Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries

Kazumichi Sato[1] keisuke Ishibashi[1] Tsuyoshi Toyono[2] Nobuhisa Miyake[1]

[1]*NTT Information Sharing Platform Laboratories, NTT Corporation*
{*sato.kazumichi, ishibashi.keisuke, miyake.nobuhisa*}*@lab.ntt.co.jp*

[2]*Internet Multifeed co.*
toyono@mfeed.ad.jp

Abstract

The Botnet threats, such as server attacks or sending of spam email, have been increasing. A method of using a blacklist of domain names has been proposed to find infected hosts. However, not all infected hosts may be found by this method because a blacklist does not cover all black domain names. In this paper, we present a method for finding unknown black domain names and extend the blacklist by using DNS traffic data and the original blacklist of known black domain names. We use co-occurrence relation of two different domain names to find unknown black domain names and extend a blacklist. If a domain name co-occurs with a known black name frequently, we assume that the domain name is also black. We evaluate the proposed method by cross validation, about 91 % of domain names that are in the validation list can be found as top 1 %.

1 Introduction

The Botnet [21, 5, 18] threats have been increasing. When infected hosts receive a command from a Command & Control (C&C) server, they launch a distributed denial-of-service (DDoS) attack, send spam email, and steal personal information [1, 22]. Moreover, a bot propagates through networks [7]. Therefore, network operators are required to find infected hosts in their networks to stop malicious activities.

To find infected hosts, methods of monitoring DNS traffic has been proposed [10, 4, 6]. When a bot attempts to connect to a C&C or malware-hosting server, it sends a DNS query to resolve domain names of these server. Therefore, given domain names of C&C or malware-hosting servers, we can find infected hosts by monitoring DNS traffic and finding the hosts that send queries of these domain names. In the rest of this paper, a black domain name denotes a domain name of a C&C or malware-hosting server. To stop malicious activities of bots, we should block connections from infected hosts to black domain names.

A blacklist can be created by analyzing bots. We need to capture bots to analyze them, and honeypots [2, 16, 14] have been proposed to capturing bots. A honeypot, e.g., operating system without applying security patches, is a vulnerable system. When a bot attempts to infect other hosts, it uses the vulnerabilities of the target host. Therefore, a honeypot can capture bots effectively.

However, it is hard to create a blacklist that covers all black domain names because of following reasons:

- The large number of new bots are observed in a day, honeypots can hardly capture all bots and we need the large amount of time to analyze bots.
- There is a bot that sends queries of many different black domain names (e.g., Conficker worm [15]), we can hardly take hold of all black domain names.

Therefore, we can not block all connections from infected hosts to black domain names and stop malicious activities with the blacklist. Our objective of study is to stop malicious activities that can not be stopped with the blacklist.

In this paper, we propose a method for finding unknown black domain names in order to stop malicious activities of bots. We focus on DNS queries sent by infected hosts. One bot may send several queries of black domain names because of a C&C server redundancy. Therefore, we assume as follows:

Assumption: If two different queries of domain names that are sent by many hosts exist and one is black, the other domain name is also black.

Using this assumption, we attempt to find unknown black domain names and extend a blacklist by using co-occurrence relation [11] between two different domain names. We define the co-occurrence relation between domain names as different domain names in queries sent

by the same host. If a domain name co-occurs with a known black domain name frequently, we assume that the domain name is an unknown black domain name. We define the score of a domain name by using co-occurrence relation with domain names of a given blacklist. However, when we applied the scoring method by using naive co-occurrence relation, we found that we did not classify domain names as black or not correctly. Therefore, we improved the scoring method by using weight of the number of hosts or domain names in queries sent by hosts. If the score of a domain name is large, we classify the domain name as black and extend the blacklist by adding the domain name to the original blacklist. Moreover, when we find hosts that send queries of unknown black domain names, we expect that we can find unknown infected hosts. Finding unknown infected hosts is our secondary objective of study.

We applied the proposed method to DNS traffic data and blacklist of known black domain names. The results show that we can validate our assumption and find unknown black domain names that are not in the blacklist to extend the blacklist. As a result, we can find unknown infected hosts by using the extended blacklist.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 describes the proposed method for finding unknown black domain names. Evaluation and experimental results are presented in Section 4. Section 6 summarizes our study.

