Detection of Host Search Activity in PTR Resource Record Based DNS Query Packet Traffic

Yasuo Musashi,* Florent Hequet,† Dennis Arturo Ludeña Romaña,† Shinichiro Kubota,* and Kenichi Sugitani*

**Center for Multimedia and Information Technologies*
Kumamoto University
2-39-1 Kurokami, Kumamoto, JAPAN, 860-855
musashi@cc.kumamoto-u.ac.jp

†Graduate School of Science and Technology
Kumamoto University
2-39-1 Kurokami, Kumamoto, JAPAN, 860-855

***Abstract* - 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 December 31st, 2009. The obtained results are: (1) We observed fourteen host search (HS) activities in which 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 and random IP address based queries in the PTR RR based DNS query request packet traffic through the days of January 8th and 21st, 2009, respectively. Also (3), we calculated Euclidean distances between the observed IP address and the last observed IP address as the DNS query keywords and we detected two kinds of HS activities by employing both threshold ranges of 1.0-2.0 and 150.2-210.4, respectively. Therefore, these results show that we can detect the HS activity by calculating the Euclidean distances between the currently- and the last-observed IP addresses in the inbound PTR RR based DNS query request packet traffic.**

Keywords: DNS based detection, host search, host name harvesting attack, anomaly detection, bots

I. 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 of 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.

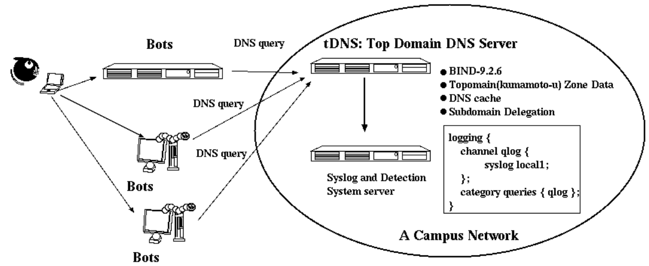

Fig. 1 A schematic diagram of a network observed in the present study.

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

II. OBSERVATION

A. 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 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.

B. Estimation of DNS Query Traffic Entropy

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

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

where X is the data set of the frequency freq(j) of a unique IP address or that of a unique DNS query keyword in the DNS query packet traffic from the Internet, and the probability P(i) is defined, as

$$P(i) = freq(i)/(\sum_{j} freq(j)) \quad (2)$$

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].

C. 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 or 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, simultaneously.

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].

D. Estimation of Euclidean Distances of IP addresses as DNS Query Keywords

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

$$d(IP_i, IP_{i-1}) = \left(\sum_{j=1}^{4} (x_{i,j} - x_{i-1,j})^2\right)^{\frac{1}{2}} \quad (3)$$

where both IP_i and IP_{i-1} are the current IP address i and the last IP address i-1 of the DNS query keywords, respectively,

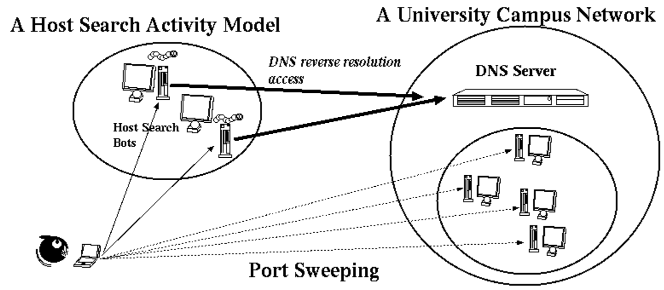

Fig. 2 A host search (HS) activity model.

and where $x_{i,1}$, $x_{i,2}$, $x_{i,3}$, and $x_{i,4}$ correspond to an IPv4 address like A.B.C.D, respectively. For instance, if an IP address is 192.168.1.1, the vector ($x_{i,1}$, $x_{i,2}$, $x_{i,3}$, $x_{i,4}$) can be represented as (192.0, 168.0, 1.0, 1.0). The detection is decided by thresholds d_{min}=1.0 and d_{max}=2.0, as

$$d_{min} \le d(IP_i, IP_{i-1}) \le d_{max} \quad (4)$$

III. Results and Discussion

A. 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 record (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2009, as shown in Figure 3.

