Entropy Study on A Resource Record DNS Query Traffic from the Campus Network

Kazuya Takemori,[†] Wei Juan Kong,[††] Dennis Arturo Ludeña Romaña,[††]
Shinichiro Kubota,[‡] Kenichi Sugitani,[‡] and Yasuo Musashi [‡]

Abstract: We investigated the source IP address (SIP)- and query keyword (QK)-based entropy changes in the A and PTR resource records (RRs) based DNS query traffic between the DNS clients and the campus DNS server through January 1st to December 31st, 2008. The results are: (1) The both entropies simultaneously decrease when the targeted attack activity is high. (2) The SIP-based entropy increases while the QK-based one decreases, simultaneously, when the random attack activity is high. (3) The SIP-based entropy decreases while the QK-based one increases, at the same time, when the host search activity is high. Therefore, we can get important information for the security incidents by only observing the DNS query traffic.

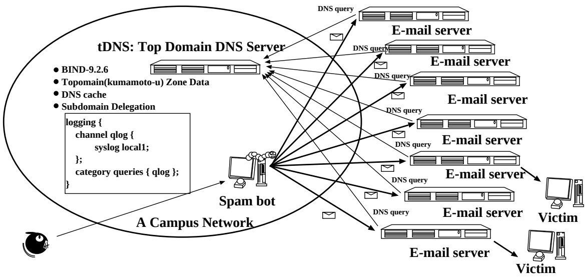
Keywords: DNS based detection, DNS traffic entropy, spam bots, host search

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]

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]

Then, we reported previously that the unique DNS query keyword based entropy in the PTR resource record (RR) based DNS query packet traffic from the outside for the campus network decreases considerably while the unique source IP addresses based entropy increases when the random spam bots activity is high in the campus network.[6] This is probably because the PTR RR based DNS query packet traffic was generated by the spam bots activity sensors like the spam filters of the E-mail server and/or the intrusion detection/prevention system (IDS/IPS) on the Internet. Therefore, we can detect spam bots activity, especially, a random spam bot (RSB) in the campus network, by watching the DNS query packet traffic from the other sites on the Internet (see Figure 1). We also reported that we observed not only an increase in the unique DNS query keyword based entropy in the PTR RR based DNS query packet traffic from the Internet but a decrease in the unique source IP address based one in the DNS query packet traffic when performing host search activity from the Internet.[7]



```
logging {
   channel qlog {
         syslog local1;
   };
   category queries { qlog };
}
```

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

In this paper, (1) we carried out entropy analysis on the A and the PTR resource records (RRs) based DNS query packet traffic from a university campus network and the Internet through January

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

1st to December 31st, 2008, and (2) we assessed the bot attack detection rate among the entropies for the A RR-, and the PTR-RR based DNS query packet traffic from the campus network and the Internet.

2. Observations

2.1 Network Systems and DNS Query Packet Capturing

We investigated on the DNS query request packet traffic between the top domain (**tDNS**) DNS server and the DNS clients. Figure 1 shows an observed network system in the present study, which consists of the **tDNS** server and the PC clients as a random spam bots in the campus network, and the victim E-mail mail servers on the Internet. The **tDNS** server is one of the top level domain name (kumamoto-u) system servers and plays an important role of domain name resolution including DNS cache function 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 the 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 the **tDNS** server, the BIND-9.2.6 program package has been employed as a DNS server daemon.[8] The DNS query packet and their query keywords have been captured and decoded by a query logging option (see Figure 1 and the named.conf manual of the BIND program in more detail). The log of DNS query packet access has been recorded in the syslog files. All of the syslog files are daily updated by the cron system. The line of syslog message 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, or a mail exchange (MX RR) type.

