社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

DNS based Entropy and Forensic Analysis on the PCs for Learners in a University

Dennis A. LUDEÑA ROMAÑA†, Shinichiro KUBOTA††, Kenichi SUGITANI††, and Yasuo MUSASHI††

† Graduate School of Science and Technology, Kumamoto University　2–39–1, Kurokami, Kumamoto, 860–8555 Japan

†† Graduate School of Science and Technology, Kumamoto University　2–39–1, Kurokami, Kumamoto, 860–8555 Japan

Abstract　We performed an entropy study on the DNS query traffic from the outside of a university campus network to the top domain DNS server when querying about reverse resolution on the PCs for learners through January 1st, 2007 to February 29th, 2008. The following interesting results are given: (1) The total DNS query traffic changes in a mild manner until January 16th, 2008, however it drastically changes after January 17th, 2008. (2) In January 17th, 2008, the DNS query traffic is mainly dominated by several specific IP addresses as their query keywords. (3) We carried out forensic analysis on the PCs for learners in which IP addresses are found in the several specific keywords and it is concluded that the PCs become spam bots when inserting USB based key disk storage.

Key words　DNS based Detection, Spam Bots, Entropy, DNS traffic

1. Introduction

It is of considerable importance to raise up a detection rate of spam bots (SBs), since they become components of the bot networks that are used to send a lot of unsolicited mails like spam, phishing, and mass mailing activities and to execute distributed denial of service attacks[1-6].

Recently, Wagner *et al.* reported that entropy based analysis was very useful for anomaly detection of the random IP and TCP/UDP addresses scanning activity of internet worms (IWs) like an W32/Blaster or an W32/Witty worm, respectively, since the both worms drastically changes entropy when after starting their activity[7].

Previously, we reported that the DNS query keywords based entropy in the DNS query packet traffic from the outside of the campus network decreases considerably while the source IP addresses based entropy increases when the spam bots activity is high in the campus network[8]. This is probably because the spam bots activity can be easily to be sensed by the spam filter and/or the IDS/IPS on the internet. Therefore, we can detect spam bots activity on the campus network, by only watching the DNS query packet traffic from the other sites on the internet.

Also, we recently reported that in the MX resource record based DNS query packet traffic from the inside of the campus network, we observed two types of changes in the DNS query keywords and the source IP addresses based entropies[9]. In the former one, the source IP addresses and the DNS query keywords based entropies decreases when the targeted spam bots activity is high, while in the latter one, the source IP addresses based entropy decreases but the DNS query keywords based one increases when the random spam bots activity is high. Therefore, we can detect a type of spam bots activity on the campus network, by only watching the DNS query packet traffic from the campus network. However, it is likely that we can find no entropy study on the PTR resource record (RR) based DNS query packet traffic. In this paper, (1) we carried out statistical and entropy analysis on the total- and the PTR RR-based DNS query packet traffics from the outside of the campus network, (2) we discuss on the difference in the entropy analysis between the total- and the PTR RR-based DNS query packet traffics, and (3) on the detected spam bots in the PC for learners.

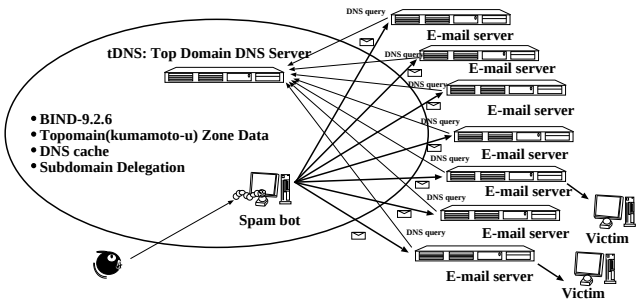

Fig. 1. A schematic diagram of a network observed in the present study.