2 Related work

In order to find botnets, many approaches of monitoring or analyzing DNS traffic data have been proposed. Our previous work [10] proposed method based on Bayes estimation [8, 20] for finding mass-mailing worm infected hosts by using a blacklist of known black domain names. This method calculate the black degree of a DNS query by using a ratio of the number of infected hosts that send the query or the number of non-infected hosts that send the same. Moreover, a black degree of a query sent by many infected hosts is large and a black degree of a query sent by many non-infected hosts is small. However, the number of non-infected hosts that send a query of an unknown black domain name may be greater than the number of infected hosts that send the same. In this case, a black degree of the unknown black domain name is small. Our proposed method may classify the domain as black because the method focuses on only infected hosts mainly. Therefore, we believe that our proposed method can find infected host more efficiently. In order to find botnets, Choi *et al.* [4] calculated the score of a domain name by comparing hosts that sent query of the domain name with hosts that sent the same query at a different time. However, in order to reduce calculation cost, if the number of hosts that send a query of a domain name is smaller than a threshold, a score of the domain name is not calculated and classified as legitimate. Therefore, this method may classify an unknown black domain name in query sent by a few hosts as legitimate. Our proposed method may find an unknown black domain name in a query sent by a few hosts because we do not determine this threshold.

In our previous work [9], we proposed a method for improving accuracy of a blacklist and finding unknown black domain names by using the DNS query graph that represents a relation between hosts and a domain name. A blacklist created by analyzing bots often includes legitimate domain names because bots send a query of legitimate domain name to confirm network connectivity. Therefore, we must remove legitimate domain names from a blacklist. Using this method with our proposed method, we expect that we can find more unknown infected hosts because legitimate domain names are removed from a blacklist and black domain names are added to the blacklist.

3 Approach

In this section, we describe our method of finding unknown black domain names and infected hosts to stop malicious activities of bots.

3.1 Method overview

In order to find unknown black domain names and infected hosts, we use DNS traffic data and a blacklist of known black domain names. In this paper, DNS traffic data denotes DNS user queries that hosts send to a DNS cache server. Monitoring DNS user queries, we can obtain source IP address of hosts and domain names. Our proposed method is to extend the blacklist and consists of following three steps: 1) Classify all hosts as infected or non-infected. 2) Find unknown black domain names in queries sent by infected hosts to extend the blacklist. 3) Find unknown infected hosts. Details of these steps follows:

Step 1: We classify all hosts H (set of source IP address) appeared in the traffic data as infected hosts H_I or non-infected hosts H_N by using a blacklist of known black domain names. If a host sends a query of domain name that is in the blacklist, the host is classified as H_I. Other is classified as H_N. An overview of this step is shown in Fig. 1.

Step 2: We find unknown black domain names in queries sent by H_I to extend the blacklist by using a method based on the co-occurrence relation. If a domain name d co-occurs with a black domain name frequently, we assume that d is an unknown black domain name. An

overview of this step is shown in Fig. 2, and a detail of this step is described in Section 3.2.

Step 3: Finally, we find unknown infected hosts. We find hosts which send a query of a black domain name found in the extended blacklist. An overview of this step is described in Figure 3.

3.2 Scoring method

As mentioned step 2, we attempt to find unknown black domain names in queries sent by H_I. However, a user of a infected host may use a DNS server (e.g., browsing web sites, sending email) or a bot to confirm network connectivity. Therefore, domain names in queries sent by H_I consist not only of black domain names but also legitimate domain names. Here, a legitimate domain name are not domain names of a C&C server or malware hosting server. We must be careful not to classify a legitimate domain name as black. Therefore, we need to define the criterion for classifying domain names as black or legitimate. The criterion is whether a score is higher or lower than a threshold. In this section, we describe a scoring method for classifying domain names.

