Detection of Host Search Activity in Domain Name Reverse Resolution Traffic

Min Lei,[†1] Yasuo Musashi,[†2] Dennis Arturo Ludeña Romaña,[†1]
Kazuya Takemori,[†1] Shinichiro Kubota[†2]
and Kenichi Sugitani[†2]

We statistically investigated the total PTR resource record (RR) based DNS query request packet traffic from the Internet to the top domain DNS server in a university campus network through January 1st to August 31st, 2009. The obtained results are: (1) We observed twelve host search (HS) activities that we can observe rapid decreases in the unique source IP address based entropy of the inbound PTR RR based the DNS query packet traffic and significant increases in the unique DNS query keyword based one. (2) We found the consecutive IP addresses in the PTR RR based DNS query request packet traffic through July 9th, 2009. hAlso (3), we calculated Euclidian distances between the observed IP address and the next observed IP address as the DNS query keywords and we detected eleven HS activities by a range of thresholds for 1.0-2.0. Therefore, these results show that we can detect more easily the inbound HS activity by calculating the Euclidian distances among the observed IP addresses in the inbound PTR RR based DNS query request packet traffic.

1. Introduction

It is of considerable importance to raise up a detection rate of Bots, since they become components of the bot clustered networks that are used to transmit a lot of unsolicited mails including 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 search activity of Internet worms (IWs) like an W32/Blaster or an W32/Witty worm, since the both worms drastically change entropy after starting their activity.[5)] Then, we reported previously that in the inbound PTR resource record (RR) based DNS query request packet traffic, the unique source IP address based entropy decreases considerably while the unique DNS query keyword based one increases when the host search (HS) activity is high.[6),7)] The HS activity is recognized to be a pre-investigation activity or a harvesting activity for collecting fully qualified domain names (FQDNs) of the university campus and/or enterprise networks *i.e.* after the HS activity, the attacker can concentrate to check out the vulnerabilities in the targeted servers or hosts.

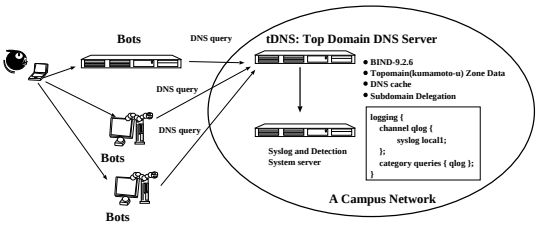

Fig. 1 A schematic diagram of the observed network in the present study.

In this paper, (1) we carried out entropy and Euclidian distance based analyses on the total PTR resource record (RR) based DNS query request packet traffic from the Internet through January 1st to August 31st, 2009, and (2) we assessed the both results for entropy and Euclidian distance based analyses on the the PTR-RR based DNS query packet traffic.

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 bots like a host search bot or a spam bot in the campus or enterprise network, and the victim hosts like

†1 Graduate School of Science and Technology, Kumamoto University
†2 Center for Multimedia and Information Technologies, Kumamoto University

the DNS servers on the campus network. 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 network 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.

2.2 Estimation of DNS Query Traffic Entropy

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

$$\mathbf{H}(\mathbf{X}) = -\sum_{\mathbf{i}\in\mathbf{X}} \mathbf{P}(\mathbf{i}) \log_{\mathbf{2}} \mathbf{P}(\mathbf{i}) \qquad (1)$$

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

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

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

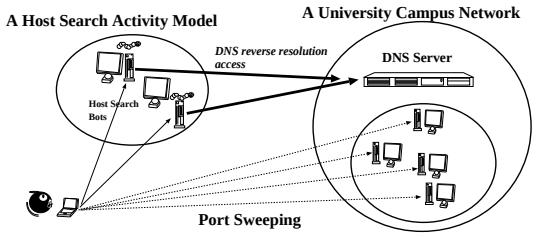

Fig. 2 A host search (HS) activity model.

