS for normal domains and then for C&C domains, and then we conduct data mining.
We extract the features for each domain by using a text mining tool called "UserLocal" [20].
Figures 4 and 5 show the co-occurrence network, which is the appearance pattern for words used in e-mail addresses, for normal domains and C&C domains, respectively.
The co-occurrence network shows relations by structuring the word patterns used in the text. Similar words in the appearance pattern are connected by a line.
We reveal the structures of the e-mail addresses for the domains and extract the features by using the word patterns.

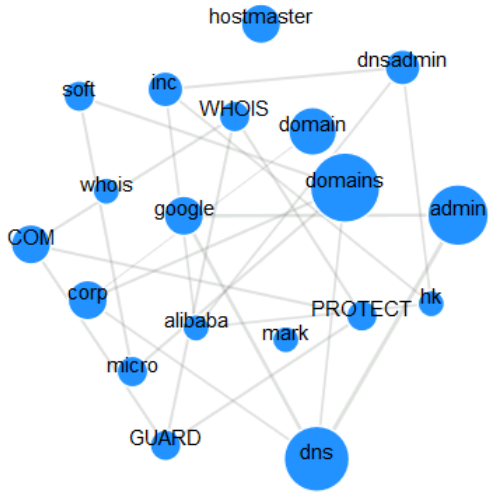

Figure 4 Co-occurrence network for normal domains

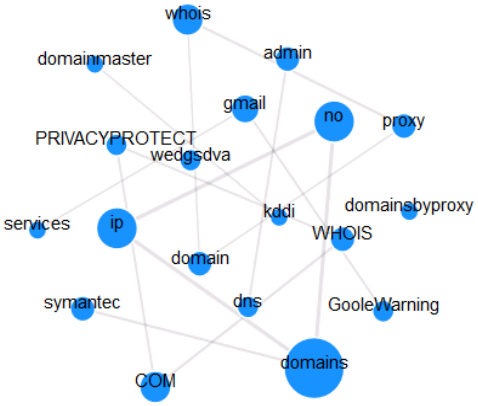

Figure 5 Co-occurrence network for C&C domains

The co-occurrence network for normal domains has a large structure connected with a plurality of words, and two patterns connected with four types of words. On the other hand, the co-occurrence network for C&C domains has three patterns connected with three types of words. When examined closely, "no", "proxy", and "PRIVACYPROTECT", which are usually used by WHOIS registration agency services, are included in Figure 5.
We should point out that the WHOIS registration agency services used for normal domains and C&C domains are different.

Therefore, we choose three features of WHOIS information: domain name, e-mail address, and valid term.

3.4 Features of DNS

DNS is a system that translates domain names into IP addresses. Technical specifications and operational rules of DNS in RFC1034 [21] and RFC1035 [22] are determined.
We can obtain the following records from the DNS.

a) Address (A) record
b) Start of authority (SOA) record
c) Host information (HINFO) record
d) MX record
e) NS record
f) Canonical name (CNAME) record
g) Well-known services (WKS) record
h) Text (TXT) record

The numbers of registered records for the NS record and the MX record show a remarkable difference. Figure 6 shows the number of NS records for normal domains and C&C domains. Figure 7 shows the number of MX records for normal domains and C&C domains.