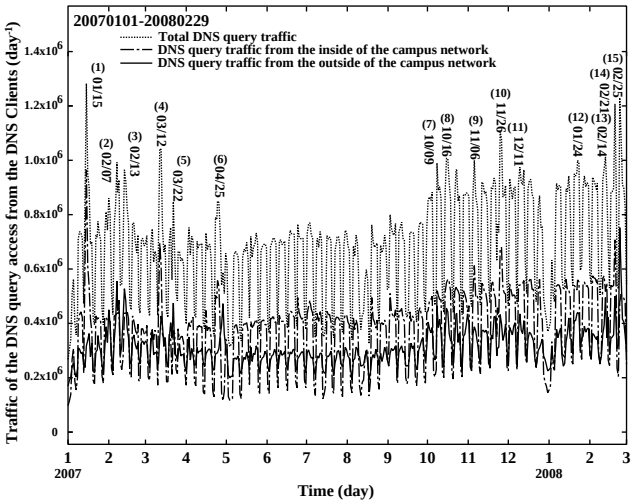

Fig. 2. Traffic of the DNS query packets to the top domain DNS server (**tDNS**) and the traffic from the inside- and the outside-DNS clients in a university through January 1st, 2007 to February 29th, 2008 (day^{-1} unit).

2. Observations

2. 1 Network System

We investigated traffic of DNS query accesses between the top domain DNS server (**tDNS**) and the DNS clients. Figure 1 shows an observed network system in the present study and optional configuration of the BIND-9.2.6 DNS server program daemon [10] of the **tDNS** server. The **tDNS** server is one of the top level domain name (kumamoto-u) system servers and plays an important role of domain name resolution and subdomain name delegation services for many PC clients and the subdomain networks servers, respectively, and the operating system is Linux OS (CentOS 4.3 Final) in which kernel-2.6.9 is currently employed with the Intel Xeon 3.20 GHz Quadruple SMP system, the 2GB core memory, and Intel 1000Mbps EthernetPro Network Interface Card.

2. 2 DNS Query Packet Capturing

In **tDNS**, BIND-9.2.6 program package has been employed as a DNS server daemon [10]. The DNS query packets and their query keywords have been captured and decoded by a query logging option (Figure 1, see % man named.conf in more detail). The log of DNS query access has been recorded in the syslog files. All of the syslog files are daily updated by the crond system. The line of syslog message mainly consists of the content of the DNS query packet like a time, a source IP address of the DNS client, a fully qualified domain name (A and AAAA resource record (RR) for IPv4 and IPv6 addresses, respectively) type, an IP address (PTR RR) type, and a mail exchange (MX RR) type.

2. 3 DNS Query Traffic

Firstly, we can demonstrate the total DNS query traffic from the inside and outside of the campus network through January 1st, 2007 to February 29th, 2008, as shown in Figure 2.

In Figure 2, the six large peaks are observed through the first and second quarters and the nine peaks are found through the third and last quarter in 2007 and the first half quarter in 2008. The former six peaks have been grouped, as the first group consists of (1) January 15th and (4) March 12th, 2007, the second one consists of (2) February 7th and (3) 13th, (6) April 25th, 2007, and the last one consists of (5) March 22nd, 2007. The first group {(1),(4)} is fixed to be the big domain name resolution traffic from the campus network, in which the DNS query traffic were generated by crashes of the NIS-based authentication systems in the campus network. The second group {(2),(3),(6)} is taken place by the DNS misconfiguration. And the last one {(5)} is based on spam bots activity.

The latter nine peaks are assigned to be two groups. The first group consists of (7) October 9th, (8) 16th, (9) November 6th, (10) 26th, (11) December 12th, 2007, (13) February 14th, (14) 21st, and (15) 25th, 2008 and the other group is (12) January 24th, 2008. In the first group {(7),(8),(9),(10),(11),(13),(14),(15)}, we observed the high frequencies for several specific query keywords like IP addresses and fully qualified domain names relating with the spam bots and the local E-mail servers. In the other group {(12)}, we observed unexpectedly several query keywords that are belonging to the PCs for learners in the DNS query traffic from the outside of the campus network.

From these results, we further carried out entropy analysis on the total DNS query traffic, the PTR resource record (RR) based DNS query traffic, and the PTR RR based DNS query traffic including only the subnetwork addresses of PCs for learners from the outside of the campus network.

