International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

DNS Based Spam Bots Detection in a University

Dennis Arturo Ludeña Romaña[1] **Shinichiro Kubota**[2] **Kenichi Sugitani**[2] **Yasuo Musashi**[2]

[1]Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan
dennis@st.cs.kumamoto-u.ac.jp

[2]Center for Multimedia and Information Technologies, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan
{s-kubota,sugitani,musashi}@cc.kumamoto-u.ac.jp

Abstract: We carried out 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 PC room terminals through April 1st, 2007 to April 30th, 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 PC room terminals 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.

Keywords: Computer Network Security, Computer Network Management, Computer Viruses and Threats, Log Analysis.

1. Introduction

It is of considerable importance to boost up a detection rate of spam bots (SBs), since they become components of the bot networks that 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 unique 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 packets 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 unique DNS query keywords and the unique source IP addresses based entropies [9]. In the former one, the unique source IP addresses- and the unique DNS query keywords based-entropies decreases when the targeted spam bots activity is high, while in the latter one, the unique 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 packets 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

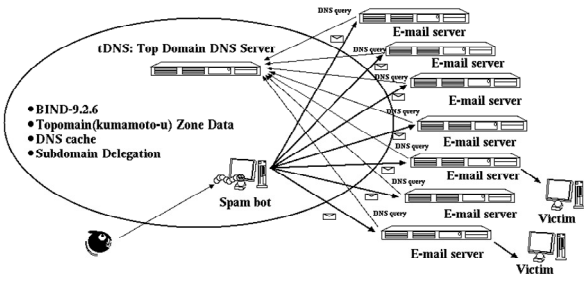

Figure.1 A schematic diagram of a network observed in the present study

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 room terminals.

2. Observations

2.1. Network System and DNS Query Packets Capturing

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.

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.

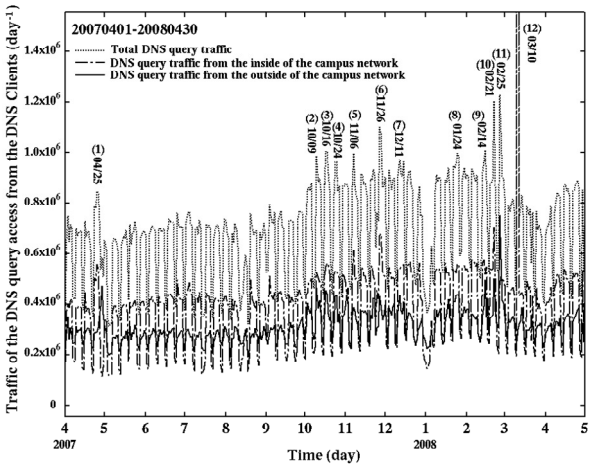

Figure.2 Traffic of the DNS query packets to the top domain DNS server (tDNS) and the traffic from the inside- and the outside-DNS clients in a university through April 1st, 2007 to April 30th, 2008 (day^{-1} unit).

2.2. Observed DNS Query Traffic

Firstly, we can demonstrate the total DNS query traffic from the inside and outside of the campus network through April 1st, 2007 to April 30th, 2008, as shown in Figure 2.

In Figure 2, the twelve large peaks are observed. The large peaks have been grouped, as follows; the first group consists of (1) April 25th, 2007 and (12) March 10th, 2008, the second one consists of (2) October 9th, (3) 16th, and (4) 24th, (5) November 6th and (6) 26th, (7) December 11th, 2007, (9) February 14th, (10) 21st, and (11) 25th, 2008, and the last one consists of (8) January 24th, 2008.

The first group {(1), (12)} is taken place by the DNS misconfiguration. In the second group {(2), (3), (4), (5), (6), (7), (9), (10), (11)}, 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 {(8)}, we observed the DNS query traffic from the outside of the campus network including several IP addresses as their query keywords belonging to the PC room terminals under our administration.