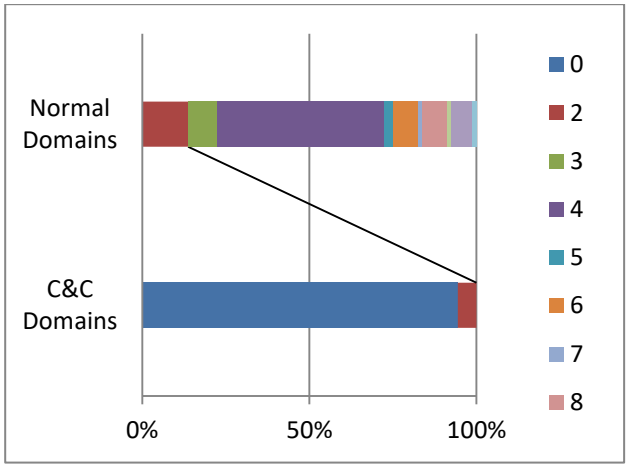

Figure 6 Number of NS records

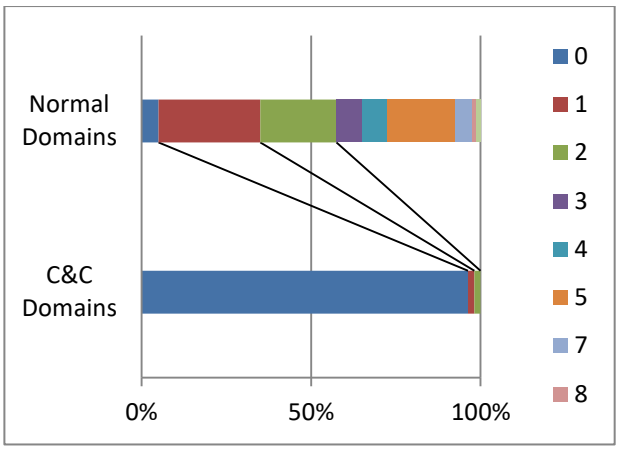

Figure 7 Number of MX records

Almost all records are not registered in the C&C domains, although many records are registered in the normal domains.
Thus, we choose these two features of DNS information: number of NS records and number of MX records.

3.5 Features of search site

A related study [9] detected an unknown C&C server by using a search engine to find the characteristics of a known C&C server in a drive-by-download attack.
In drive-by-download attacks, PCs are infected with malware by browsing websites. The malicious websites, which are infected with malware, conduct search engine optimization (SEO) to introduce more malware. It is assumed that the purpose of the SEO is to attract customers.
On the other hand, the website may not be used for malware infection in the targeted attack, because the malware is sent to the target directly by spoofing e-mail, etc. The attacker wants to hide the C&C server for the targeted attack so that it cannot be detected. In addition, a short-lived C&C server cannot be found by the crawler of the web search engine. Therefore, the C&C server may not be found by the web search engine. We note that this feature should receive particular attention.
In the present study, finding an evaluation domain by using the Google search engine was

investigated with regard to hits or non-hits. The results are shown in Figure 8.

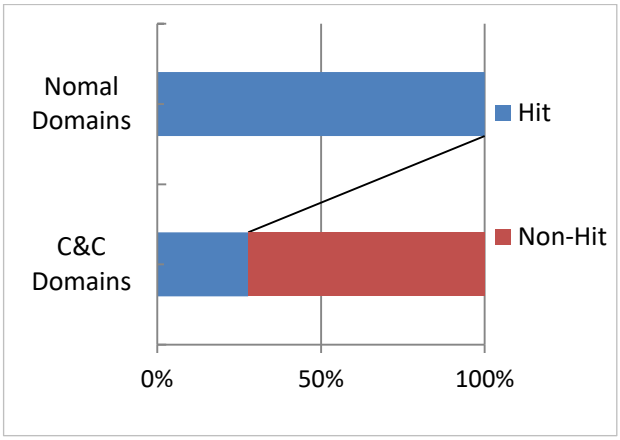

Figure 8 Search sites finding C&C servers

As shown in Figure 8, many C&C domains were not hit by the search site. Some hit C&C domains seemed to be hijacked servers. In targeted attacks, the C&C servers prepared by the attackers themselves are used.

3.6 Training model and algorithm

We construct a training model using a support vector machine (SVM) [24] and a neural network [25] as the algorithm for machine learning.
The SVM is a machine learning method that performs classification into two classes by pattern recognition [26].
A neural network is a type of supervised learning method.
It is possible to express the relation between the input and the output by mathematical modeling of some of the features found in human brain functions. The standard method is a hierarchical neural network using two layers.
We construct a training model by using a neural network with e-mail addresses and valid terms from WHOIS, the number of NS records, the number of MX records from the DNS, and the number of hits from a search site.
Table 4 shows the features included in machine learning.