2. 4 Estimation of Entropy

We employed Shannon's function in order to calculate entropy $H(X)$, as

$$H(X) = -\sum_{i \in X} P(i) \log_2 P(i) \qquad (1)$$

where X is the data set of the frequency $freq(j)$ of IP addresses or that of the DNS query keywords in the DNS query packet traffic from the campus network, and the probability $P(i)$ is defined, as

$$P(i) = \frac{freq(i)}{\sum_j freq(j)} \qquad (2)$$

where i and j $(i, j \in X)$ represent the source IP address or the DNS query keyword in the DNS query packet, and the

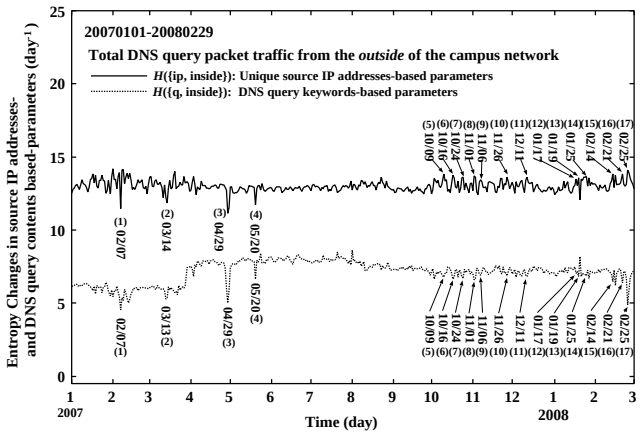

Fig. 3. Entropy changes in the total DNS query packet traffic from the outside of the campus network to the top domain name system (**tDNS**) server through January 1st, 2007 to February 29th, 2008. The solid and dotted lines show the source IP addresses and DNS query keywords based entropies, respectively (day^{-1} unit).

frequency $freq(i)$ are estimated with the following script program:

```
#!/bin/tcsh -f
cat querylog | grep "client 133\.95\." |\
tr '#' ' ' | awk '{print $7}' |\
sort -r | uniq -c | sort -r >freq-sIPaddr
cat querylog | grep "client 133\.95\." |\
awk '{print $9}' | sort -r | uniq -c |\
sort -r >freq-querycontents
```

Chart 1

where "querylog" is a syslog file including syslog messages of the BIND-9.2.6 DNS server daemon program[10]. The syslog message (one line) consists of keywords as "Month", "Day", "hours:minutes:seconds", "server name", "named[process identifier]:", "client", "source IP address#source port address:", "query:", and "a DNS query keyword". This script program consists of three program groups: (1) The first program group is a first line only including "#!/bin/tcsh -f" means that this script is a TENEX C Shell (tcsh) coded script programs. (2) The second program group estimates frequencies of the unique source IP addresses, consisting of of unix commands from "cat" to "sort -r" because the back slash "\" connects the line terminated by "\" with the next line in the tcsh program. In this program group, the "cat" shows all the syslog message-lines from the syslog file "querylog", the "grep -v" command extracts only the message-lines excluding the source IP address of "133.95.x.y", the "tr" replaces a character '#' with a white space ' ', the unix command "awk '{print $7}"' extracts only a seventh keyword as "source IP address" in the message-line, the "sort -r | uniq -c | sort -r" commands sort the dataset of "source IP addresses" into the dataset of "unique source IP addresses" and estimate the frequencies of the unique source IP addresses and the final results are written into the file "freq-sIPaddr". (3) The last program group extracts the DNS query keywords from the syslog message-lines, sorts the dataset of "DNS query keywords" into the dataset of "unique DNS query keywords" and estimates the frequencies of the unique DNS query keywords. Finally, the results of the last program group are written the file into "freq-querycontents". In the last program group, although almost the commands, arguments, and their options take the same as the second program group, the unix command "tr" and its arguments are removed and a new argument " '{print $9}' " replaces the arguments of the unix command "awk" in the second program group.

3. Results and Discussion

3. 1 Entropy Analysis on DNS Query Traffic from the Outside of the Campus Network

We illustrate the calculated the source IP addresses and the query keywords based entropies in the total DNS query packet traffic from the outside of the campus network to the top domain name system **tDNS** server through January 1st, 2007 to February 29th, 2008, as shown in Figure 3.

