Entropy Based Analysis of DNS Query Traffic in the Campus Network

Dennis Arturo Ludeña Romaña
Graduate School of Science and Technology, Kumamoto University
Kumamoto 860-8555 JAPAN

and

Yasuo Musashi
Center for Multimedia and Information Technologies, Kumamoto University
Kumamoto 860-8555 JAPAN

ABSTRACT

We carried out the entropy based study on the DNS query traffic from the campus network in a university through January 1st, 2006 to March 31st, 2007. The results are summarized, as follows: (1) The source IP addresses- and query keyword-based entropies change symmetrically in the DNS query traffic from the outside of the campus network when detecting the spam bot activity on the campus network. On the other hand (2), the source IP addresses- and query keyword-based entropies change similarly each other when detecting big DNS query traffic caused by prescanning or distributed denial of service (DDoS) attack from the campus network. Therefore, we can detect the spam bot and/or DDoS attack bot by only watching DNS query access traffic.

Keywords: Bot, Bot Worm, Detection, DNS, Entropy

1. INTRODUCTION

It is of considerable importance to raise up a detection rate of bot worms (BWs), because they compromise not only the PC clients but also hijack the compromised PC clients. After the hijacking, the BW-infected PC clients become almost components of the bot networks (bots) that are used to send a log of unsolicited mails like spam, phising, and mass mailing (SMTP proxy) and to execute distributed denial of service attacks [1-4].

Previously, we reported that the entropy based on the frequency of the DNS query keywords in the DNS query traffic from the outside campus decreases considerably when the entropy based on the frequency of the source IP addresses increases [5] *i.e.* we can detect bot worm (BW) activity, especially as spam bots on the campus network by only watching the DNS query traffic from the other sites on the internet. However, it is likely that we can find no investigation on the comparison between the entropy based on the frequency from the *inside* of the campus network.

In this paper, (1) we carried out the entropy based study on the entropy based analysis of the DNS query traffic from the campus network, and (2) we discuss on the difference between the entropy based on the frequency from the *inside* of the campus network.

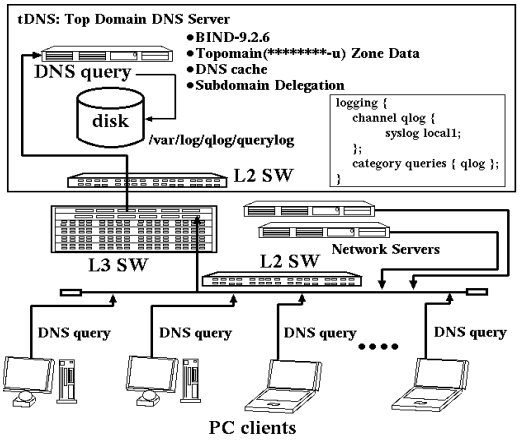

Figure 1. A schematic diagram of a network observed in the present study.

2. OBSERVATIONS

Network Systems

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

Capture of DNS Query Packets

In **tDNS**, BIND-9.2.6 program package has been employed as a DNS server daemon [6]. 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.

Estimation of Entropy

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

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

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 outside of 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 X) represent the source IP address or the DNS query keywords in the DNS query packet, and the frequency $freq(i)$ are 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[6]. 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 and 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}' "