2.3 Host Search Activity Model

We define here a host search (HS) model (See Figure 2). *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 like a directory harvesting attack. 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 the HS activity, as setting to 1,000 packets 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.[9]

2.4 Estimation of Euclidian Distances Among IP addresses as DNS Query Keywords

The Euclidian distances, $\mathbf{d}(\mathbf{IP_{i+1}}, \mathbf{IP_i})$, are calculated, as

$$\mathbf{d}(\mathbf{IP_{i+1}}, \mathbf{IP_i}) = (\sum_{\mathbf{j=0}}^{\mathbf{3}} (\mathbf{x_{i+1,j}} - \mathbf{x_{i,j}})^{\mathbf{2}})^{\frac{\mathbf{1}}{\mathbf{2}}} \qquad (3)$$

where $\mathbf{IP_i}$ and $\mathbf{IP_{i+1}}$ are the IP address $(\mathbf{i})$ and the next IP address $(\mathbf{i})$ of the DNS query keywords, respectively, and where $\mathbf{x_{i,0}}$, $\mathbf{x_{i,1}}$, $\mathbf{x_{i,2}}$, and $\mathbf{x_{i,3}}$ correspond to the IPv4 address like A.B.C.D, respectively. For instance, if an IP address is 192.168.1.1, the vector $(\mathbf{x_{i,0}}, \mathbf{x_{i,1}}, \mathbf{x_{i,2}}, \mathbf{x_{i,3}})$ can be represented as $(192, 168, 1, 1)$. The detection is decided by thresholds $\mathbf{d_{min}} = 1.0$ and $\mathbf{d_{max}} = 2.0$, as

$$\mathbf{d_{min}} \leq \mathbf{d}(\mathbf{IP_{i+1}}, \mathbf{IP_i}) \leq \mathbf{d_{max}} \qquad (4)$$

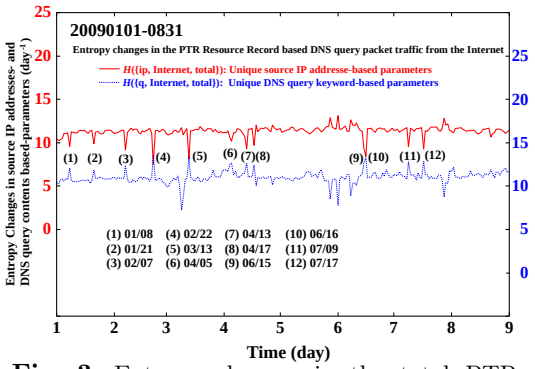

Fig. 3 Entropy changes in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server through January 1st to August 31st, 2009. The solid and dotted lines show the unique source IP addresses and unique DNS query keywords based entropies, respectively (day^{-1} unit).

3. Results and Discussion

3.1 Entropy Changes in Total PTR-RRs DNS Query Packet Traffic from the Internet

We demonstrate the calculated unique source IP address and unique DNS query keyword based entropies for the PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server through January 1st to August 31st, as shown in Figure 3.

In Figure 3, we can find twelve peaks and these peaks of (1)-(12) correspond to January 8th, 21st, February 7th, 22nd, March 13th, April 5th, 13th, 17th, June 15th, 16th, July 9th, and 17th, 2009, respectively, in which all the peaks show significant increase and decrease in the unique source IP address- and the unique DNS query keyword based entropies, respectively. This result indicates that all the peaks (1)-(12) can be assigned to the HS activity.

In the peak (11), at July 9th, 2009, we investigated the DNS query keywords in the total inbound PTR RR based DNS query packet traffic and the results are shown Figure 4. In Figure 4, we can view scenery that the IP address as DNS query keyword is consecutively incremented.

Therefore, it has a possibility that this

```
Jul  9 09:04:59 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.178 IN PTR
Jul  9 09:04:59 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.177 IN PTR
Jul  9 09:04:59 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.179 IN PTR
Jul  9 09:04:59 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.181 IN PTR
Jul  9 09:04:59 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.180 IN PTR
Jul  9 09:05:00 kun named[5017]: client 91.***.***.145#7709: query: 133.95.3.242 IN PTR
Jul  9 09:05:01 kun named[5017]: client 210.***.**.249#24099: query: 133.95.106.34 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.182 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.183 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.184 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.185 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.186 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.187 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.188 IN PTR
Jul  9 09:05:01 kun named[5017]: client 193.**.***.110#4280: query: 133.95.128.189 IN PTR
```

Fig. 4 Changes in the IP address as the DNS query keywords in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server at July 9th, 2009.