First, we define a degree of co-occurrence relation based on Jaccard index [11] between two domain names $C(d_i, d_j)$ as follows:

$$C(d_i, d_j) = \frac{\sum_{\{h \mid d_i \in D_h \wedge d_j \in D_h\}} 1}{\left|\{h \mid d_i \in D_h \vee d_j \in D_h\}\right|}. \quad (1)$$

Here, H denotes all hosts, D denotes a set of domain names, and D_h denotes domain names in queries sent by $h \in H$. The numerator of Equation (1) represents co-occurrence frequency between d_i and d_j, the denominator of Equation (1) represents the total number of hosts that send queries of d_i or d_j. $C(d_i, d_j)$ represents co-occurrence rate between d_i and d_j. If $C(d_i, d_j)$ is large, we assume that the relation between d_i and d_j is strong.

Next, we define a score S of a domain name by using Equation (1). When D_B denotes black domain names, the score of a domain name $d \in D$ is described as follows:

$$S(d) = \sum_{d_b \in D_B} C(d_b, d). \quad (2)$$

$C(d_b, d)$ represents the degree of co-occurrence relation between a domain name and a black domain name defined by Equation (1). Therefore, if $S(d)$ is large, a domain name d is related to many black domain names, it represents a black degree for domain names. We assume that a high-scored domain name is a black domain name.

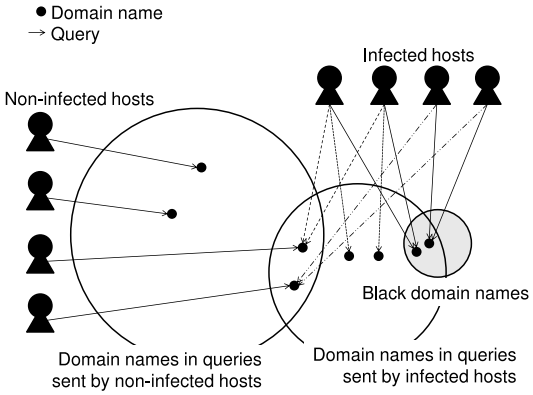

Figure 1: Classifying hosts

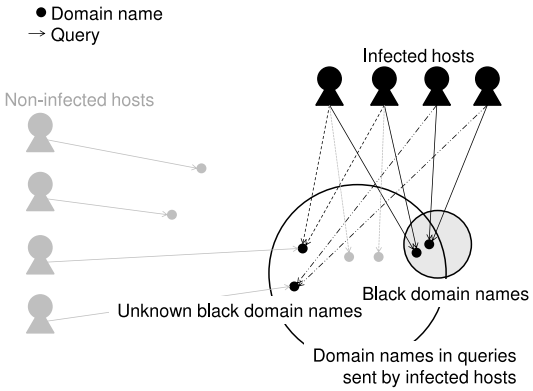

Figure 2: Finding unknown black domain names

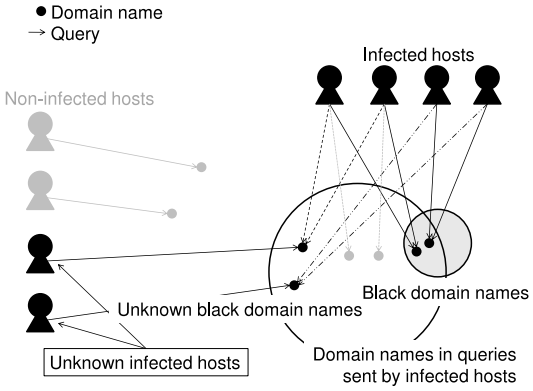

Figure 3: Finding unknown infected hosts

3.3 Problematic domain names

As mentioned above, we define the black degree for a domain name as S. However, when we apply the scoring method S based on co-occurrence relation to a DNS traffic and a blacklist in order to find unknown domain names, we may find that some high-scored legitimate domain names. Therefore, we need to improve scoring method that is defined in Section 3.2. We describe types of high-scored legitimate domain names as follows:

Popular domain names: Popular domain names, e.g., www.google.com, are those in queries sent by many hosts (Fig. 4). A user of a infected host sends queries of these domain names to a DNS server to browse or sends email, or a bot may send them to confirm network connectivity. Therefore, a popular domain name may co-occur with a black domain name frequently, and thus S of a popular domain name may be large.

