DNS によるスパムボットとホスト探索活動の検知

デニス・アルトゥロ・ルデニャ・ロマニャ[†]
杉谷 賢一[‡]　久保田 真一郎[‡]　武蔵 泰雄[‡]

概要: 2007 年 4 月 1 日〜2008 年 7 月 31 日の期間について学外からのとある大学のトップドメイン DNS サーバへの DNS クエリパケットの流量のエントロピー研究を行ったところ、次の様な結果を得た。(1) 期間中送信元 IP アドレスに関する流量エントロピーの増加と同時に DNS クエリキーワードに関する流量エントロピーの減少が度々観察されることから、そのキャンパスネットワーク上では、ランダムスパムボットが現在も尚存在し、活動していることを示しており、また (2) 学外の特定の IP アドレスやネットワークアドレスからキャンパス内部のネットワークアドレス範囲について逆引名前解決アクセスが多数行われているのが観察されている。これらの結果は、その大学のキャンパスネットワーク上に多数のランダムスパムボットが活動し、また学外から攻撃の対象となっていることを示している。

DNS Based Detection of Spam Bots and Host Search Activity

Dennis Arturo Ludeña Romaña[†]
Shinichiro Kubota,[‡] Kenichi Sugitani,[‡] and Yasuo Musashi[‡]

Abstract: We carried out an entropy study on the DNS query traffic from the outside for a university campus network to the top domain DNS server in a university through April 1st, 2007 to July 31st, 2008. The following interesting results are given: (1) The random spam bots have been still alive and/or active in the campus network because we can observe that the unique source IP addresses-based DNS traffic entropy increases as well as the unique DNS query keywords-based one decreases frequently. (2) We have also observed a lot of the reverse name resolution access from the specific site on the campus IP address range. Therefore, it can be concluded that in the campus network, the random spam bots are still active and the campus network is also targeted by the attackers.

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.[1−4]

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.[5]

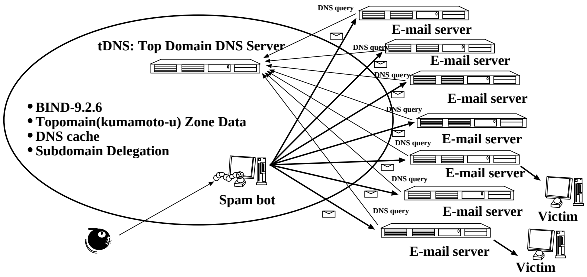

Figure 1. A schematic diagram of a network observed in the present study.

Previously, we reported that the unique DNS query keywords based entropy in the DNS query packet traffic from the outside for the campus network decreases considerably while the unique source IP addresses based entropy increases when the spam

[†] 熊本大大学院自然研究科・Graduate School of Science and Technology, Kumamoto University.
[‡] 熊本大学総合情報基盤センター・Center for Multimedia and Information Technologies, Kumamoto University.