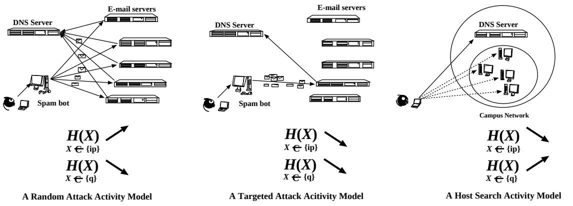


Figure 2. Random spam bots (RSB), targeted spam bots (TSB), and host search (HS) activity models

2.2 Estimation of Entropy

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

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

where X is the data set of the frequency $freq(j)$ of IP addresses or that of the DNS query keyword in the DNS query packet traffic from the outside of the campus network, and the probability $P(i)$ is defined, as

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

where i and j $(i, j \in X)$ represent the unique source IP address or the unique DNS query keyword in the DNS query packet, and the frequency $freq(i)$ are estimated with the script program, as reported in our previous work.[9]

2.3 Attack Activity Models

We define three incidents detection models for random attack (RA) activity, targeted attack (TA) activity, and host search (HS) activity (See Figure 2), respectively.

A random attack (RA) activity model – since a random spam bot (RSB), a typical example for the RA activity model, randomly attacks various victim E-mail servers, the E-mail servers can try to check IP addresses and fully qualified domain names (FQDNs) for the RSB, with referring to the top domain DNS (**tDNS**) server in the campus network. This causes an increase in the number of the unique source IP addresses in the DNS query traffic but a decrease in the number of the unique DNS query keyword *i.e.* the unique source IP addresses- and the unique DNS query keyword-based entropies simultaneously increase and decrease, respectively, when the RA activity is high in the campus network.

A targeted attack (TA) activity model – since the targeted spam bot (TSB), for example, attacks a small number of specific victim E-mail servers in the campus network or on the Internet, the E-mail servers can check IP addresses and FQDNs for the TSB, with referring to the **tDNS** server in the campus network. This causes decreases in the unique IP addresses- and the DNS query keyword-based entropies when the TA activity is high.

A host search (HS) activity model – The host search activity can be mainly carried out by a small number of IP hosts on the Internet or in the campus network like bot compromised PCs. Since these IP hosts send a lot of the DNS reverse name resolution (the PTR RR based DNS query) request packets to the **tDNS** server, the unique IP addresses- and the unique DNS query-keywords based entropies decrease and increase, respectively.

Here, we should also define thresholds for detecting these three kinds of malicious activity models, as setting to 1,000 day^{-1} for the frequencies of the top-ten unique source IP addresses or the DNS query keywords. The evaluation for threshold was previously reported.[10]

3. Results and Discussion

3.1 Entropy Changes in the A and PTR RRs DNS Query Packet Traffic from the Campus Network

We demonstrate the calculated unique source IP address and unique DNS query keywords based entropies for the A and PTR resource records (RRs) based DNS query request packet traffic from the campus network to the top domain DNS (**tDNS**) server through January 1st to December 31st, as shown in Figure 3.

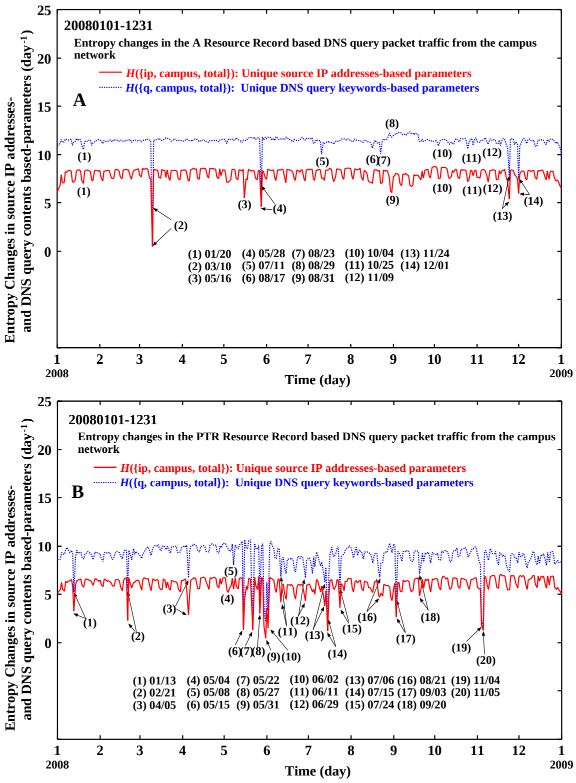


Figure 3. Entropy changes in the total A and PTR resource records (RRs) based DNS query request packet traffic from the campus network to the top domain DNS (**tDNS**) server through January 1st to December 31st, 2008. The solid and dotted lines show the unique source IP addresses and unique DNS query keywords based entropies, respectively (day^{-1} unit).