In Figure 3, we can find fourteen significant peaks and these peaks (1)-(14) correspond to January 8th, 21st, February 7th, 22nd, March 13th, April 5th, 13th, 17th, June 15th, 16th, July 9th, 17th, December 13th, and 30th, 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)-(14) can be assigned to the HS activity.

In the peak (1), at January 8th, 2009, we investigated the DNS query keywords in the total inbound PTR RR based DNS query packet traffic and the results are shown in 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 consecutive increment of the IP address can be useful to detect the HS activity in the PTR RR based DNS query request packet traffic.

B. Euclidean Distances in Consecutive Incremental DNS Reverse Queries

We illustrate the calculated Euclidean distance (1.0 < $\mathbf{d(IP_i, IP_{i-1})}$ < 2.0) between the current IP address IP_i and the last IP address IP_{i-1}, 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 December 31st, 2009, as shown in Figure 5.

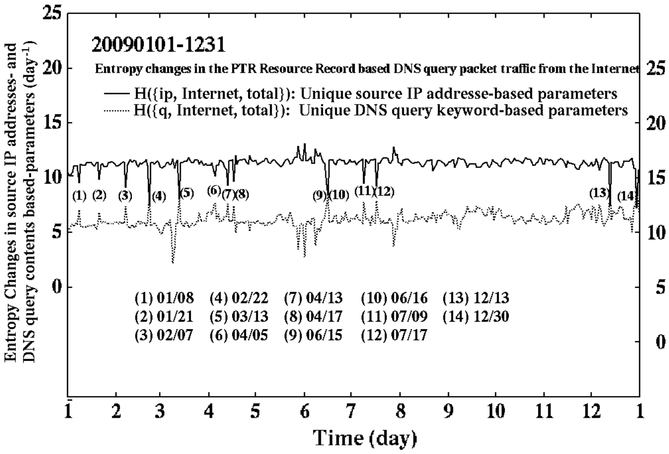

Fig. 3 Entropy changes in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2009. The solid and dotted lines show the unique source IP addresses and unique DNS query keywords based entropies,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

```
Jan 8 14:53:15 kun named[10018]: client 66.***.***.*#14981: query: 133.95.199.106 IN PTR
Jan 8 14:53:15 kun named[10018]: client 66.***.***.*#36073: query: 133.95.199.107 IN PTR
Jan 8 14:53:16 kun named[10018]: client 66.***.***.*#38238: query: 133.95.199.108 IN PTR
Jan 8 14:53:16 kun named[10018]: client 66.***.***.*#21441: query: 133.95.199.114 IN PTR
Jan 8 14:53:16 kun named[10018]: client 66.***.***.*#46887: query: 133.95.199.115 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#22561: query: 133.95.199.116 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#26389: query: 133.95.199.117 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#45189: query: 133.95.199.118 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#37530: query: 133.95.199.119 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#12411: query: 133.95.199.128 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#53097: query: 133.95.199.129 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#14727: query: 133.95.199.130 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#46036: query: 133.95.199.131 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#32521: query: 133.95.199.132 IN PTR
Jan 8 14:53:17 kun named[10018]: client 66.***.***.*#12967: query: 133.95.199.133 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 January 8th, 2009.

In Figure 5, we can observe thirteen significant peaks (1)-(13) being allocated to January 8th, February 7th, 22nd, March 13th, April 5th, 13th, 17th, June 15th, 16th, July 9th, and 17th, December 13th, and 30th, 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.

C. Euclidean Distances in Random Sequential DNS Reverse Queries

The campus IP addresses are represented as $133.95.x_i.y_i$ in which both x_i and y_i can take numbers from 0 to 255, as: $0 \le x_i \le 255$ and $0 \le y_i \le 255$ *i.e.* the following eq 5 is obtained employing the both newly defined variables (x_i y_i) and eq 3, as:

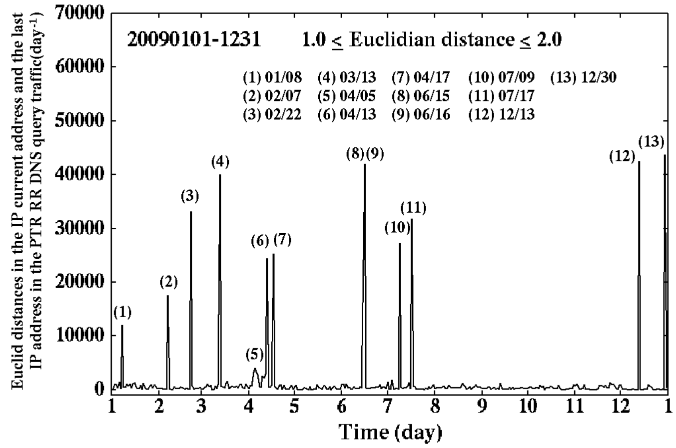

Fig. 5 Changes in Euclidean distance between the current IP address and the last IP address, as the unique DNS query keywords in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2009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