Domain names in queries sent by infected heavy user: A heavy user is a host that sends many queries to a DNS server. When we monitor DNS traffic data and use a blacklist, we may find infected heavy users (Fig. 4). Due to these users, many domain names may co-occur with a black domain name. However, not all domain names in queries sent by an infected heavy user are black.

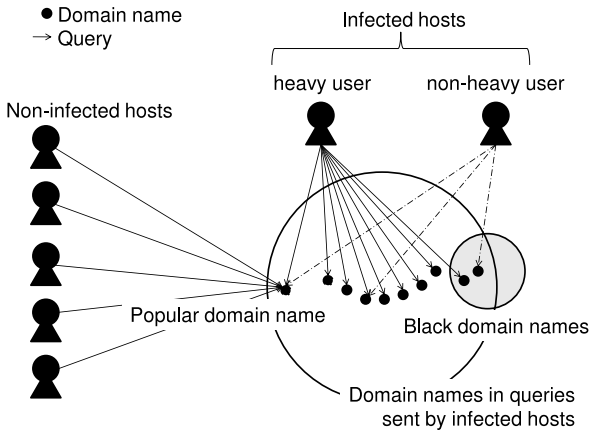

Figure 4: Heavy user

Because these problematic domain names impact the accuracy of classifying black domain names, we must eliminate their influence. In the rest of this section, we describe methods for doing so.

3.3.1 Improving score of a popular domain name

We need to reduce the score of popular domain names to prevent classifying those names as black. Here, we consider the popularity of domain names from the perspective of infected or non-infected hosts. While legitimate popular domain names are popular with both infected and non-infected hosts, black domain names are popular with only infected hosts. Therefore, we focus on the number of non-infected hosts that send a query of a domain name.

The weight of a domain name $W(d)$ is described as follows:

$$W(d) = \frac{\left|\{h \mid h \in H_I \wedge d \in D_h\}\right|}{\left|\{h \mid h \in H \wedge d \in D_h\}\right|}. \quad (3)$$

Equation (3) represents a ratio of the number of infected hosts that send a query of a domain name to the number of all hosts the send the query. Weights of popular domain names $W(d_p)$ are smaller than weights of black domain names $W(d_b)$ because a numerator of $W(d_b)$ is greater than a numerator of $W(d_p)$.

Now, we define a weighted score S_w by using Equation (3).

$$S_w(d) = S(d) \times W(d). \quad (4)$$

Using S_w, we can reduce the scores of popular domain names. An overview of this is shown in Fig. 5.

According to past surveys of the ratio of infected hosts to non-infected hosts, a ratio is about 1% [3]. Therefore, if a score of popular domain name $S(d)$ is high, $W(d)$ is very small because the denominator is much greater than the numerator. As a result, the weighted score of the popular domain name $S_w(d)$ is very small.

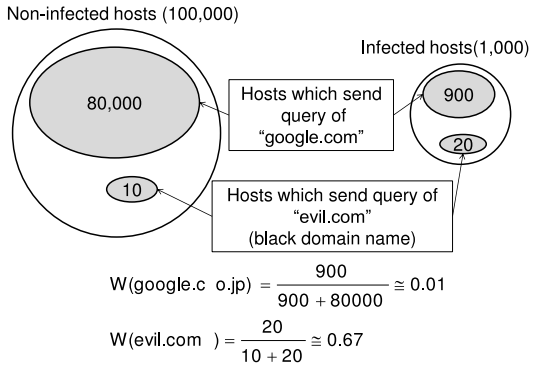

Figure 5: Overview of reducing score of popular domain name

3.3.2 Improving score of domain name in query sent by infected heavy user

We need to reduce the score of a domain name in a query sent by an infected heavy user to prevent classifying this domain name as black. Here, we consider two cases of a relation between a domain name d and a known black domain name d_b. One is a relation between d and d_b sent by a heavy user, the other is a relation between d and d_b sent by a non-heavy user. We assume that the former relation is clearly weaker than the latter. However, calculating a relation between d_i and d_j by using C based on the naive co-occurrence relation, increment of C is the same in each case. As a result, a score S of a legitimate domain name in a query sent by an infected heavy user is large.