In Figure 3A, we can find fourteen peaks and they are categorized into three groups, as: {(1)-(2), (4)-(7), (10)-(14)}, {(3)}, and {(8), (9)}. In the first peak group, all the peaks show simultaneous decreases in the unique source IP addresses- and unique DNS query keyword based entropies. This means that the peaks can be assigned to a targeted attack (TA) activity model. At the day in the peak (2), for instance, we found that a specific Windows PC transmitted 12,019,758 DNS query packets, in which we found only one specific FQDN as "download.windows.update.com". In the peak (3) of the second peak group, the unique source IP address based entropy decreases only. We investigated the cause of the peak (3) and we ultimately found a communication problem in the DNS cache system and the syslog system. In the last peak group, we can observe an increase in the DNS query keyword based entropy and a decrease in the source IP address based one for the peaks (8) and (9), respectively. In the peak (8), we found 64,894 unique FQDNs in the A RR based DNS query traffic. Simultaneously, we also found the specific E-mail server which transmitted 102,850 A RR based DNS query request packets including 18,927 unique FQDNs (usually 5,274 unique FQDNs). We interviewed the administrators for the E-mail server and we found that they tried repeatedly to optimize parameters for a spam filter.

In Figure 3B, we can observe significant twenty peaks and these peaks are grouped into one type, as: {(1)-(20)}. This result indicates that all the peaks can be assigned to a TA activity model. In the peaks (1), (5), (11), (19), and (20), we observed that a specific PC host 133.95.a1.100 transmitted 36,694, 22,600, 66,810, 302,103, and 485,241 packets, respectively, and their query keywords mainly consist of the specific IP addresses as: 202.***.51.141 (20,662 day^{-1}), 202.***.142.118 (22,497 day^{-1}), 149.***.120.34 (20,272 day^{-1}), 118.***.8.121 (297,608 day^{-1}), and 118.***.8.121 (485,172 day^{-1}), respectively. The specific PC host is an Windows machine. In the peaks (2), (3), (14), (15), (16), and (17), we obtained that a specific PC host 133.95.a2.63 sent 188,063, 58,728, 59,1630, 106,670, 37,707, and 177,790 packets, respectively. The specific PC host is a Mac OS X 10.4.11 machine. In the peaks (6), (7), (8), (9), and (10), we found that a specific PC host 133.95.a3.105 launched 277,360, 355,185, 121,941, 710,615, and 327,806 packets, respectively. This specific PC host can be a broadband router. In the peaks (12) and (13), the specific PC hosts were found as 133.95.a4.126. The PC host is a Linux machine as a VoIP server in which we found misconfiguration in the VoIP services. In the other peaks (4) and (18), we found the specific PC hosts of 133.95.a5.66 and 133.95.a6.148, respectively.

Interestingly, we can notice a fact that the dates for the peaks are almost mutually different *i.e.* we can get useful but different information from the entropy changes in the A and PTR RRs based DNS query request packet traffic from the campus network.