```
Jan 21 07:33:40 kun named[11427]: client 216.***.**.102#36321: query: 133.95.196.199 IN PTR
Jan 21 07:33:40 kun named[11427]: client 216.***.**.102#32131: query: 133.95.236.239 IN PTR
Jan 21 07:33:40 kun named[11427]: client 216.***.**.102#37894: query: 133.95.246.249 IN PTR
Jan 21 07:33:44 kun named[11427]: client 216.***.**.102#50210: query: 133.95.72.66 IN PTR
Jan 21 07:33:48 kun named[11427]: client 216.***.**.102#39770: query: 133.95.36.171 IN PTR
Jan 21 07:33:50 kun named[11427]: client 216.***.**.102#35163: query: 133.95.61.196 IN PTR
Jan 21 07:33:51 kun named[11427]: client 216.***.**.102#9028: query: 133.95.71.206 IN PTR
Jan 21 07:33:59 kun named[11427]: client 216.***.**.102#28488: query: 133.95.171.112 IN PTR
Jan 21 07:34:01 kun named[11427]: client 216.***.**.102#60624: query: 133.95.197.138 IN PTR
Jan 21 07:34:03 kun named[11427]: client 216.***.**.102#5157: query: 133.95.217.158 IN PTR
Jan 21 07:34:06 kun named[11427]: client 216.***.**.102#12436: query: 133.95.252.4 IN PTR
Jan 21 07:34:11 kun named[11427]: client 216.***.**.102#18754: query: 133.95.28.42 IN PTR
Jan 21 07:34:11 kun named[11427]: client 216.***.**.102#22970: query: 133.95.198.212 IN PTR
Jan 21 07:34:17 kun named[11427]: client 216.***.**.102#51025: query: 133.95.6.28 IN PTR
```

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.

$$d(IP_i, IP_{i-1}) = ((x_i - x_{i-1})^2 + (y_i - y_{i-1})^2)^{\frac{1}{2}} \qquad (5)$$

where $(x_i-x_{i-1})^2$ or $(y_i-y_{i-1})^2$ takes a range from 0 to 255^2 *i.e.* the range of the Euclid distance, $\mathbf{d(IP_i, IP_{i-1})}$, should be between 0.0 to $\sqrt{255^2+255^2}$ (~ 360.6).

If the both variables are random sequences, the Euclid distance, $\mathbf{d(IP_i, IP_{i-1})}$, can also take a random sequence. Also, if the random sequence follows the Gaussian distribution, the probability for the Euclid distance takes a maximum value between at 180.3 (~ 360.6/2) with a standard deviation of 30.1 (~ 360.6/12) because of the central limit theorem *i.e.* d_{min} and d_{max} should take values of 150.2 (~ 180.3-31.1) and 210.4 (~ 180.3+31.1).