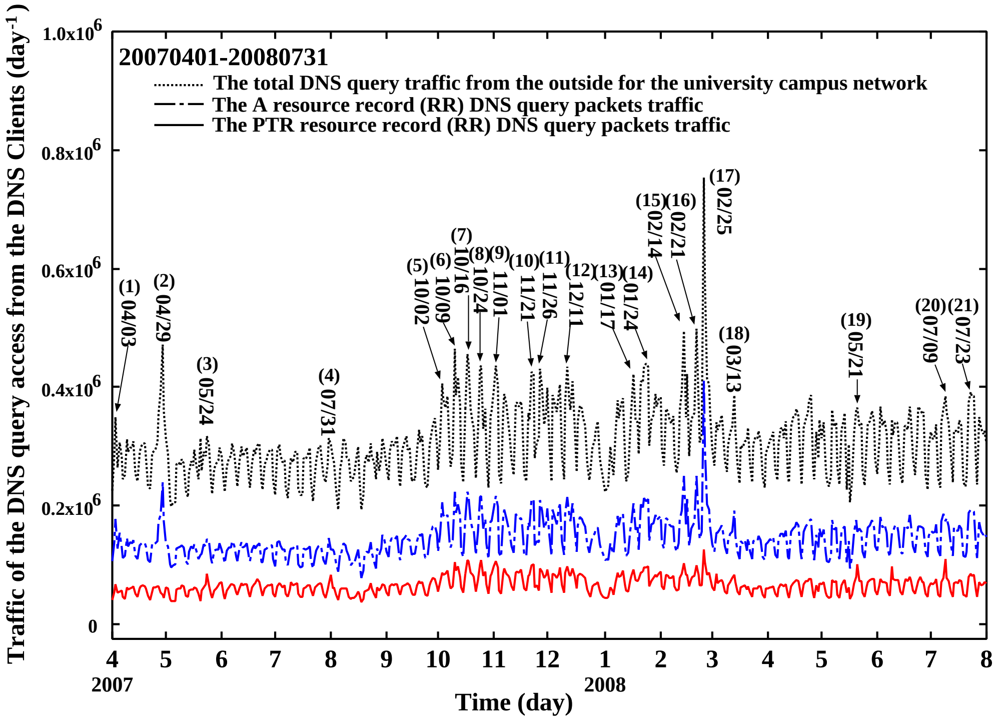

Figure 2. The total DNS query packets traffic from the DNS clients to the top domain DNS (**tDNS**) server, the A resource record (RR) based DNS query packets traffic, and the PTR RR based DNS query packets traffic from the outside for the campus network in a university through April 1st, 2007 to July 31st, 2008 (day^{-1} unit).

bots activity is high in the campus network.[6] This is probably because the spam bots activity can be easily 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 (see Figure 1).

In this paper, (1) we carried out statistical and entropy analysis on the PTR resource record (RR) based DNS query packets traffic from the outside for the university campus network, and (2) we discuss on the detection of the spam bots and the host search activity in the campus network.

2. Observations

2.1 Network System and DNS Query Packets Capturing

We investigated on the DNS query packets traffic 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[8] 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 Ethernet-Pro Network Interface Card.

BIND-9.2.6 program package has been employed as a DNS server daemon.[8] The DNS query packets and their query keywords have been captured and decoded by a query logging option (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 contents 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.

2.2 Observed DNS Query Traffic

Firstly, we can demonstrate the total DNS query packets traffic, the A resource record (RR) based DNS query packets traffic, and the PTR RR based DNS query packets traffic from the outside for the campus network through April 1st, 2007 to July 31st, 2008, as shown in Figure 2.

In Figure 2, we can observe significant peaks of (1) April 3rd, (2) 29th, (3) May 24th, (4) July 31st, (5) October 2nd, (6) 9th, (7) 16th, (8) 24th, (9) November 1st, (10) 21st, (11) 26th, (12) December 11th, 2007, (13) January 17th, (14) 24th, (15) February 14th, (16) 21st, (17) 25th, (18) March 13th, (19) May 21st, (20) July 9th, and (21) 23rd, 2008.

In the first group consisting of the peaks (1), (3)-(18), and (21), the similar peaks can be observed in the A and the PTR RRs based DNS query packets traffic curves. In the second group consisting of the peak (2), however, no similar peak can be found in the PTR RR based DNS query packets traffic curve like those in the first group. In the last group consisting of the peak (19) and (20), the peaks in the PTR RR based traffic curve is more sharper than those in the A RR based traffic one. These results show that in Figure 2, the peaks can be categorized into three groups. Previously, we reported these groups corresponding to *random spam bots* (**RSB**), *targeted spam bots* (**TSB**), and *host search* (**HS**), respectively.[9] In other words, the results also indicate that we can detect three security incident models like **RSB**, **TSB**, and **HS**, only observing the DNS resolution traffic from the outside for the campus network.

Consequently, we further carried out entropy analysis on the the PTR RR based DNS query traffic from the outside for the campus network by taking the three incident models into consideration.

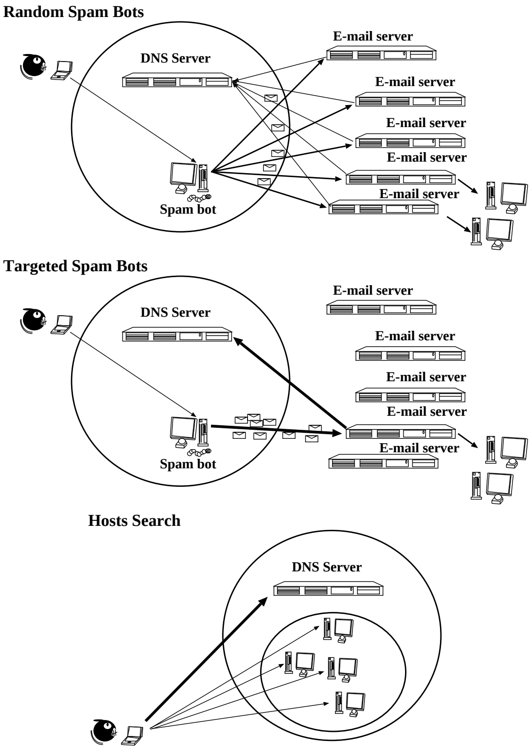

Figure 3. Random spam bots (RSB), targeted spam bots (TSB), and host search (HS)

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.[8] 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.

