Entropy Study on MX Resource Record-Based DNS Query Packet Traffic

Dennis Arturo Ludeña Romaña, † Yasuo Musashi,‡ and Kenichi Sugitani ‡

†Graduate School of Science and Technology and ‡Center for Multimedia and Information Technologies, Kumamoto University,
2-39-1 Kurokami, Kumamoto-City, 860-8555, JAPAN

Abstract: We carried out entropy study on the source IP addresses- and query keywords in the MX resource record (RR) type DNS query packet traffic between the top domain DNS server and the DNS clients in a university through January 1st, 2004 to July 31st, 2007. The interesting results are summarized, as follows: (1) The source IP addresses- and query keywords-based entropies change symmetrically when detecting random spam bots activity. On the other hand (2), the source IP addresses- and query keywords-based entropies changes similarly each other when detecting targeted spam bots activity. Therefore, it can be concluded that we can distinguish two types of spam bots activity in the campus network by only observing the DNS query packet traffic.

1. Introduction

It is of considerable importance to raise up a detection rate of bot worms (BWs), since they compromise not only the PC clients but also hijack the compromised PC clients. After the hijacking, the BW-compromised PC clients become almost components of the bot networks (bots) that are used to send a lot of unsolicited mails like spam, phishing, and mass mailing (spam bots activity) and to execute distributed denial of service attacks (DDoS bots activity).[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[8] when the BW activity is high. This is probably because the BW activity like a spam bot one is very easily to be sensed by the spam filter and/or the IDS on the internet. Therefore, we can detect bot worm (BW) activity, especially as spam bots on the campus network, by only watching the DNS query packet traffic from the other sites on the internet.

Also, we very recently reported that in the DNS query packet traffic from the inside of the campus network, the DNS query keywords based entropy considerably decreases while the source IP addresses based one decreases when the BW activity is high.[9]

However, it is likely that we can find no entropy study on the MX record resource (RR) based DNS query packet traffic. In this paper, (1) we carried out the entropy study on the MX RR based DNS query packet traffic from the campus network, and (2) we discuss on the difference between the MX RR based DNS query packet traffic from the random spam bots and the targeted spam ones.

2. Observations

2.1 Network systems

We investigated traffic of the DNS query packet access between the top domain DNS (**tDNS**) server and the DNS clients. Figure 1 shows an

†Graduate School of Science and Technology, Kumamoto University.
‡Center for Multimedia and Information Technologies, Kumamoto University.

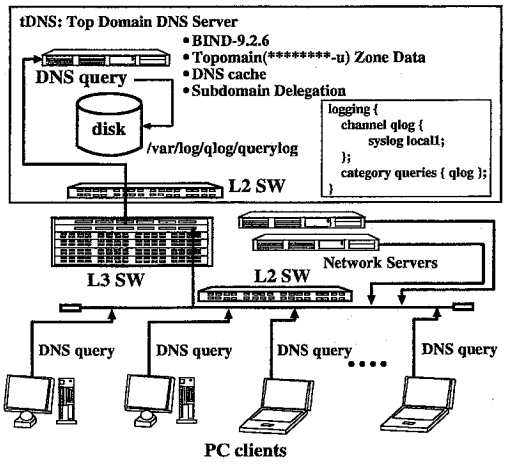

Figure 1. A schematic diagram of a network observed in the present study.

observed network system in the present study, an optional configuration of the BIND-9.2.6 server program daemon in **tDNS**. The DNS server, **tDNS**, is one of the top level DNS (kumamoto-u) servers and plays an important role of domain name resolution and subdomain delegation services for many PC clients and the subdomain network servers in the university, respectively, and the operating system is CentOS 4.3 Final and is currently employed kernel-2.6.9 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 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 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 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