We investigated the frequency distribution of the Euclid distance $\mathbf{d(IP_i, IP_{i-1})}$ for the specific source IP address in the inbound PTR RR DNS query traffic (Figure 7).

In Figure 7, the aggregated peaks can be observed in the range of $\mathbf{d(IP_i, IP_{i-1})}$ ~ 180. This feature indicates that although the frequency distribution takes an incomplete shape of Gauss distribution, it has a possibility that we detect the random sequence based HS activity in a high rate when employing the central limit theorem.

We demonstrate the calculated Euclidean distance (150.2

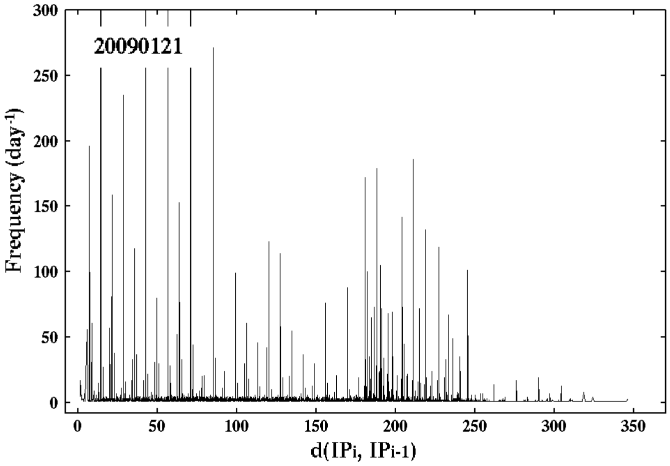

Fig. 7 Frequency distribution of the Euclidean distance **d(IP_i, IP_{i-1})** between the current IP address and the last IP one, as the DNS query keywords in the specific inbound PTR-resource records (RR) based DNS query request packet traffic from the HS attacker in January 21st, 2009 (day^{-1} unit).

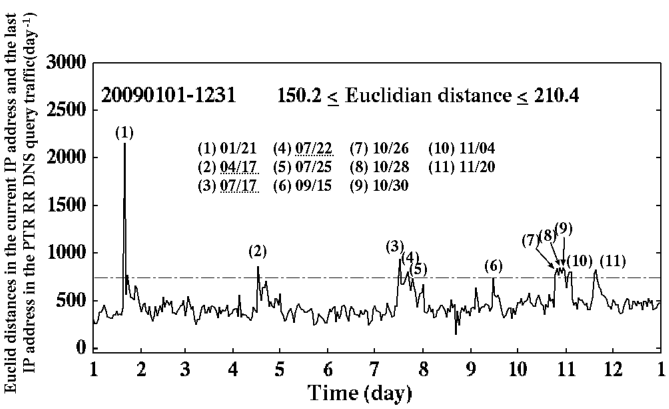

Fig. 8 Changes in Euclidean distance between the current IP address and the last IP address, as the unique DNS query keywords in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2009 (day^{-1} unit).

≤ **d(IP_i, IP_{i-1})** ≤ 210.4) between the current IP address IP_i and the last IP address IP_{i-1}, 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 December 31st, 2009, as shown in Figure 8.

In Figure 8, we can observe eleven significant peaks (1)-(11) being allocated to January 21st, April 17th, July 17th, 22nd, 25th, October, 26th, 28th, 30th, November 4th, and 20th, 2009, respectively. Especially, the peak (1) is very sharp and it is also assigned to the peak at January 21st, 2009, which is observed in Figure 3 but disappeared in Figure 5.

From these results, it can be concluded that in order to raise the detection rate of the HS activity, we should combine the both consecutive- and random-IP address sequences based detection technologies.