Table 4 Features of machine learning

Input		Type
Label		Normal or C&C
Domain		String
WHOIS	Admin mail address	String
	Registered mail address	String
	Technical mail address	String
	Valid term	Number
DNS	NS record	Number
	MS record	Number
Search site		Hit or Non-Hit

4 Results

For evaluation, 80 normal and 54 C&C domains were used.
Because the amount of data was small, the accuracy could be low depending on how we chose the test data.
The amount of provided data for a particular domain used for targeted attacks was small. Thus, we evaluated the data with a cross-validation method, because it can reduce the error margin even for a small amount of data.
The cross-validation method divides the original data into block units [27]. One of the blocks is the test data, and the others are the learning data for evaluation.
The evaluation consisted of calculating the average of each evaluation result as the estimated accuracy (Figure 9).
This evaluation method can increase the estimation accuracy even for a small amount of data. The accuracy is calculated as follows.
Let N^{ts} be the total number of test data, and t^{ts} be the total number of data classified accurately, such that $A^{ts}(d^n) = \frac{t^{ts}}{N^{ts}}$. The *n*-th evaluation accuracy is estimated as follows:

$$A^{CV}(d) = \frac{1}{n}\sum_{i}^{n} A^{ts}(d^i) \quad (1)$$

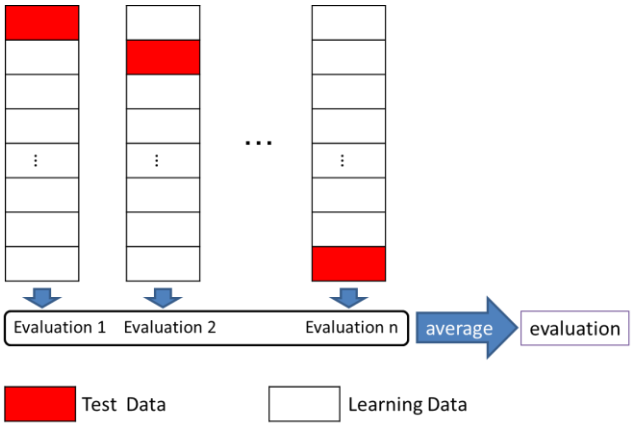

Figure 9 Cross-validation method

The results of the evaluation of each combination of WHOIS, DNS, and search site by the SVM and the neural network using the cross-validation method are shown in Table 5.

Table 5 Detection rates by cross-validation

Combination	SVM	Neural network
WHOIS only	88.8%	88.8%
DNS only	96.3%	95.5%
Search site only	88.8%	88.8%
WHOIS + DNS	97.8%	98.5%
WHOIS + Search site	91.8%	92.5%
DNS + Search site	99.3%	99.3%
WHOIS + DNS + Search site	99.3%	99.3%

As a result, the SVM and neural network achieved a superior detection rate of 99.3%.
The WHOIS only or the search site only achieved an 88.8% detection rate. However, the DNS only achieved a higher detection rate. The results of the WHOIS and the search site indicate that the DNS is an important element.
The result of adding the search site to the WHOIS and the DNS also improved the detection rate. Therefore, it is effective to use the search site. Moreover, removing WHOIS from the WHOIS, the DNS and the search site did not change the detection rate.
The C&C server constructed by the attacker was not hit by the search site. Therefore, the search site is effective for detection of the C&C server. However, the search is not valid if the attacker hijacks a server, so an attacker hijacking a server is detected by the combination of WHOIS and the DNS.

5 Conclusion

In this paper, we collected the feature points of e-mail addresses used for C&C domains and proposed a method to determine C&C servers by using machine learning with well-known information such as WHOIS, DNS, and the result of a web search engine.
We clarified the features of WHOIS registration agency services used for C&C domains by illustrating the relation of words in extracted e-mail addresses in co-occurrence networks. Moreover, the use of search sites for detection of C&C servers was found to be effective.
Finally, we evaluated domain names and e-mail addresses. The valid terms from WHOIS, the number of NS records, the number of MX records from the DNS, and the number of search sites returned by Google were input for machine learning. As a result, we were able to identify the C&C server at a high detection rate of 99.3%. In future work, we intend to improve the accuracy by revising the machine learning algorithms, input values, and preprocessing.