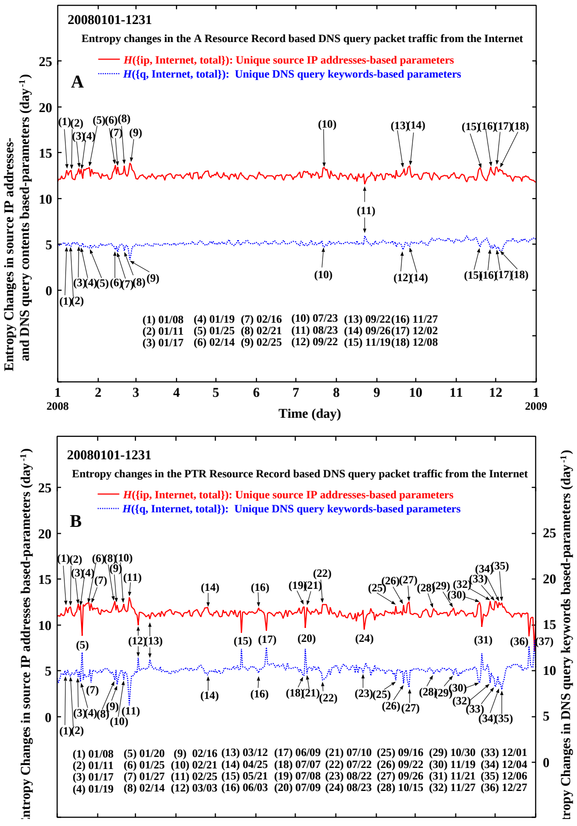


Figure 4. Entropy changes in the total A and PTR resource records (RRs) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server through January 1st to December 31st, 2008. The solid and dotted lines show the unique source IP addresses and unique DNS query keywords based entropies, respectively (day^{-1} unit).

3.2 Entropy Changes in the A and PTR RRs DNS Query Packet Traffic from the Internet

We illustrate the calculated unique source IP address and unique DNS query keywords based entropies for the A and PTR resource records (RRs) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server through January 1st to December 31st, as shown in Figure 4.

In Figure 4A, we can observe eighteen peaks and they can be grouped into two types, as: {(1)-(10), (13)-(18)} and {(11)}. We can observe the seventeen peaks in which all the peaks show an increase in the unique source IP addresses based entropy and a decrease in the unique DNS query keywords based one *i.e.* this feature indicates random attack (RA) activity like a random spam bot (RSB) activity.

In the latter group, we can find only a peak (11) which demonstrates a decrease in the unique source IP addresses based entropy but a increase in the unique DNS query keywords based one. Although this feature seems to be host search (HS) activity, we can observe the HS activity only in the PTR RR based DNS query request packet traffic. In fact, we had a half-day blackout through August 22nd to 23rd, 2008, because the campus electrical facilities were inspected for defects. Therefore, the one half-day blackout affected the DNS query request packet traffic.

As shown in Figure 4B, we can observe thirty seven peaks and they can be categorized into three groups, as: {(1)-(4), (6)-(11), (14), (16), (18)-(19), (21)-(22), (25)-(30), (32)-(35)}, {(5), (12)-(13), (15), (17), (20), (31), (36)-(37)}, and {(23), (24)} in which the first, the second, and the last groups take twenty six, nine, and two peaks, respectively.

In the first peak group, all the peaks show an increase in the unique source IP address based entropy and a decrease in the unique DNS query keywords based one, showing the RA like a RSB activity. Also, we can observe the similar peaks for the RA activity at the same dates in Figures 4A and 4B. Furthermore, the peaks in Figure 4B are slightly sharper than those in Figures 4A *i.e.* the found RA activity can be a random spam bot (RSB) activity. This is because the RSB transmits a lot of spam E-mails to the E-mail servers on the Internet, the E-mail server gets an IP address of the connected SMTP engine and it tries to get an FQDN for the IP address (reverse name resolution) and to get an IP address again from the obtained FQDN. Therefore, if the E-mail server has received a spam E-mail from the campus network, the top domain DNS server can receive at least two DNS query packets from the DNS cache server for the victim E-mail server in which the former and latter packets can be PTR and A RRs based DNS query request packets. In addition, these results are very different from those for the entropy analysis on the A and PTR RRs based DNS query request packet traffic from the campus network (see Figures 3A and 3B). In the peak (3), (4), (6), and (7), for instance, we investigated the detected PC hosts which are PC room terminals and administrated by a computer center, however, no evidence can be directly found in the PC hosts. Fortunately, we successfully interviewed an account holder and finally, we detected an *auto.inf* file, a Win32/Agent.BUL Trojan Horse (TH), in the USB stick type disk storage for the account holder. In the peak (16), we detected the IP address for a broadband router in the campus laboratory in which there were three Windows XP PCs. We investigated all the PCs by the typical anti-virus software but no virus infection could be found. Then, we checked the SMTP activity in the PC by executing a netstat command on the DOS window and we could detected a lot of SMTP connections. Therefore, we could conclude that the PC hosts was hijacked to be a random spam bot (RSB).