Therefore, we focus on the number of domain names sent by a host and define weighted degree of co-occurrence relation between two domain names $C'(d_i, d_j)$ as follows:

$$C'(d_i, d_j) = \frac{\sum_{\{h \mid d_i \in D_h \wedge d_j \in D_h\}} 1/\left|D_h\right|}{\left|\{h \mid d_i \in D_h \vee d_j \in D_h\}\right|}. \quad (5)$$

The numerator of this equation is weighted by the number of domain names in queries sent by a host ($|D_h|$). Therefore, if a heavy infected user send queries of a domain name and a black domain name, co-occurrence frequency may fairly increase. Using Equation (5), we define a new score of the domain name $S'(d)$ as follows:

$$S'(d) = \sum_{d_b \in D_B} C'(d_b, d). \quad (6)$$

Using Equation (6), we can reduce a score of a domain name in a query sent by an infected heavy user. An overview of this is shown Fig. 6.

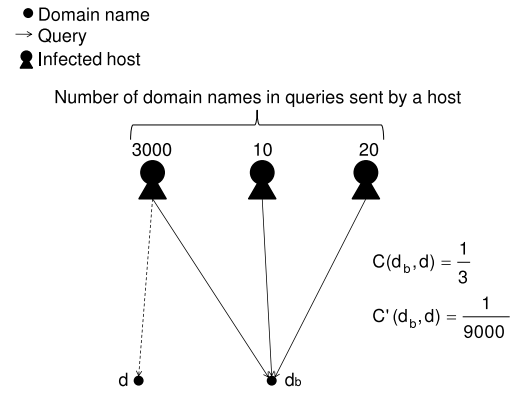

Figure 6: Overview of reducing score of domain name in query sent by infected heavy user

3.3.3 Improved Scoring method

Using Equations (3) and 5, we can eliminate problems related to popular domain names and infected heavy users. Finally, we define a score S'_w as follows:

$$S'_w(d) = \left(\sum_{d_b \in D_B} C'(d_b, d) \right) \times W(d). \quad (7)$$

We evaluate the effectiveness of these scoring methods (Equations (2), (4), and (7)) in next section.

4 Evaluation

In this section, we describe evaluation of our proposed method.

4.1 Data set

In order to evaluation our proposed method, we used DNS traffic data captured during an hour on February, 2009. We also used a blacklist of about 270 known domain names create by a honeypot during the same period.

4.2 Validation of correctness of our assumption

Our proposed method is based on the assumption in Section 1. We used 10-fold cross validation [12] to validate this assumption. Firstly, we split a blacklist into 10 lists, we create a learning list by using $9/10$ of the lists and a validation list by using $1/10$ of the rest. We can create ten types of learning and validation lists. Secondly, we applied our proposed method to DNS traffic data and each learning list and took average of each scores of ten validations. Finally, we extracted domain names in top n % of average scores and calculated each ratio of the number of these domain names in the validation list to the total number of domain names in the validation list. If our assumption is correct, this ratio will be high. The results of the validation are shown in Fig. 7.

Fig. 7 shows that each ratio of the number of domain names in top n % of scores to the total number domain names in validation list. This result shows that scoring method S'_w can find unknown black domain names effectively. Moreover, the greatest difference of effectiveness is appeared in about top 1 % of scores. Table 1 shows that a ratio of the number of domain names in top 1 % of scores.

The result shows that when we score a domain name by using naive scoring method S, we find only about 23 % of domain names in validation list. In Section 3, we describe that score of problematic domain names is

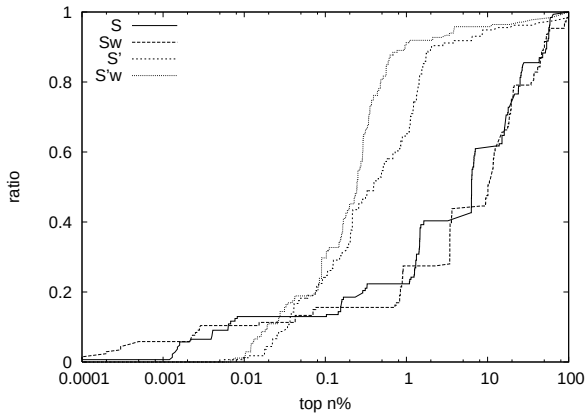

Figure 7: A ratio of # of domain names in validation list to # of domain names in top n % score