Therefore, we further carried out entropy analysis on the total DNS query traffic from the outside of the campus network, the PTR resource record (RR) based DNS query traffic from the outside of the campus network, and the PTR RR based DNS query traffic including only the subnetwork addresses of PC room terminals from the outside of the campus network.

2.3. Estimation of Entropy

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

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

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

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

where *i* and j ($i, j \in X$) represent the unique source IP address or the unique DNS query keywords in the DNS query packets, and the frequency *freq(i)* is estimated with the following script program:

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

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 "DNS query keywords". 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 backslash "\" 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" (or "grep") command extracts only the message-lines excluding (or including) 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 into the file "freq-querykeywords". 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 April 1st, 2007 to April 30th, 2008, as shown in Figure 3.
In Figure 3, we can observe significant peaks of (1) April 3rd and (2) 29th, (3) May 20th, (4) October 9th, (5) 16th, and (6) 24th, (7) November 1st, (8) 6th, and (9) 26th, (10) December 11th, 2007, (11) January 17th, (12) 19th, (13) 20th, and (14) 25th, (15) February 14th, (16) 21st, and (17) 25th, 2008. Expectedly, we have already observed the same peaks (4), (5), (6), (8), (9), (10), (15), (16), and (17) in Figure 3 corresponding to (2), (3), (4), (5), (6), (9), (10), (11), and (12) 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 (1), (2), (3), (7), (11), (12) and (13) 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 the DNS based detection rate of security incidents.

Also, in Figure 3, almost all the peaks are simply assigned to usual spam bots activity because

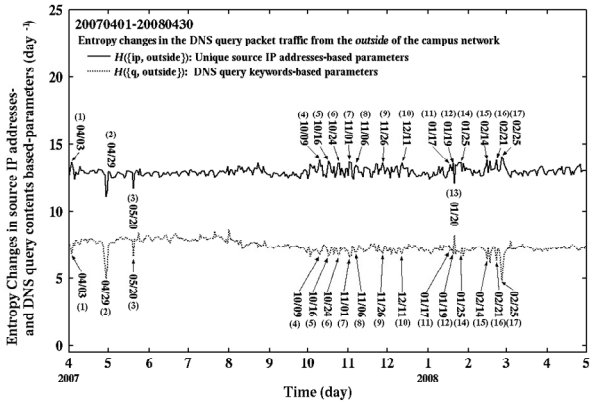

Figure.3 Entropy changes in the total DNS query packets traffic from the outside of the campus network to the top domain name system (tDNS) server through April 1st, 2007 to April 30th, 2008. The solid and dotted lines show the unique source IP addresses- and the unique DNS query keywords based-entropies, respectively (day^{-1} unit).

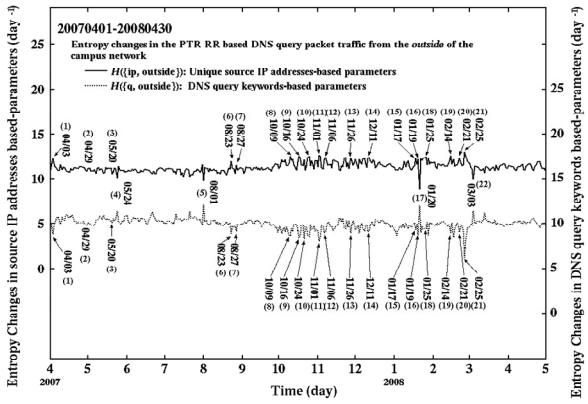

Figure.4 Entropy changes in the total PTR resource record (RR) based DNS query packets traffic from the outside of the campus network to the top domain name system (tDNS) server through April 1st, 2007 to April 30th, 2008. The solid and dotted lines show the unique source IP addresses- and the unique DNS query keywords-based entropies, respectively (day^{-1} unit).

we detected the same or similar IP addresses and FQDNs of the local vulnerable E-mail servers. However, the several peaks (11), (12), 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 authors administrated specific subnet addresses.