In the second peak group, nine peaks can be found. The unique source IP addresses based entropy decreases while the unique DNS query keywords based one increases. This feature indicates the host search (HS) activity. It is very important to detect the HS activity because the HS activity is mainly performed as pre-investigation on the campus network for the next cyber attack.

In the last peak group, we can observe two peaks (23) and (24). The peaks can be assigned to August 22nd and 23rd, 2008, respectively. This is because we had a half-day blackout through August 22nd to 23rd, 2008, because of inspection of electrical defects, affecting the entropy changes.

4. Conclusions

We carried out entropy analyses on the total A and the PTR resource records (RRs) based DNS query packet traffic from the campus network and the Internet through January 1st to December

31st, 2008. The following results are obtained, as: (1) We can observe 14, 20, 18, and 37 security incidents in the entropy changes of the A and PTR resource records (RRs) based DNS query traffic from the campus network and the A and PTR RRs based DNS query one from the Internet, respectively. (2) In the entropy changes of the A RR based DNS query packet traffic from the campus network, the incidents consist of the TA activity of 12 incidents (86%) and the other 2 incidents (14%). (3) In the entropy changes of the PTR RR based DNS query packet traffic from the campus network, the incidents consist of the TA activity of 20 incidents (100%). (4) In the entropy changes of the A RR based DNS query packet traffic from the Internet, the incidents consist of the RA activity of 17 incidents (94%) and the other 1 incident (6%). (5) In the entropy changes of the PTR RR based DNS query packet traffic from the Internet, the incidents consist of the RA activity of 26 incidents (70%), the HS activity of 9 incidents (24%), and the other 2 incidents (6%).

From these results, it is concluded that we can detect security incidents like the TA activity in the campus network by observing entropy changes of the A and PTR RRs based DNS query traffic from the campus network, and the RA activity and the HS activity in the campus network by observing the entropy changes of the A and the PTR RRs based DNS query packet traffic from the Internet.

Acknowledgement

All the studies were carried out in CMIT of Kumamoto University and this study is supported by the Grant aid of Graduate School Action Scheme for Internationalization of University Students (GRASIUS) in Kumamoto University.

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) 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) Ludeña Romaña, D. A., Sugitani, K., and Musashi, Y.: DNS Based Security Incidents Detection in Campus Network, *International Journal of Intelligent Engineering and Systems*, Vol. 1, No.1, pp.17-21 (2008).

7) Ludeña Romaña, D. A., Kubota, S., Sugitani, K., and Musashi, Y.: Entropy Study on A and PTR Resource Record-Based DNS Query Traffic, *IPSJ Symposium Series*, Vol. 2008, No.13, pp.55-61 (2008).

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

9) Ludeña Romaña, D. A., Musashi, Y., Matsuba, R., and Sugitani, K.: Detection of Bot Worm-Infected PC Terminals, *Information*, Vol. 10, No.5, pp.673-686 (2007).

10) Ludeña Romaña, D. A., 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).

11) Musashi, Y., Matsuba, R., and Sugitani, K.: Development of Automatic Detection and Prevention Systems of DNS Query PTR record-based Distributed Denial-of-Service Attack, *IPSJ SIG Technical Reports*, Vol. 2004, No.77, pp.43-48 (2004).