Table 1: Ratio of number of domain names in validation list to number of domain names in top 1 % of scores

Heavy hosts \ Popular domain	Not reduce	Reduce
Not Reduce	23.0 %(S)	27.4 %(S_w)
Reduce	65.2 %(S')	91.2 %(S'_w)

large. Therefore score of domain names in validation list is relatively small. As a result, the ratio is also small.

While we find about 27 % of domain names in validation list by using the scoring method S_w which reduce scores of popular domain names only, we find about 65 % of domain names in validation list by using the scoring method S' which reduce scores of domain names in queries sent by heavy infected hosts only. This result shows that when we use the scoring method S, many high-scored domain names are in queries sent by infected heavy hosts. Moreover, we find that only a few infected heavy hosts send queries of these domain names. When we score a domain name by using the scoring method S'_w which reduce scores of popular domain names and domain names in queries sent by infected heavy hosts, we find about 91 % of domain names in the validation list. This result shows that if we reduce scores of problematic domain names, we can find unknown black domain names.

4.3 Validation of effectiveness of proposed method

As mentioned Section 4.2, we found that the scoring method S'_w was most effective to find unknown black domain names. Therefore, we applied S'_w to DNS traffic data and all known black domain names to validate our proposed method. We classified high-scored domain names as black, legitimate, unclear by using a web search engine. If information sites about threats [13] are included in the search results for a domain name, we classify the domain name as black, and if web sites, e.g., company sites, are include in the search results for a domain name, we classify the domain name as legitimate. Other domain names are classified as unclear.

When we classified domain names that have the top 100 score by using a web search engine, we found that these domain names consisted of 39 % black, 4 % legitimate, and 57 % unclear. In domain names classified as black, there were domain names including malformed characters. If bots sends queries of these domain names to DNS server, responses from DNS server may be NXDOMAIN (no such domain). Therefore, the bots can not connect C&C server and perform malicious activities. However, our secondary objective of study is to find infected hosts. We thus classify these domain names as black to find infected hosts. Detail of these domain names are described as follows:

<black domain names>:<port number>: We found, e.g., "`helsinki.fi.eu.undernet.org:6669`", domain names including port number. Moreover, we found that all port number appeared in these domain names were used for Internet Relay Chat (IRC) protocol. When a C&C server attempt to send command to a bot, IRC protocol is often used. Therefore, we assumed that the left side of ":" was a domain name of a C&C server.

<black domain name>/<directory name>: We found a domain name that was like URL ("`google-analitucs.com/loader/`"). When we examined "google-analitucs.com" by using a web search engine, we found that the domain name was black. Therefore, we assumed that "`google-analitucs.com/loader/`" was black.

We believe that unknown black domain names are included in unclear domain names. When we examined unclear domain names, we found that there were suspicious domain names in unclear as follows:

1. Domain names whose subdomain differ from a subdomain of a known black domain name:
We found, e.g., "`china.alwaysproxy.info`", domain names whose subdomain differ from a subdomain of a known black domain name. For example, "`{newss|ofat|ports}.alwaysproxy.info`" are known black domain names. As mentioned Section 1, one bot may be send several queries of black domain names. Therefore, we assumed that a domain name of

this case was one of domain names of C&C servers. However, we must be careful, e.g., "xxx.3322.org" domain names by using hosting service. There are many subdomain of domain names by using hosting service. Therefore, if "xxx.3322.org" is black, "yyy.3322.org" may not be black. We did not classify these domain names in unclear as black.

2. <black domain name>.<legit domain name>: We found, e.g., "www.h7smcnrwlsdn34fgv.info.<legit domain name>", domain names that a legit domain name followed a known black name. In this case, we found that all black domain names that were a left side of the above format were domain names that the bot (Sality.Q) attempt to connect.