In Figure 3, we can observe significant peaks of (1) February 7th, (2) March 13th and 14th, (3) April 29th, (4) May 20th, (5) October 9th, (6) 16th, (7) 24th, (8) November 1st, (9) 6th, (10) 26th, (11) December 11th, 2007, (12) January 17th, (13) 19th, (14) 25th, (15) February 14th, (16) 21st, and (17) 25th, 2008.

Expectedly, we have already observed the same peaks (1), (5), (6), (9), (10), (11), (15), (16), and (17) in Figure 3 corresponding to (2), (7), (8), (9), (10), (11), (13), (14), and (15) in Figure 2, respectively in which these peaks are fixed to several spam bots activities. Interestingly, on the other hand, we can find new peaks (2), (3), (4), (7), (8), (12), (13), and (14) in Figure 3.

These features show that entropy analysis on the DNS query packet traffic is useful for extracting the hidden security incidents in the DNS query packet traffic from the outside of the campus network, *i.e.* the entropy analysis has a possibility which can raise up the detection rate of security incidents in the DNS query packet traffic.

Also, in Figure 3, almost all the peaks are simply assigned to usual spam bots activity because almost the same IP addresses or fully qualified domain name (FQDN) of the local vulnerable E-mail servers. However, the several peaks (12), (13), and (14) are very difficult to identify what kinds of spam bots since the detected IP addresses are variable daily and/or hourly. Fortunately, the detected IP addresses in the peaks are easily identified because they are belonging to the

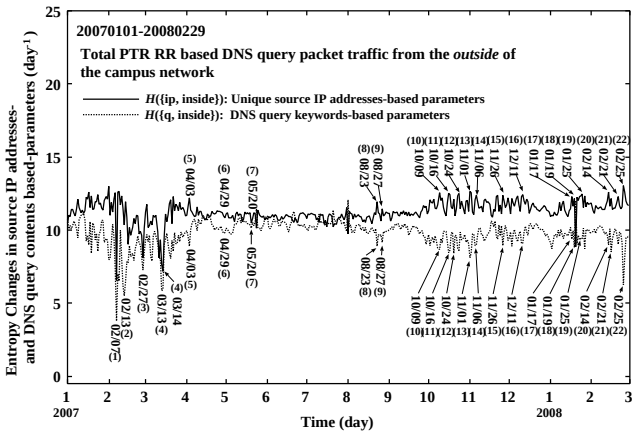

Fig. 4. Entropy changes in the total PTR resource record (RR) based DNS query packet traffic from the outside of the campus network to the top domain name system (**tDNS**) server through January 1st, 2007 to February 29th, 2008. The solid and dotted lines show the source IP addresses and DNS query keywords based entropies, respectively (day^{-1} unit).

authors administrated specific subnet addresses.

3.2 Entropy Analysis on PTR RR-DNS Query Traffic from Outside of Campus Network

We performed entropy analysis on the PTR resource record (RR) based DNS query packet traffic (reverse name resolution traffic) from the outside of the campus network through January 1st, 2007 to February 29th, 2008 (Figure 4).

In Figure 4, we can find interesting peaks of (1) February 7th, (2) 13th, (3) 27th, (4) March 13th and 14th, (5) April 3rd, (6) 29th, (7) May 20th, (8) August 23rd, (9) 27th, (10) October 9th, (11) 16th, (12) 24th, (13) November 1st, (14) 6th, (15) 26th, (16) December 11th, 2007, (17) January 17th, (18) 19th, (19) 25th, (20) February 14th, (21) 21st, and (22) 25th, 2008.

And these peaks are categorized into two types, as: {(1), (2), (3), (4), (6), (7)} and {(5), (8), (9), (10), (11), (12), (13), (14), (15), (16), (17), (18), (19), (20), (21), (22)}. In the former group, the source IP addresses and the DNS query keywords based entropies decrease simultaneously. This feature means that the spam bots attacks only to the specific E-mail serves on the internet. In the latter group, on the other hand, the source IP addresses based entropy increases but the DNS query keywords based one decreases. This shows that the spam bots attack randomly targeted E-mail servers on the internet.