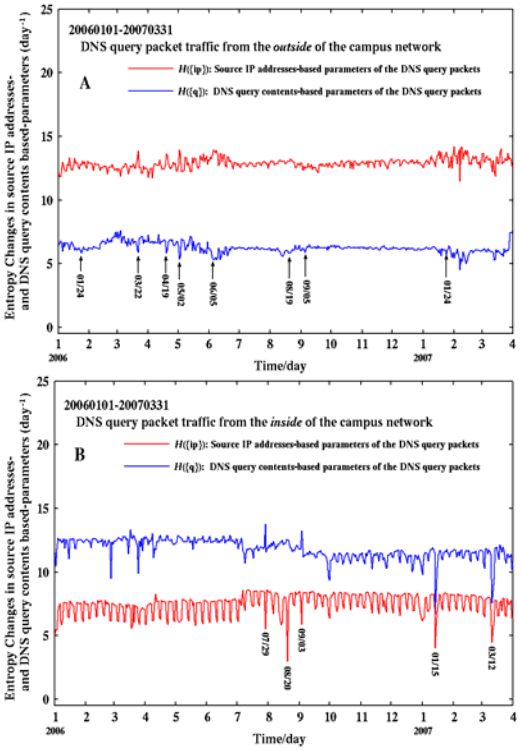

Figure 2. Entropy changes in the DNS query traffic from the outside (A) and the inside (B) of the campus network to the top domain name system (tDNS) server through January 1st, 2006 to March 31st, 2007 (day^{-1} unit). The both solid and dotted lines show entropies based on the data set of the number of the unique source IP addresses and on the frequency of the unique DNS query keywords, respectively.

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 "freqquerykeywords". 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. Entropy based packet traffic analysis was suggested by Wagner and Plattner , recently [7].

3. RESULTS AND DISCUSSION

Entropy Analysis on DNS Query Traffic

We illustrate the calculated entropy for the frequencies of the unique source IP addresses and the DNS query

keywords in the DNS traffic from the inside and the outside of the campus network to the top domain DNS (**tDNS**) server through January 1st, 2006 to March 31st, 2007, as shown in Figure 2.

In Figure 2A, we can observe several significant peaks of (i) January 24th, (ii) March 22nd, (iii) April 29th, (iv) May 2nd, (v) June 5th, (vi) August 19th, (vii) September 5th, 2006, and (viii) January 24th, 2007. Fortunately, since we have received a lot of complaining E-mail against the described peaks from the other sites, the peaks have been assigned to the security incidents, as follows: the spam bot activities for (i)-(vi), the misconfiguration in the campus subdomain DNS server for (vii), and the unknown for (viii), respectively. Interestingly, we can also notice that the both entropy curves change symmetrically at each peaks. These results show that security incidents in the university campus network can be detectable when only observing the frequency of the domain name resolution access from the *outside* of the campus network.

In Figure 2B, on the other hand, we can find several peaks of (a) July 29th, (b) August 20th, (c) September 9th, 2006, (d) January 15th, and (e) March 15th, 2007. Also, these peaks have already fixed, as: E-mail spamming activity for (a), a crash of the local E-mail server by the big SMTP traffic for (b), the DNS misconfiguration for (c) in the local subdomain DNS servers, the big historical domain name resolution traffic for (d) and (e) in which the DNS query traffic were generated by crashes of the NIS-based authentication systems.

Furthermore, we can obtain new findings when comparing the unique source IP addresses- and DNS query keywords based entropy curves each other in Figures 2A and 2B, respectively. This is because the symmetrical changes emerges when detecting the spam bot activity, however, simultaneous changes take place when receiving the unusual big DNS query traffic from the campus network because of DNS related misconfiguration at the local subdomain and/or overloaded crash at the local E-mail servers.

As a result, it can be clearly concluded that entropy based analysis on the DNS query traffic provides an important information on the security incidents in the campus network.

4. CONCLUSIONS

We investigated on the DNS query traffic from the campus network in a university through January 1st, 2006 to March 31st, 2007 employing entropy based statistical analysis method. The following interesting results are obtained, as: (1) The source IP addresses- and query keyword-based entropies change symmetrically in the DNS query traffic from the outside of the campus network when detecting the spam bot activity on the campus network. (2), the source IP addresses- and query keyword-based entropies change similarly each other when detecting big DNS query traffic caused by prescanning or distributed denial of service (DDoS) attack from the campus network.

From these results, it can be concluded that we can detect the spam bot and/or DDoS attack bot in the campus network by only watching DNS query access traffic.

We continue to develop detection technology based on the results of the present paper and to evaluate of the detection rate.

5. 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/botskristoff_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”, Proceeding of the Network and Distributed System Security (NDSS) Symposium 2006; http://www.isoc.org/isoc/conferences/ndss/06/proceedings/html/2006/

[5] D. A. Ludeña R., 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)

[6] BIND-9.2.6: Internet Systems Consortium http://www.isc.org/products/BIND/

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