3. Domain names in queries for DNSBL lookups: DNSBL (DNS Blacklist) [17] is a DNS-based database consisting of malicious IP address and often used for spam filter. Sending a query of a domain name including IP address of a sender to a DNS server, we can check whether received email is spam or not. We found, e.g., "<IP address>.zen.spamhaus.org", domain names in queries for DNSBL lookups. Bots may send queries for DNSBL lookups to check that the bots themselves are in blacklist [19]. Therefore, a possibility exists that these domain names are black domain names.

In this validation, we classified only suspicious domain names in case 1 as black. A possibility exists that these domain names are true unknown black domain names because these domain names are not appeared in search results. Domain names in case 2 and 3 also may be true unknown black domain names, we will validate whether these domain names are black or legit in our future work. Table 2 shows the classification results for domain names that have the top 20 score. These results show that 80 % of the domain names in top 20 scores are black domain names that are not in the given blacklist and that no legitimate domain name is included in top 20. This indicates that our proposed method can find unknown black domain names.

Our secondary objective of study was to find unknown infected hosts. Therefore, we created an extended blacklist by adding unknown black domain names that had the top 100 scores to original blacklist. We then compared the number of infected hosts found by using the extended blacklist with the number of hosts found by using the original blacklist. The results showed that the number of infected hosts found by using the extended blacklist was 3 % higher than the number found by using the original blacklist. We consider that this rate is not enough to achieve our secondary objective of study.

Table 2: Domain names in top 20 scores

Score	Domain name	Evaluation
0.571	spy.nerashti.com	Black
0.571	bla.bihsecurity.com	Black
0.571	aaaaaaaaaaaaa.locop.net	Black
0.500	icq-msg.com	Black
0.319	mail.tiktikz.com	Black
0.300	x.zwned.com	Black
0.300	evolutiontmz.sytes.net	Unclear
0.300	dcom.anxau.com	Black
0.292	usa.lookin.at	Unclear
0.292	rewt.buyacaddi.com	Black
0.250	unkn0wn	Unclear
0.250	google-analitucs.com/loader/	Black
0.222	netspace.err0r.info	Unclear
0.203	win32.kernelupdate.info	Black
0.203	free.systemupdates.biz	Unclear
0.200	zjjdtc.3322.org	Black
0.200	ykln.3322.org	Unclear
0.200	dr27.mcboo.com	Black
0.189	china.alwaysproxy.info	Black
0.167	home.najd.us	Black

5 Discussion

In this section, we discuss a possibility of countermeasure against our proposed method, as well as the evaluation results for our two objectives; finding unknown black domain names and finding unknown infected hosts.

As mentioned Section 4, our proposed method can find black domain names effectively. However, if bots send the large number of queries of legitimate domain names, our proposed method can be polluted. This is because we reduce co-occurrence frequency between a domain name and a black domain name in queries sent by a infected heavy host in Equation (5). Even though a domain name co-occurs with known black domain names frequently, the score of the domain name is small and classified as legitimate when many bots send the large number of queries. Considering these cases, we will improve our proposed method in our future work.

As for our primary objective, finding unknown domain names, our experimental results shows that a bot send several queries of black domain names which is not included in original blacklist. Therefore, we can stop malicious activities of bots by using extended blacklist more effectively because extended blacklist includes more domain names of C&C servers than original blacklist. However, it remains that we have 4 % of false positives (legitimate domains) when we manually inspect domain names whose scores are in top 100. Though we can not inspect all the domain names of top 1 % (1600) do-

main names with which we can achieve 91.2 % coverage, the false positive ratio may increase. It is a future work to tune the threshold on how many high-scored domains to be used to find unknown infected hosts.

As for our secondary objective, finding unknown infected hosts by using extended blacklist, we found that the number of unknown infected hosts is relatively small compared to the number of unknown black domain names. While a rate of increase in the number of unknown domain names is 18 %, a rate of increase in the number of unknown infected hosts is 3 %. In this paper, we focus on domain names in queries sent by infected hosts to extend blacklist. Therefore, although unknown black domain names can be found, most hosts which send those domain names are known infected hosts found by using original blacklist. We should improve our proposed method to find unknown infected hosts more effectively and achieve secondary objective of study.