Interestingly, in the peak (13), we detected large PTR resource record (RR) based DNS query packets traffic from a specific site including the IP addresses of the university campus network as their query keywords. Probably, this site tried to collect the live IP addresses of the campus network and in order to carry out port-scan for the live hosts in the campus network. This result indicates that we can detect the live hosts harvesting activity with entropy analysis on the DNS query traffic from the outside of the campus network.

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 April 1st, 2007 to April 30th, 2008 (Figure 4).

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

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

Previously, we reported the similar insights for entropy analysis on the MX RR based DNS query packet traffic from the campus network [9] in which we described two types of spam bots; *random spam bots* (RSB) and *targeted spam bots* (TSB) in Figure 6. In the last group, we can observe that the unique source IP addresses based entropy decreases but the unique DNS query keywords based one increases.

Furthermore, we also detected several specific sites trying to search the live IP addresses of the campus network. This shows that the unique source IP addresses based entropy decreases when increasing the frequency of the unique source IP addresses. And the unique DNS query keywords based entropy increases when increasing the number of unique DNS query keywords (Figure 6). Therefore, it can be concluded that we can easily

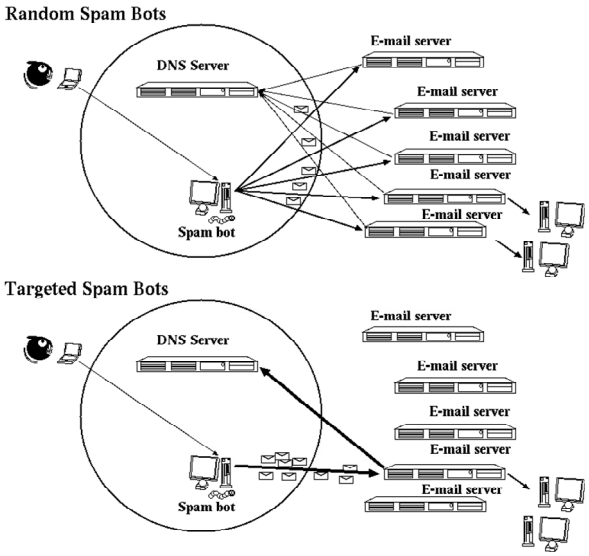

Figure.5 Random spam bots (RSB) and targeted spam bots (TSB).

detect the live hosts harvesting activity with entropy analysis on the DNS query traffic from the outside of the campus network *i.e.* only observing the DNS query packets traffic from the outside of the campus network.

3.3. Entropy Analysis on DNS Query Traffic including IP addresses of PC room terminals

We demonstrate the calculated the unique source IP addresses- and the unique query keywords-based entropies in the PTR resource record (RR) based DNS query packet traffic including only the IP addresses of PCs in the PC rooms as their query keywords from the outside of the campus network to the top domain name system (**tDNS**) server through April 1st, 2007 to April 30th, 2008, as shown in Figure 7.

In Figure 7, we can observe several interesting peaks of (1) August 9th, 2007, (2) January 17th, (3) 19th, (4) 21st, (5) 22nd, (6) 23rd, (7) 24th, (8) 25th, (9) 27th, and (10) March 3rd, 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 in Table 1: where the above top IP addresses are obtained when the frequency takes more than 1,000/day..

Surely, we obtained a couple of the top IP addresses of 133.95.a1.173 and 133.95.a2.181, in