```
1  #!/bin/tcsh -f
2  set Threshold=10
3  # Step 1 Learning to produce a low-diemnsianl
4  cat /var/log/querylog | \ clgrep -v -cclients.conf
5  grep "IN PTR" | arpa | \
6  sdis 0.0 0.0 | qdis 1.0 2.0 150.2 210.4 | tr '#' ' ' | \
7  awk '{print $7}' | sort -r | uniq -c | sort -r | \
8  awk '{printf("%s\t%s\n",$2,$1);}' |  qdos $Threshold | \
9  awk '{print $1}' >tmpfile
10 # Step 2 Detection
11 cat /var/log/querylog | clgrep -ctmpfile | \
12 grep "IN PTR" | arpa >HSdet.log
13 # Step 3 Scoring
14 cat HSdet.log | wc -l >>HSdetScore.txt
15 exit 0
```

Fig. 9 Suggested Algorithm and Script Code.

D. Detection Algorithm for Host Search Activity

We suggest the following detection algorithm of the Host Search (HS) activity and we show a prototype program (see Figure 9):

── Step 1 *Learning to produce a low-dimensional*── In this step, the **clgrep** and **grep** commands extract inbound PTR RR based DNS query request packet messages from the DNS query log file (*/var/log/querylog*), the **arpa** command converts the reverse query format "D.C.B.A.in-addr.arpa" into the usual IPv4 format "A.B.C.D" (A, B, C, and D represent digit numbers of {0-255}), the **sdis** command prints out a syslog message if the Euclidean distance between the two source IP addresses is calculated to be zero, the **qdis** command prints out the syslog message if the Euclidean distance **d(IP_i, IP_{i-1})** takes ranges of 1.0-2.0 and 150.2-210.4, respectively, and the **awk**, **sort**, **uniq**, and **qdos** commands (lines 7 to 9 in Figure 9) compute the frequencies of the Euclidean distance **d(IP_i, IP_{i-1})** and if the frequency exceeds a threshold value (*Threshold*=10), they write out the candidate IP addresses into a *tmpfile* as training data.

── Step 2 *Detection* ── In the next step, the **clgrep**, **grep**, and **arpa** commands extract the HS activity related messages in the DNS query log file (*/var/log/querylog*), using the training data (*tmpfile*) and they generate only an HS activity related DNS query log file (*HSdet.log*).

── Step 3 *Scoring* ── In the final step, the **wc** command calculates the score for the detection of the HS activity in the file *HSdet.log*, and it writes out the detection score into a score file (*HSdetScore.txt*).

E. Evaluation

We calculated the score for the detection of the HS activity in the inbound PTR RR based DNS query request packet traffic through January 1st to December 31st, 2009 (Figure 10).

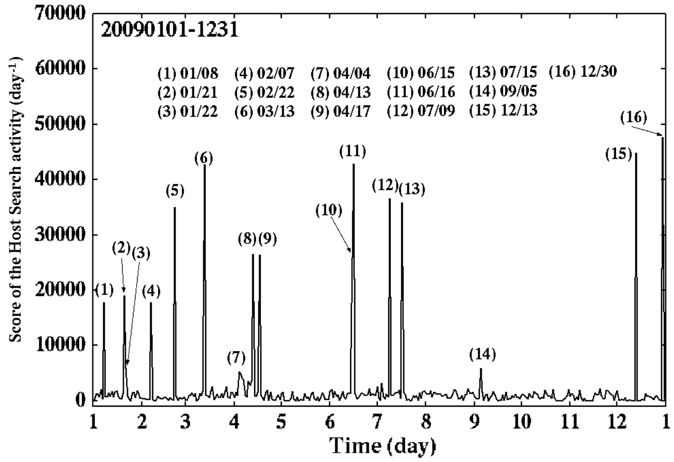

Fig. 10 Changes in score of the Host Search activity detection in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2009 (day^{-1} unit).

In Figure 10, we can observe sixteen peaks (1)-(16) that are assigned to January 8th, 21st, 22nd, February 7th, 22nd, March 13th, April 5th, 13th, 17th, June 15th, 16th, July 9th, 17th, September 5th, December 13th, and 30th, 2009, respectively.