2.4 Entropy Changes in Spam Bots and Host Search

We define three incidents detection models for random spam bots (RSB) activity, targeted spam bots (TSB) activity, and host search (HS) activity, respectively, as follows:

Random Spam Bots Activity

The unique source IP addresses and the DNS query keywords based entropies simultaneously increase and decrease (in a symmetrical manner), respectively, when the RSB activity increases *i.e.* the RSB activity increases not only randomness for the unique source IP addresses but also frequencies for the specific IP addresses or fully qualified domain names (FQDNs), as query keywords in the DNS query packets traffic. This is because when E-mail servers on the internet are randomly attacked, the E-mail servers will check it out by carrying out the DNS reverse and/or standard resolution access on the detected source IP addresses or FQDNs to the top domain name DNS (**tDNS**) server in the campus network.

Targeted Spam Bots Activity

The unique IP addresses and the DNS query keywords based entropies decrease in a parallel manner when the TSB activity increases *i.e.* the TSB activity increases both frequencies for the specific fully qualified domain names (FQDNs) and the IP addresses. This is because when E-mail servers on the internet undergo a denial of service (DoS) attack, it increases the DNS query packets traffic from the specific spam bots on the campus network.

Host Search Activity

The unique IP addresses and the unique DNS

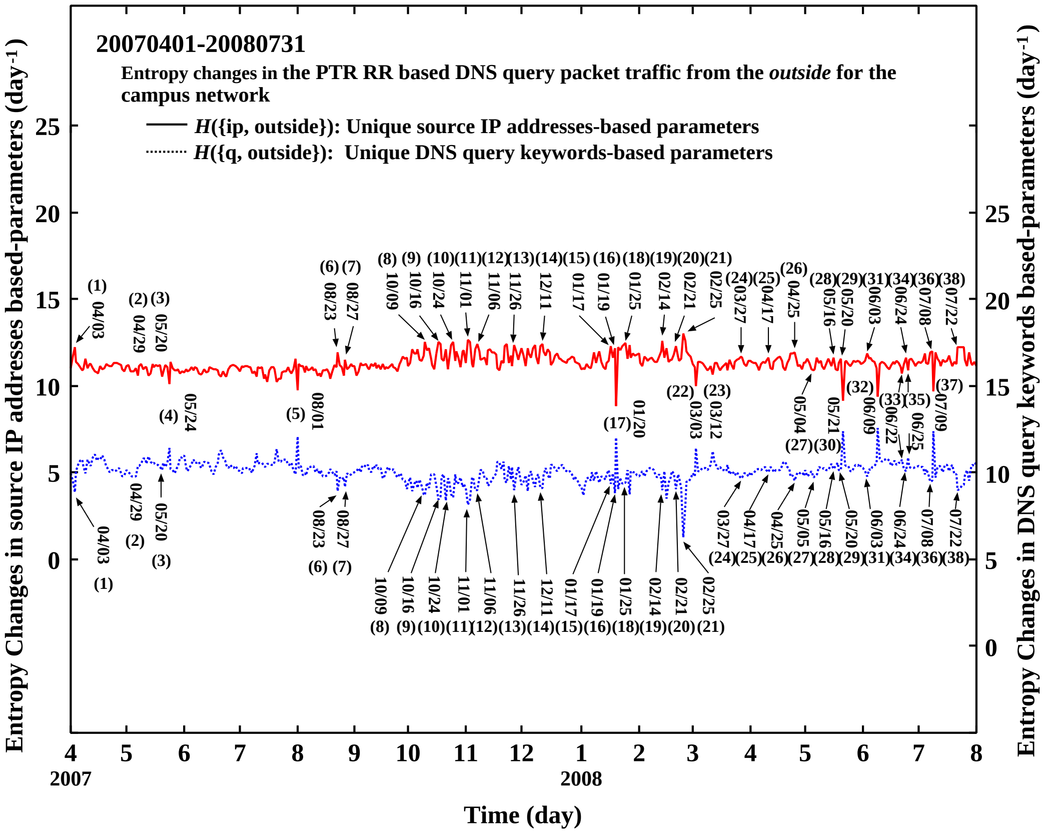

Figure 4. Entropy changes in the total PTR resource record (RR) based DNS query packet traffic from the outside for the campus network to the top domain name system (**tDNS**) server through April 1st, 2007 to July 31st, 2008. The solid and dotted lines show the unique source IP addresses and unique DNS query keywords based entropies, respectively (day^{-1} unit).

query keywords based entropies decrease and increase in anti-symmetrical manner when the host search activity increases. The host search activity is carried out before portsweep which is performed when planning security attack. Usually, the host search activity is achieved from the specific bots/bots clustering network so that the randomness for the unique source IP addresses is low but the randomness for the DNS query keywords is considered to be high.

3. Results and Discussion

3.1 Entropy Changes in the DNS Query Packets Traffic

We performed entropy analysis on the PTR resource record (RR) based DNS query packets traffic (reverse name resolution traffic) from the outside for the campus network through April 1st, 2007 to July 31st, 2008 (Figure 4).

In Figure 4, we can find interesting peaks of (1) April 3rd, (2) 29th, (3) May 20th, (4) 24th, (5) August 1st, (6) 23rd, (7) 27th, (8) October 9th, (9) 16th, (10) 24th, (11) November 1st, (12) 6th, (13) 26th, (14) December 11th, 2007, (15) January 17th, (16) 19th, (17) 20th, (18) 25th, (19) February 14th, (20) 21st, (21) 25th, (22) March 3rd, (23) 12th, (24) 27th, (25) April 17th, (26) 25th, (27) May 4th, 5th, (28) 16th, (29) 20th, (30) 21st, (31) June 3rd, (32) 9th, (33) 22nd, (34) 24th, (35) 25th, (36) July 8th, (37) 9th, (38) 22nd, 2008.