Previously, we reported the similar insights for entropy analysis in the MX RR based DNS query packet traffic from the campus network[9] in which there are two types of spam bots; *random spam bots* (RSB) and *target spam bots* (TSB).

Note that also in Figure 4, the new peaks (2), (3), (5), (7), and (8) are added and the peaks in the both entropy curves become more clear and sharp than those in Figure 3.

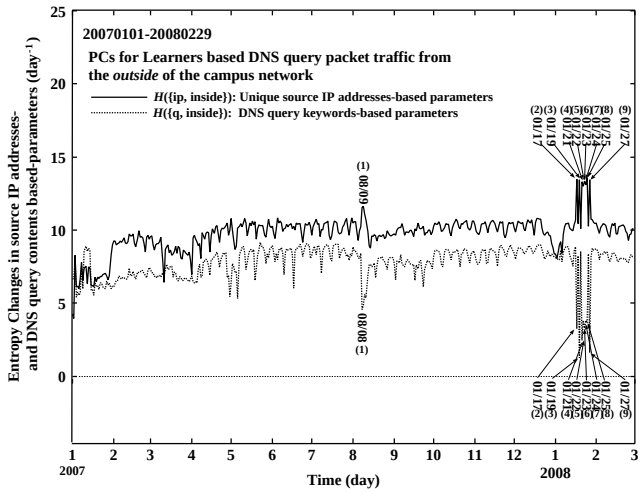

Fig. 5. Entropy changes in the DNS query packet traffic including the IP addresses of PCs for learners as query keywords from the outside of the campus network to the top domain name system (**tDNS**) server through January 1st, 2007 to February 29th, 2008. The solid and dotted lines show the source IP addresses and DNS query keywords based entropies, respectively (day^{-1} unit).

3.3 Entropy Analysis on DNS Query Traffic including IP addresses of PCs for Learners

We demonstrate the calculated the source IP addresses and the query keywords based entropies in the PTR resource record (RR) based DNS query packet traffic including only the IP addresses of PCs for learners as their query keywords from the outside of the campus network to the top domain name system **tDNS** server through January 1st, 2007 to February 29th, 2008, as shown in Figure 5.

In Figure 5, we can observe several interesting peaks of (1) August 9th, 2007, (2) February 17th, (3) 19th, (4) 21st, (5) 22nd, (6) 23rd, (7) 24th, (8) 25th, and (9) 27th, 2008.

Currently, the peak (1) is unknown but probably fixed to DNS misconfiguration in the specific home directories server system for the university students.

In the peaks (2)-(9), we carried out statistics on the query keywords in the total PTR RR based DNS query packets traffic at February 17th, 2008 (the peak (1)) and the results are shown, as follows:

IPv4 address	Frequency/day
133.95.a1.173	11,263
133.95.a2.181	2,359
133.95.**.1	1,943
133.95.***.103	1,761
133.95.***.11	1,737
133.95.**.1	1,721
133.95.**.209	1,623
133.95.**.3	1,538
133.95.**.100	1,317
133.95.**.100	1,224

where the above top IP addresses are obtained when the fre-

quency takes more than 1,000/day.

Surely, we can obtained a couples of the top IP addresses of 133.95.a1.173 and 133.95.a2.181,in which the both IP addresses are assigned to the PCs for learners-subnetwork addresses, 133.95.a1.0/24 and 133.95.a2.0/24, respectively.

After the peak (1), we also performed statistics on the query keywords in the total PTR RR based DNS query packets traffic at the peaks (2)-(9) and the following top and/or second top query keywords are obtained, as:

Date	IPv4 address	Frequency/day
Jan. 17th	133.95.a1.173	11,263
	133.95.a2.181	2,359
Jan. 19th	133.95.a1.172	13,954
Jan. 21st	133.95.a1.172	13,158
Jan. 22nd	133.95.a1.148	8,861
Jan. 23rd	133.95.a1.145	12,047
Jan. 24th	133.95.a1.144	8,894
Jan. 25th	133.95.a3.137	7,601
	133.95.a3.144	6,405
Jan. 27th	133.95.a1.131	14,557