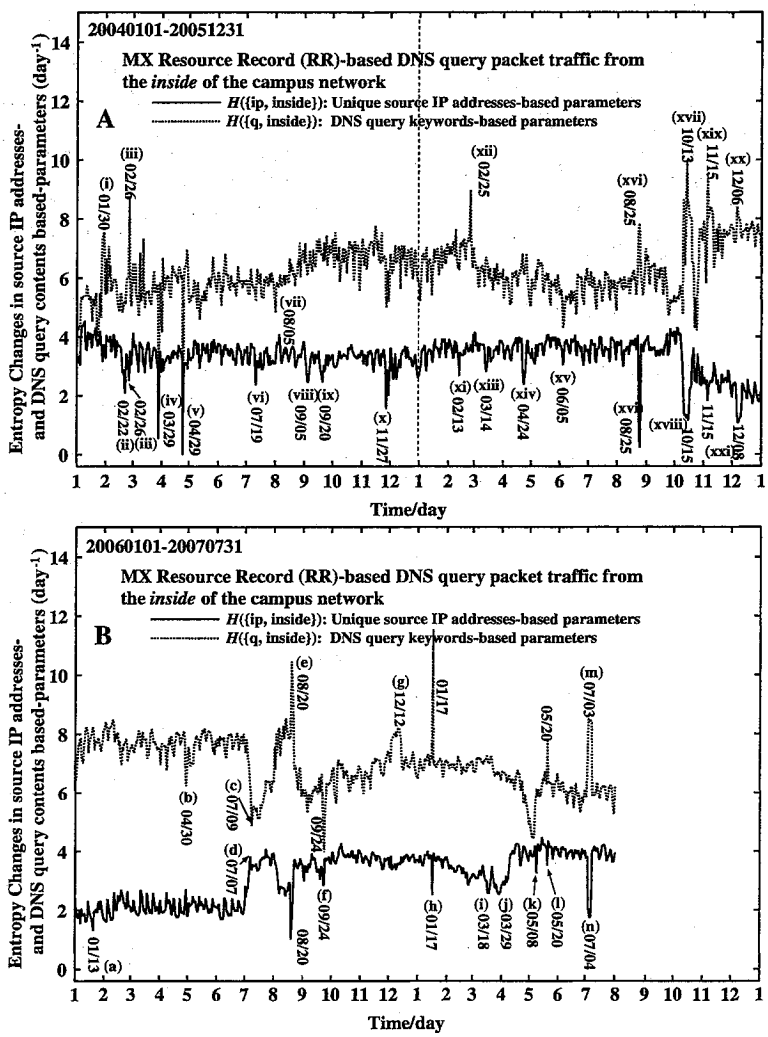

Figure 2. Entropy changes in the MX resource record based DNS query packet traffic from the campus network to the top domain name system (**tDNS**) server through January 1st, 2004 to December 31st, 2005 (A), and January 1st, 2006 to July 31st, 2007. The solid and dotted lines show the source IP addresses and DNS query keywords based entropies, respectively (day^{-1} unit).

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

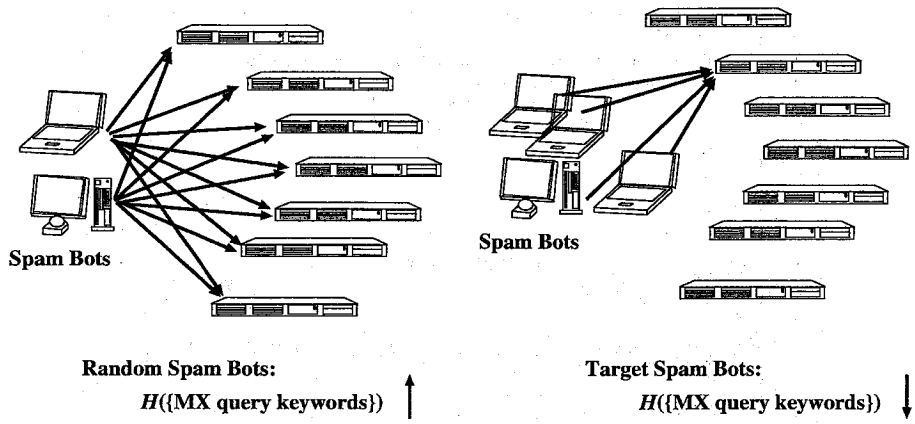

Figure 3. Random spam bots (RSB) and target spam bots (TSB). RSB raise up the entropy value of the MX query keywords while TSB take down the entropy one of the MX query ones.

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

We illustrate the calculated the source IP addresses and the query keywords based entropies in the MX resource record (RR) based DNS query packet traffic from the campus network to the top domain name system **tDNS** server through January 1st, 2004 to December 31s, 2005 and through January 1st, 2006 July 31st, 2007, as shown in Figure 2 (2A and 2B, respectively)

In Figure 2, we can observe the following significant peaks of (i) January 30th, (ii) February 22nd and (iii) 26th, (iv) March 29th, (v) April 29th, (vi) July 19th, (vii) August 5th, (viii) September 5th and (ix) 20th, (x) November 27th, 2004, (xi) February 13th and (xii) 25th, (xiii) March 14th, (xiv) April 24th, (xv) June 5th, (xvi) August 25th, (xvii) October 13th and (xviii) 15th, (xix) November 15th, (xx) December 6th, and (xxi) 8th, 2005, (a) January 13th, (b) April 30th, (c) July 7th and (d) 9th, (e) August 20th, (f) September 24th, (g) December 12th, 2006, (h) Janury 17th, (i) March 18th and (j) 29th, (k) May 8th and (l) 20th, (m) July 3rd, and (n) 4th, 2007.