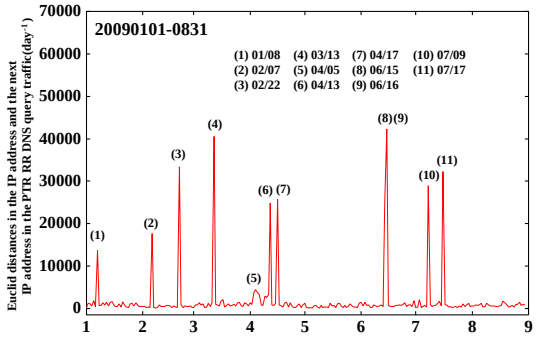

Fig. 5 Changes in Euclidian distance between the IP address and the next IP address, as the unique DNS query keywords in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server through January 1st to August 31st, 2009 (day^{-1} unit).

consecutive increment of the IP address can be useful to detect the HS activity in the PTR RR based DNS query request packet traffic.

3.2 Euclidian Distance among IP addresses as DNS Query Keywords

We illustrate the calculated Euclidian distance, $\mathbf{distance}(\mathbf{IP_{i+1}}, \mathbf{IP_i})$, between the IP address and the next IP address, as the unique DNS query keywords in the PTR resource record (RR) based DNS query request packet traffic from the

Internet to the top domain DNS (**tDNS**) server through January 1st to August 31st, as shown in Figure 5.

In Figure 5, we can observe eleven significant peaks (1)-(11) being allocated to January 8th, February 7th, 22nd, March

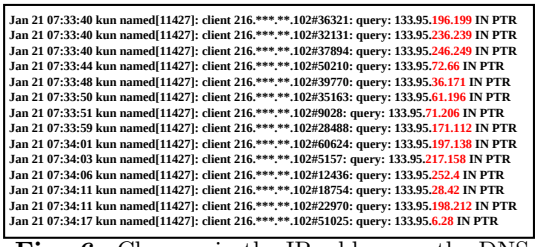

Fig. 6 Changes in the IP address as the DNS query keywords in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (**tDNS**) server at January 21st, 2009.

13th, April 5th, 13th, 17th, June 15th, 16th, July 9th, and 17th, 2009, respectively.

In Figure 3, we can observe the peak (2), corresponding to January 21st, 2009, however, we can find no peak for January 21st, 2009, in Figure 5. Thus, we investigated in more detail the DNS query keywords in the total inbound PTR RR based DNS query packet traffic at January 21st, 2009, and the results are shown Figure 6. In Figure 6, we can watch a scene that the IP address as DNS query keyword is discontinuously or randomly changed.

As a result, the discontinuous or randomized DNS query keywords cause a factor for disappearing of the peak at January 21st, 2009, in Figure 5.

4. Conclusions

We investigated entropy and Euclidian distance based analyses on the total inbound PTR resource record (RR) based DNS query request packet traffic through January 1st to August 31st, 2009. The following interesting results are found: (1) we observed twelve HS activities in the entropy change in the PTR RR based DNS query request packet traffic, and (2) we found eleven HS activities in the Euclidian distances between the IP address and the next IP address in the PTR RR based DNS query request packet traffic *i.e.* the relative detection rate is unfortunately decreased when employing the Euclidian distance based method for detecting the HS activity in the total inbound PTR-RR based DNS query packet traffic.

From these results, it is concluded that we should continue further study to develop and improve the Euclidian distance based HS activity detection technology .

Acknowledgments 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) No. 165240040213 in Kumamoto University.

References

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 2005)*, Linköping, Sweden, 2005, pp.172-177
6) Ludeña Romaña, D. A. and Musashi, Y.: DNS Based Analysis of DNS Query Traffic in the Campus Network, *Journal of Systemics, Cybernetics and Informatics*, Vol. 6, No.5, pp.42-44 (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).