We performed packet capturing the outbound traffic through January 23rd, 16:34:45-48 ($\sim$ 3 sec: 25,680KB) by Ethereal-0.10.14[11] in order to confirm whether or not the PTR RR based DNS query traffic is related with spam bots activity. We can show the following SMTP TCP decoded stream (133.95.a1.145 $\rightarrow$ a victim host:25), as:

```
EHLO ******
250-mail38-***.*****.com
250-PIPELINING
250-SIZE 150000000
250-ETRN
250-STARTTLS
250 8BITMIME
MAIL FROM:<lynxes@the.*********llage.com>
RCPT TO: <francis@*******.******.com>
RCPT TO: <fady@*******.******.com>
RCPT TO: <unrzegcg@*******.******.com>
RCPT TO: <tasziqv@*******.******.com>
RCPT TO: <stevelind@*******.******.com>
RCPT TO: <sbachman@*******.******.com>
DATA
250 Ok
250 Ok
250 Ok
250 Ok
250 Ok
250 Ok
354 End data with <CR><LF>.<CR><LF>
250 Ok: queued as 7*83***D807*
```

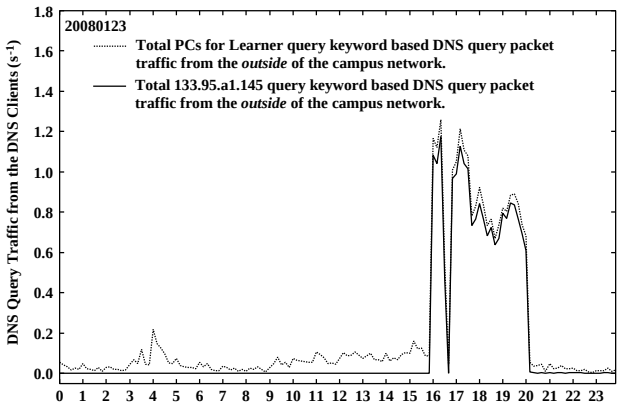

Fig. 6. The total DNS query packet traffic including the IP addresses of PCs for learners as query keywords from the outside of the campus network to the top domain name system (**tDNS**) server through January 23rd, 2007. The dotted and solid lines show the total traffic and the traffic including only 133.95.a1.145 as their query keywords, respectively (s^{-1} unit).

In this TCP stream, we can expectedly observe the spam bots activity in the PC for learners (133.95.a1.145). This is because the PC for learners is normal Windows PC and it has no function to perform E-mail delivery service.

Also, it is found that the specific account (login ID) for the PCs for learners since the account can be observed in the syslog files of the student account servers and the PTR RR based DNS query traffic can be observed through when carrying out login into the PCs for learners.

Therefore, we made contact with the account holder about the security incident and we investigated the PCs for learners. However, we cannot find any evidence and/or trace in the PCs for learners. After the interview with the account holder, it is found that the account holder always uses a USB key disk storage to save his/her document and/or spreadsheet data.

Then, we investigate the USB key disk storage with anti-virus scanners (Trendmicro Viurs Baster). Finally, we successfully detected an auto.inf file in the USB key disk storage and AV-scanners pointed out an W32/Agent.BUL Trojan horse (TH) at February 28th, 2007 in which the TH is a down loader type bot virus[12].

Therefore, it can be concluded that the bot virus infected USB key disk storage kicks auto.inf if opened by user and bot virus down loading a spam bot from the other site. And it starts spam bots activity.

4. Conclusions

We investigated statistical and entropy analyses on the total and the PTR resource record (RR) based DNS query packet traffic from the *outside* of the campus network through January 1st, 2007 to February 29th, 2008. The following interesting results are obtained, as follows: (1) We can

