DNS Based Spam Bots Detection in a University

Dennis Arturo Ludeña Romaña,[1] Shinichiro Kubota,[2] Kenichi Sugitani,[2] and Yasuo Musashi[2]
[1]*Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan*
dennis@st.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) In January 17th, 2008, the DNS query traffic is mainly dominated by several specific IP addresses as their query keywords. (2) 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.

1. Introduction

Recently, 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 [1].

In this paper, we carried out entropy analysis on the PTR resource record (RR)-based DNS query packets traffic from the outside of the campus network.

2. Observations

2.1. Network System, DNS Query Packets Capturing, and Estimation of Entropy

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

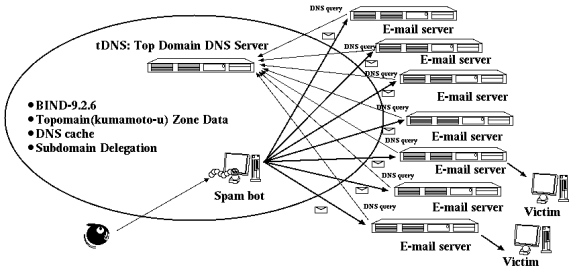

Figure 1. A schematic diagram of a network observed in the present study.

of the BIND-9.2.6 DNS server program daemon [2] of the top domain name system (**tDNS**) server. 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. The line of syslog message 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 (FQDN)(A or AAAA RR) type, an IP address (PTR RR) type, or a mail exchange (MX RR) type.

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 frequencies $\{freq(j)\}$ of IP addresses or that of the DNS query keywords in the DNS query packets traffic from the outside of the campus network, and the probability $P(i)$ is defined, as

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

978-0-7695-3391-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICINIS.2008.54

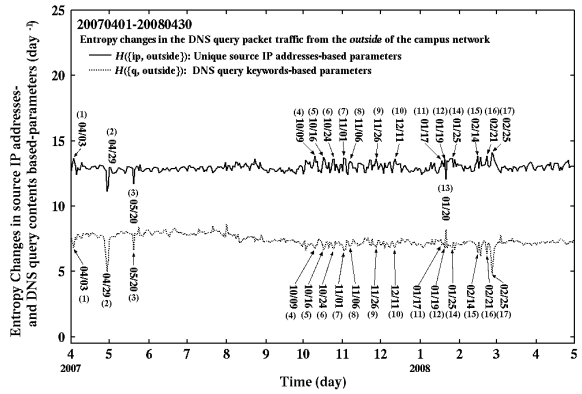

Figure 2. Entropy changes in the total DNS query packets traffic from the outside of the campus network to tDNS server through April 1st, 2007 to April 30th, 2008 (day^{-1} unit).

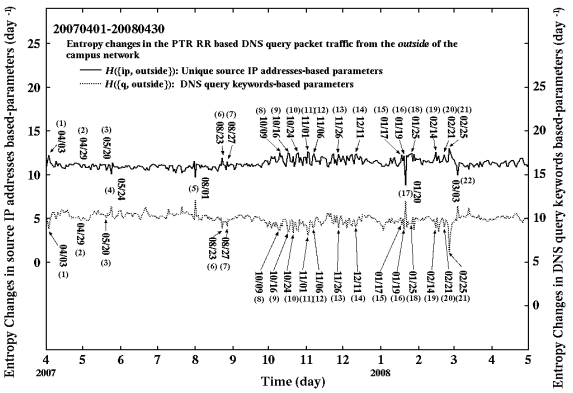

Figure 3. Entropy changes in the total PTR RR based DNS query packets traffic from the outside of the campus network to tDNS server through April 1st, 2007 to April 30th, 2008 (day^{-1} unit).

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 script program shown in [1].

3. Results and Discussion

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

We illustrate the calculated 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 **tDNS** server through April 1st, 2007 to April 30th, 2008, as shown in Figure 2.

In Figure 2, 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. In Figure 2, almost all the peaks are simply assigned to usual spam bots activity because 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 kind of spam bots are attacking, since the detected IP addresses are variable daily or hourly. Fortunately, the detected IP addresses in the peaks are easily identified because they belong to administrated specific subnet addresses.

Interestingly, in the peak (13), we detected large PTR 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 carry out active host search in the campus network.

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

We performed entropy analysis on the total PTR RR based DNS query packet traffic from the outside of the campus network through April 1st, 2007 to April 30th, 2008 (Figure 3).

In Figure 3, 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, showing that the spam bots attack randomly targeted E-mail servers on the internet (*random spam bots*: **RSB**). In the second group, on the other hand, the unique source IP addresses- and the unique DNS query keywords-based entropies decrease simultaneously, indicating that the spam bots attacks only to the specific E-mail serves on the internet (*targeted spam bots*: **TSB**).

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,

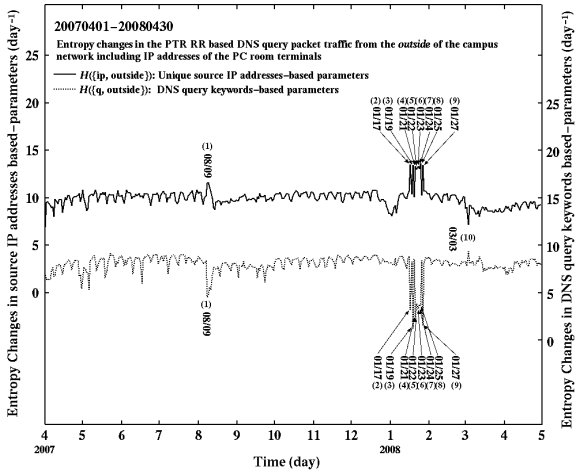

Figure 4. Entropy changes in the DNS query packets traffic including the IP addresses of PC room terminals as their query keywords from the outside of the campus network to tDNS server through April 1st, 2007 to April 30th, 2008 (day^{-1} unit).

probably indicating the same host search at the peak (13) in Figure 2.

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-RR based DNS query packets traffic including only the IP addresses of the PC room terminals as their query keywords from the outside of the campus network to the **tDNS** server through April 1st, 2007 to April 30th, 2008, as shown in Figure 4.

In Figure 4, 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, in which the above top IP addresses are obtained when the frequency takes more than 1,000/day. Surely, we 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 PC room terminals-subnet addresses: 133.95.a1.0/24 and 133.95.a2.0/24, respectively.

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

IPv4 address	Frequency (day^{-1})
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

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

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

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 showing in Table 2.

We performed packets capturing the outbound traffic through January 23rd, 16:34:45-48 (~3 sec: 25,680KB) by Ethereal-0.10.14 [3] 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:

```
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*

In this SMTP 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 a standard Windows PC and it has no function to perform E-mail delivery services.

Also, it is found that there is a specific account (login ID) used in 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 packets traffic can be observed through when carrying out login into the PC room terminals.

Therefore, we made contact with the account holder about above 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 [4].

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 bots components from the other site. And it starts spam bots activity.

Note that at the peak (10) in Figure 4, we detected the hosts search activity.

4. Conclusions

We investigated entropy analyses on the total and the PTR 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 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. (2) 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. Then, we further detected several specific IP addresses of the PC room terminals through January 17th to 27th, 2008. It is also found that all the detected specific IP addresses concern only one account holder. Therefore, 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 when 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 again triggered by the Trojan Horse in the USB key disk storage from the subnet 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.

References

[1] D. A. Ludeña Romaña, Y. Musashi, R. Matsuba, and K. Sugitani, "Detection of Bot Worm-Infected PC Terminals," *Information*, Vol. 10, No.5, 2006, pp.673-686.

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

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

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