REFERENCES

[1] Cyber GRID View vol.1 English Edition http://www.lac.co.jp/security/report/pdf/apt_report_vol1_en.pdf

[2] M.Kuyama, Y.Kakizaki, R.Sasaki, "Method for Detecting a Malicious Domain by using WHOIS and DNS features" The Third International Conference on Digital Security and Forensics (DigitalSec2016), pp. 74-80(2016).

[3] D.I.Jang, M.Kim, H.C. Jung, B.N. Noh, "Analysis of HTTP2P Botnet: Case Study Waledac"2009 IEEE 9th Malaysia International Conference on Communications (Micc), pp. 409-412(2009).

[4] W.Lu, M.Tavallaee, A.A.Ghorbani, "Automatic Discovery of Botnet Communities on Large-Scale Communication Networks" ASIACCS '09 Proceedings of the 4th International Symposium on Information, Computer, and Communications Security(2009).

[5] T.Ikuse, K.Aoki, T.Yagi, T.Hariu, "Identifying C&C Server Based on Modified Data Descent Analysis" Proceedings of the 2014 IEICE SOCIETY Conference (2014).

[6] M.H.Tsai, K.C.Chang, C.C.Lin, C.H.Mao, H.M.Lee, "C&C Tracer: Botnet Command and Control Behavior Tracing" IEEE International Conference on Systems, Man and Cybernetics (SMC), Anchorage, AK, pp.1859-1864(2011).

[7] M.Felegyhazi, C.Kreibich, V.Paxson, "On the Potential of Proactive Domain Blacklisting" USENIX Conference on Large-scale Exploits and Emergent Threats, pp.6 (2010).

[8] J.Ma, L.K.Saul, S.Savage, G.M.Voelker,"Beyond Blacklists. Learning to Detect Malicious Web Sites from Suspicious URLs" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1245-1254(2009).

[9] L.Invernizzi, S.Benvenuti, P.M.Comparetti, M.Cova, C.Kruegel, G.Vigna. EvilSeed, A Guided Approach to Finding Malicious Web Pages" IEEE Symposium on Security and Privacy, pp.428-442(2012).

[10] H.Mihara, R.Sasaki, "Proposal and Evaluation of Technique to Detect C&C Server on Botnet Using Attack Data (CCCDATAset2009) and Quantification Methods Type II" Journal of Information Processing Society of Japan Vol. 51, No. 9, pp. 1579-1590(2010).

[11] S.Okayasu, R.Sasaki, "Proposal and Evaluation of Methods Using the Quantification Theory and Machine Learning for Detecting C&C Server Used in a Botnet" Computer Software and Applications Conference (COMPSAC) 2015 IEEE 39th Annual Vol. 03, pp. 24-29(2015).

[12] N,Nakamura, R.Sasaki, "Evaluation of Technique to Detect C&C Server of Botnet Using Accumulated Data" Proceedings of Computer Security Symposium 20111(CSS2011), pp. 456-461(2011).

[13] Alexa Top 500 Global Siteshttp://www.alexa.com/topsites

[14] Targeted Attack Trends 2014 Annual Reporthttps://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/reports/rpt-targeted-attack-trends-annual-2014-report.pdf

[15] Trendmicro Press http://www.trendmicro.co.jp/jp/about-us/press-releases/articles/20150409062703.html

[16] VirusTotal https://www.virustotal.com/

[17] LastLine https://www.lastline.com/

[18] RFC954 NICNAME/WHOIS https://www.ietf.org/rfc/rfc954.txt

[19] RFC3912WHOIS Protocol Specification http://www.ietf.org/rfc/rfc3912.txt

[20] User Local http://textmining.userlocal.jp/

[21] DOMAIN NAMES - CONCEPTS AND FACILITIES http://www.ietf.org/rfc/rfc1034.txt

[22] DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION http://www.ietf.org/rfc/rfc1035.txt

[23] Google https://www.google.com/

[24] V.Vapnik, A.Lerner, "Pattern recognition Using Generalized Portrait Method" Automation and Remote Control24, pp.774-780(1963).

[25] Multilayer Perceptronhttp://deeplearning.net/tutorial/mlp.html

[26] P.John, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines" Technical Report MSR-TR-98-14, pp.1-21(1998).

[27] R.Kohavi, "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection" Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12), pp.1137-1143 (1995).