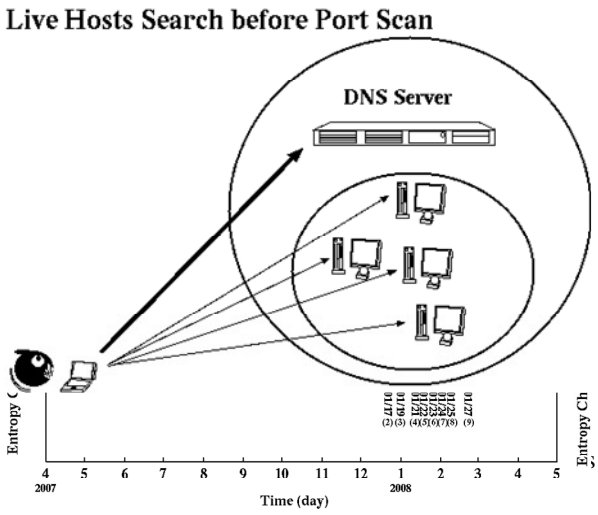

Figure.7 Entropy changes in the DNS query packets traffic including the IP addresses of PC room terminals as query keywords from the outside of the campus network to the top domain name system (**tDNS**) server through April 1st, 2007 to April 30th, 2008. The solid and dotted lines show the source IP addresses- and DNS query keywords-based entropies, respectively (day^{-1} unit).

which the both IP addresses are assigned to the PC room terminals-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 shown in Table 2.

Table 1. Detected unique IP addresses and their Frequency at January 17th, 2008.

IPv4 Address	Frequency
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

We performed packet capturing the outbound traffic through January 23rd, 16:34:45-48 (~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 → a victim host:25), as:

Table 2. Detected top/2nd-top unique IP addresses and their Frequency through January 17th to 27th, 2008.

Date	IPv4 Address	Frequency (day^{-1})
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

```
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*
```

Chart 2. Packet capturing

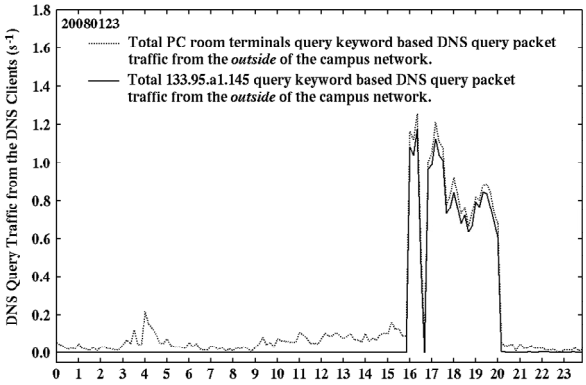

Figure.8 The total DNS query packets traffic including the IP addresses of PC room terminals as query keywords from the outside of the campus network to the top domain name system (**tDNS**) server through January 23rd, 2007. The dotted and solid lines show the total traffic and the traffic including 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 room terminals (133.95.a1.145). This is because the PC room terminal is normal Windows PC and it has no function to perform E-mail delivery services.

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

Therefore, we made contact with the account holder about the security incident and we investigated the PC room terminals. However, we cannot find any evidence and/or trace in the PC room terminals. 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.

Note that at the peak (10) in Figure 7, we detected the live hosts harvesting activity in order to carry out the next port-scan on the PC room terminals.

4. Conclusions

We investigated statistical and entropy analyses on the total and the PTR resource record (RR) based DNS query packets traffic from the outside of the campus network through April 1st, 2007 to April 30th, 2008. The following interesting results are obtained, as follows: (1) We can observe 12 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 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 PC room terminals in the query keywords of the PTR RR based DNS query packets traffic from the outside of the campus network at 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 PC room terminals as their query keywords. From the analysis, we further detected several specific IP addresses of the PC room terminals 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 PC room terminals but no trace or signature of spam bots in the PC room terminals. Finally, we found that the USB key disk storage of the account holder kicks to download spam bots components from the internet and performs the spam bots activity through inserting the USB key disk storage into the PC room terminals. As a result, the W32/Agent.BUL Trojan Horse was detected in the USB key disk storage. From these results, we took a simple countermeasure (OP25B: Outbound Port 25 Blocking) in order to suppress the spam bots activity triggered by the Trojan Horse in the USB key disk storage from the subnetwork addresses of the PC room terminals.

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

Acknowledgments

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 InterInternet 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_jul05.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,” The Proceedings 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, 2003, pp.87-90.D

[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, 2006, pp.39-46

[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, 2007, pp.21-26.
[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