Expectedly, in Figure 10, the score peaks (2) emerge again, corresponding to random IP address query based HS activity. Also, we can find the other new score peaks (3) and (14). The score peak (3) is assigned to be the next day of the score peak (2), however, the HS activity based DNS query request packet traffic mainly consists of the random IP address query based one as well as the consecutive IP address based one, partially. The last peak (14) is allocated to be consecutive IP address based HS activity.

Furthermore, we calculated the number of the HS activity in 2009 versus the thresholds and the results are shown in Figure 11. In Figure 11, the number of the HS activity starts to change gradually or in a milder manner, after a threshold value of 2,000 day^{-1}. This feature indicates that the threshold value should take more than 2,000 day^{-1}. This is because that the number of the HS activity drastically increases when decreasing the threshold value *i.e.* this means that the false positive considerably increases when employing less than 2,000 day^{-1}, as a threshold value.

IV. CONCLUSIONS

We investigated entropy and Euclidean distance based analyses on the total inbound PTR resource record (RR) based DNS query request packet traffic through January 1st to December 31st, 2009. The following interesting results are found: (1) we observed fourteen considerable host search (HS) activities in the entropy change in the PTR RR based DNS query request packet traffic, and (2) we found the thirteen consecutive incremental IP address- and the eleven random sequence of the IP address-based host search (HS) activities in the Euclidean distances between the current IP

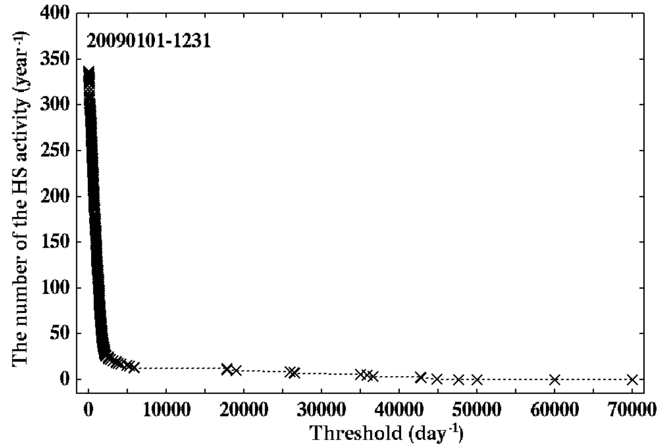

Fig. 11 The number of the HS activity in 2009 versus the thresholds and the data through January 1st to December 31st, 2009.

address and the last IP address in the PTR RR based DNS query request packet traffic.

From these results, it is concluded that we can detect the two typical HS activities by observing the Euclidean distance between the current and the last IP addresses in the total inbound PTR RR based DNS query packet traffic to the campus top level domain DNS server, from the Internet.

ACKNOWLEDGMENT

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] 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] J. Kristoff: Botnets, *North American Network Operators Group (NANOG32)*, Reston, Virginia (2004), http://www.nanog.org/mtg-0410/kristoff.html
[4] B. McCarty: Botnets: Big and Bigger, *IEEE Security and Privacy*, No. 1, 2003, pp. 87-90.
[5] A. Wagner and B. Plattner: 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] D. A. Ludeña Romaña and Y. Musashi: DNS Based Analysis of DNS Query Traffic in the Campus Network, *Journal of Systemics, Cybernetics and Informatics*, Vol. 6, No. 5, 2008, p.42-44.
[7] D. A. Ludeña Romaña, S. Kubota, K. Sugitani, and Y. Musashi: Entropy Study on A and PTR Resource Record-Based DNS Query Traffic, I*PSJ Symposium Series*, Vol. 2008, No. 13, 2008, p.55-61.
[8] BIND-9.2.6: http://www.isc.org/products/BIND/
[9] D. A. Ludeña Romaña, Y. Musashi, R. Matsuba, and K. Sugitani: Detection of Bot Worm-Infected PC Terminals, *Information*, Vol. 10, No. 5, 2007, pp. 673-686.