Fortunately, the following three peaks (i), (iii), and (iv) can be assigned for W32/Mydoom.A mass mailing worm (MMW) activity,[11] spam bots activity, and W32/Netsky.Q MMW activity,[12] in which we have previously reported.[13]

Interestingly, the source IP addresses based entropy decreases and the query keywords based one increases simultaneously in the peaks (i) and (iii). In other words, the both entropies changes symmetrically. The increase in the query keywords based entropy clearly indicates *random spam bots* (RSB) activity like a mass mailing worm (MMW) attack. On the other hand, the source IP addresses and the query keywords based entropies decrease similarly in the peak (iv). This result clearly shows *targeted spam bots* (TSB) activity like a targeted SMTP-DoS attack.

Therefore, we can categorize the peaks (i)-(xxi),(a)-(n) into two groups: RSB group {(i), (iii), (vi), (vii), (viii), (ix), (xii), (xiii), (xiv), (xvii), (xviii), (xix), (xx), (xxi), (a), (d), (e), (g), (h), (i), (j), (k), (l), (m), (n)} and TSB group {(ii), (iv), (v), (x), (xi), (xv), (b), (c)}.

As a result, it can be clearly concluded that en-

tropy analysis on the MX resource record (RR) based DNS query packet traffic provides us very important information on the spam bots activity in the campus network.

4. Conclusions

We performed entropy based analysis on the MX resource record (RR) type DNS query packet traffic from the campus network toward the top domain name system (**tDNS**) server through January 1st, 2004 to July 31st, 2007. The following interesting results are obtained, as: (1) The source IP addresses and the DNS query keywords based entropies considerably decrease and increase, respectively, when the random spam bots activity takes place, while (2) the source IP addresses and the DNS query keywords based entropies decrease similarly when the targeted spam bots activity occurs.

From these results, it can be concluded that we can detect two kinds of spam bot activity by observing the source IP addresses and query keywords based entropy changes of the MX RR based DNS query packet traffic.

We further continue to develop detection technology according to the results of the present paper and to evaluate of the detection rate.

Acknowledgement. All the calculations and investigations were carried out in Center for Multimedia and Information Technologies (CMIT), Kumamoto University. We gratefully thank to all the CMIT staffs and system engineers of MQS (Kumamoto) for daily supports and constructive cooperations.

References and Notes

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

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

3) (a) Kristoff, J., Botnets, detection and mitigation: DNS-based techniques, *Northwestern University*, 2005, http://www.it.northwestern.edu/-bin/docs/bots_kristoff_jul05.ppt. (b) Kristoff, J., Botnets, *North American Network Operators Group (NANOG32)*, Reston, Virginia (2004), http://www.nanog.org/mtg-0410/kristoff.html

4) David, D., Zou, C., and Lee, W., 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) Schonewille, A. and v. Helmond, D. - J., 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) McCarty, B.: Botnets: Big and Bigger, *IEEE Security and Privacy*, No.1, pp.87-90 (2003).

7) Wagner, A. and Plattner, B., 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) A. Ludeña Romaña, D., Nagatomi, H., Musashi, Y., Matsuba, R., and Sugitani, K.: 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) A. Ludeña Romaña, D. and Musashi, Y.: Entropy Based Analysis of DNS Query Traffic in the Campus Network, *Proceeding for the 4th*

International Conference on Cybernetics and Information Technologies, System and Applications (CITSA2007), Orlando, FL USA 2007, pp.162-164.

10) http://www.isc.org/products/BIND/

11) http://www.trendmicro.com/vinfo/virus-encyclo/default5.asp?VName=WORM_MYDOOM.A

12) http://www.trendmicro.com/vinfo/virus-encyclo/default5.asp?VName=WORM_NETSKY.Q

13) (a) Matsuba, R., Musashi, Y., and Sugitani, K.: Detection of Mass Mailing Worm-infected IP address by Analysis of Syslog for DNS server, *IPSJ SIG Technical Reports, Distributed System and Management 32nd (DSM32)*, Vol. 2004, No.37, pp.67-72 (2004). (b) Musashi, Y. and Rannenberg, K.: Detection of Mass Mailing Worm-infected PC terminals by Observing DNS Query Access, *IPSJ SIG Technical Reports, Computer Security 27th, (CSEC27)*, Vol. 2004, No.129, pp.39-44 (2004).