observe 15 incidents in the total DNS query packets traffic but 17 incidents in the entropy change of the total DNS query packets traffic. This result indicates that entropy analysis on the DNS query packets traffic can raise up a detection rate of the security incidents in the campus network. (2) We can more clearly observe 22 incidents in the entropy change in the total PTR RR based DNS query packets traffic. This means that the entropy analysis on the PTR RR based DNS query packets traffic is more superior to that on the total DNS query packets traffic. In the entropy change of the PTR RR based DNS query packets traffic, the peaks for the random spam bots (RSB) become to be very sharpened. Probably, this result is interpreted in terms of discarding the specific query keywords such as fully qualified domain names of the local E-mail servers in the total PTR RR based DNS query traffic. (3) We found the specific IP addresses of the PCs for learners in the query keywords of the PTR RR based DNS query packets traffic from the outside of the campus network through January 17th, 2008 so that we also carried out entropy analysis on the total DNS query packets from the outside of the campus network including the IP addresses of the PCs for learners as their query keywords. From the analysis, we further detected several specific IP addresses of the PCs for learners through January 17th to 27th, 2008. It is found that all the detected specific IP addresses concern only one account holder. We contacted the account holder and investigated the PCs for learners but no trace or signature of spam bots in the PCs for learners. Finally, we found that the USB key disk storage kicks to download spam bot from the internet and performs the spam bots activity through inserting the USB key disk storage into the PCs for learners. After a survey, the W32/Agent.BUL Trojan Horse was found in the USB key disk storage. From these results, we took a simple countermeasure (OP25B) to suppress the spam bots activity kicking by the Trojan Horse in the USB key disk storage from the subnetwork addresses of the PCs for learners.

We further continue to develop spam bots activity detection technology according to the results of the present paper and to raise up the detection rate.

Acknowledgement

All the studies were carried out in CMIT of Kumamoto University. We gratefully thank to all the CMIT and MQS staffs.

References

[1] P. Barford and V. Yegneswaran, An Inside Look at Botnets, Special Workshop on Malware Detection, *Advances in Information Security*, Springer Verlag, 2006.

[2] J. Nazario, Defense and Detection Strategies against Internet Worms, I Edition; *Computer Security Series*, Artech House, 2004.

[3] (a) J. Kristoff, Botnets, detection and mitigation: DNS-based techniques, *Northwestern University*, 2005, http://www.it.northwestern.edu/bin/docs/bots_kristoff_jul-05.ppt. (b) J. Kristoff, Botnets, *North American Network Operators Group (NANOG32)*, Reston, Virginia (2004), http://www.nanog.org/mtg-0410/kristoff.html

[4] D. David, C. Zou, and W. Lee, Model Botnet Propagation Using Time Zones, *Proceeding of the Network and Distributed System Security (NDSS) Symposium 2006;* http://www.isoc.org/isoc/conferences/ndss/06/proceedings/html/2006/

[5] A. Schonewille and D. -J. v. Helmond, The Domain Name Service as an IDS. How DNS can be used for detecting and monitoring badware in a network, 2006; http://staff.science.uva.nl/˜delaat/snb-2005-2006/p12/report.pdf

[6] B. McCarty: Botnets: Big and Bigger, IEEE Security and Privacy, No.1, pp.87-90 2003.

[7] A. Wagner and B. Plattner, Entropy Based Worm and Anomaly Detection in Fast IP Networks, *Proceedings of 14th IEEE Workshop on Enabling Technologies: Infrastracture for Collaborative Enterprises (WETICE 2006)*, Linköping, Sweden, 2005, pp.172-177

[8] D. A. Ludeña Romaña, H. Nagatomi, Y. Musashi, R. Matsuba, and K. Sugitani: A DNS-based Countermeasure Technology for Bot Worm-infected PC terminals in the Campus Network, Journal for Academic Computing and Networking, Vol. 10, No.1, pp.39-46 2006.

[9] D. A. Ludeña Romaña, Y. Musashi, and K. Sugitani: Entropy Study on MX Resource Record-Based DNS Query Packet Traffic, IPSJ Symposium Series, Vol. 2004, No.13, pp.21-26 2007.

[10] BIND-9.2.6: http://www.isc.org/products/BIND/

[11] Ethereal-Network Protocol Analyzer: http://http://www.ethereal.com/

[12] W32/Agent.BUL Trojan Horse (TH): http://www.trendmicro.com/vinfo/virusencyclo/default5.asp?VName=TROJ_AGENT.BUL