6 Conclusion

In this paper we proposed a method for finding unknown black domain names and infected hosts by using DNS traffic data and a blacklist of known black domain names. Our proposed method based on the co-occurrence relation extends the blacklist. Using the extended blacklist, we can find unknown infected hosts.

We applied our proposed method to DNS traffic data and a blacklist. The results of cross validation show that about 91 % of domain names that are in the validation list can be found. The results of effectiveness of the method show that the number of hosts found by the extended blacklist is 3 % higher than the number found by the original blacklist.

We will set a threshold to determine whether a domain name is black or legit and improve our proposed method to find more unknown infected hosts in our future work.

References

[1] U. Bayer, I. Habibi, D. Balzarotti, E. Krida, and C. Kruege. A View on Current Malware Behaviors. In *Proceedings of the 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats*, Boston, MA, April 2009.

[2] P. Beacher, M. Koetter, T. Holz, M. Dornseif, and F. Freiling. The Nepenthes Platform: An Efficient Approach To Collect Malware. In *Proceedings of the 9th International Symposium on Recent Advances In Intrusion Defection*, Hamburg, Germany, September 2006.

[3] Cyber Clean Center. Fiscal Year 2008 Activity Report on Cyber Clean Center. `https://www.ccc.go.jp/en_report/h20ccc_en_report.pdf`, 2009.

[4] H. Choi, H. Lee, H. Lee, , and H. Kim. Botnet Detection by Monitoring Group Activities in DNS Traffic. In *Proceedings of the 7th IEEE International Conference on Computer and Information Technology*, Washington, DC, October 2007.

[5] E. Cooke, F. Jahanian, and D. McPherson. The Zombie Roundup: Understanding, Detecting, and Disrupting Botnets. In *Proceedings of the Steps to Reducing Unwanted Traffic on the Internet Workshop*, Cambridge, MA, July 2005.

[6] D. Dagon. Botnet detection and response. In *OARC Workshop*, 2005.

[7] D. Dagon, C. Zou, and W. Lee. Modeling Botnet Propagation Using Time Zones. In *Proceedings of the 13th Annual Network and Distributed System Security Symposium*, San Diego, CA, February 2006.

[8] P. Graham. A Plan for Spam. `http://www.paulgraham.com/spam.html`.

[9] K. Ishibashi, T. Toyono, and M. Iwamura. Improving accuracy of black domain list by using DNS query graph. In *Proceedings of 1st Internet Workshop on Information Network Design*, Fukuoka, Japan, December 2008.

[10] K. Ishibashi, T. Toyono, K. Toyama, M. Ishino, H. Ohshima, and I. Mizukoshi. Detecting Mass-Mailing Worm Infected Hosts by Mining DNS Traffic Data. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data*, Philadelphia, PA, August 2005.

[11] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[12] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of 40th International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada, August 1995.

[13] Trend Micro. Virus Information. `http://threatinfo.trendmicro.com/vinfo`.

[14] J. Nazario. PhoneyC: A Virtual Client Honeypot. In *Proceedings of the 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats*, Boston, MA, April 2009.

[15] P. Porras, H. Saïdi, and V. Yegneswara. A Foray into Conficker's Logic and Rendezvous Points. In *Proceedings of the 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats*, Boston, MA, April 2009.

[16] The Honeynet Project. `http://www.honeynet.org`.

[17] The Spamhaus Project. `http://www.spamhaus.org`.

[18] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. A Multifaceted Approach to Understanding the Botnet. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, Rio de Janeriro, Brazil, October 2006.

[19] A. Ramachandran, N. Feamster, and D. Dagon. Revealing Botnet Membership Using DNSBL Counter-Intelligence. In *Proceedings of the 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet*, Can Jose, CA, July 2006.

[20] G. Robinson. A Statistical Approach to the Spam Problem. *Linux Journal*, 2003(107), March 2003.

[21] S. Staniford, V. Paxson, and N. Weaver. How to Own the Internet in Your Spare Time. In *Proceedings of the 11th USENIX Security Symposium*, San Francisco, CA, August 2002.

[22] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar. Characterizing Botnets from Email Spam Records. In *Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats*, San Francisco, CA, April 2008.