And these peaks are categorized into three types, as: {(1), (6), (7), (8), (9), (10), (11), (12), (14), (15), (16), (18), (19), (20), (21), (24), (25), (26), (28), (29), (31), (34), (36), (38)}, {(2), (3), (27)}, and {(4), (5), (17), (22), (23), (27), (30), (32), (33), (35), (37)}. In the first grouped peaks, the unique source IP addresses based entropy increases but the unique DNS query keywords based one decreases. This shows the random spam bots activity and totally twenty four incidents are detected. In the second grouped peaks, 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 and finally the three targeted spam bots incidents are detected. In the latter group, the unique source IP addresses based entropy decreases but the unique DNS query keywords based one increases. This shows the host search activity and totally the eleven incidents are detected.

From these results, it can be concluded that the random spam bots activity and the host search activity are major incidents and the targeted spam bots activity are minor incidents in the campus network.

4. Conclusions

We carried out entropy analysis on the PTR resource record (RR) based DNS query packet traffic from the *outside* for the campus network through April 1st, 2007 to July 31st, 2008. The following interesting results are obtained, as follows: (1) We can observe totally 38 incidents in the entropy change of the PTR RR based DNS query packets traffic. (2) The security incidents consist of the random spam bots (RSB) of 24 incidents (63%), the targeted spam bots (TSB) of 3 incidents (8%), and the host searches (HS) of 11 incidents (29%). From these results, we can clearly realize that the spam bots and host search activity can be still active and the campus network is also targeted by the attackers in a higher manner.

Acknowledgement

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

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) (b) Kristoff, J., Botnets, *North American Network Operators Group (NANOG32)*, Reston, Virginia (2004), http://www.nanog.org/mtg-0410/kristoff.html

4) McCarty, B.: Botnets: Big and Bigger, *IEEE Security and Privacy*, No.1, pp.87-90 (2003).

5) 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

6) 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).

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

8) BIND-9.2.6:
http://www.isc.org/products/BIND/

9) A. Ludeña Romaña, D., Kubota, S., Sugitani, K., and Musashi, Y.: DNS based Entropy and Forensic Analysis on the PCs for Learners in a University, *IPSJ SIG Technical Reports, the 1st Internet and Operational Technologies (IOT01)*, Vol. 2008, No.37, pp.103-108 (2008).