Trends in Host Search Attack in DNS Query Request Packet Traffic

Nobuhiro Shibata

Graduate School of Science and Technology
Kumamoto University
2-39-1 Kurokami, Central W., Kumamoto,
JAPAN, 860-855
shibatan@st.cs.kumamoto-u.ac.jp

Yasuo Musashi, Dennis Arturo Ludena Romana,
Shinichiro Kubota, and Kenichi Sugitani

Center for Multimedia and Information Technologies
Kumamoto University
2-39-1 Kurokami, Central W., Kumamoto,
JAPAN, 860-855
{musashi,dennis,kubota,sugitani}@cc.kumamoto-u.ac.jp

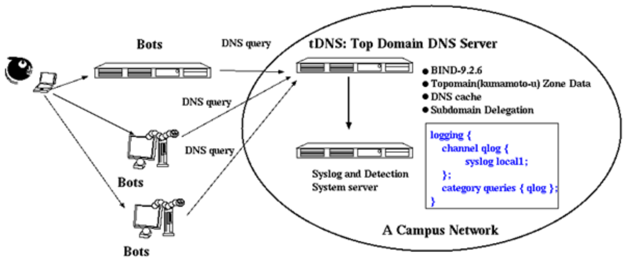
***Abstract*— We statistically investigated the total PTR resource record (RR) based DNS query request packet traffic from the Internet to the top domain DNS server in a university campus network through January 1st to December 31st, 2011. The obtained results are: (1) We found twelve host search (HS) attacks in the scores for detection method using the calculated Euclidean distances between the observed IP address and the last observed IP address in the DNS query keywords by employing both threshold ranges of 1.0-2.0 (consecutive) and 150.2-210.4 (random). However, we found nineteen HS attacks in the scores using the calculated cosine distance between the DNS query IP addresses (threshold ranges of 0.75-0.83 and 0.9-1.0). (3) In the newly found HS attacks, we observed that the source IP addresses of the HS attack DNS query packets are distributed. Therefore, it can be concluded that the cosine distance based detection technology has a possibility to detect the source IP address-distributed host search attack.**

Keywords-DNS Host Search Attack; DNS Log Analysis; Advanced Persistent Threats

I. INTRODUCTION

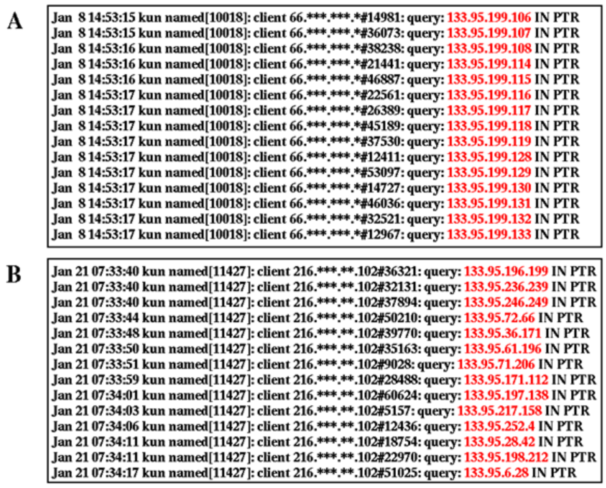
The host search (HS) attack [1] is recognized to be a pre-investigation attack or a harvesting attack of fully qualified domain names (FQDNs) of the university campus and/or enterprise networks i.e. after the HS attack, the attacker can concentrate to check out the vulnerabilities in the targeted servers or hosts in order to carry out the advanced persistent threat (APT) attack [2].

Previously, we reported development and evaluation of the Euclidian distance based detection model system for the HS attack against the campus top domain name system (tDNS) server [3] and it currently works still well for detecting the single source IP address-based HS attack, however, recently, the attackers upgraded their strategies that they started to a distributed source IP address HS attack to evade the detection system i.e. it is required to develop a new system for detecting the distributed source IP address based HS attack.



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

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



A

```
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
```

B

```
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
```

Figure 2. 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 (A) and 21st (B), 2009.

In this paper, (1) we carried out Euclidean and cosine distances based analyses on the source IP address and the DNS query IP address in the total PTR resource record (RR) based DNS query request packet traffic from the Internet through January 1st to December 31st, 2011, and (2) we assessed the both results for the Euclidean and cosine distances 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 5.5 Final) in which the kernel-2.6.18 is currently employed with the Intel Xeon X5660 2.8 GHz 6 Cores dual node system, the 16GB core memory, and Intel Corporation EthernetPro 82575EB Gigabit Ethernet Controller.

In the tDNS server, the BIND-9.3.6-P1 program package has been employed as a DNS server daemon [4]. The DNS query request 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 request 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 request 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 an E-mail exchange (MX RR) type.

B. *Host Search Attack Model*

— A host search (HS) attack model — the host search (HS) attack can be mainly carried out by a small number of IP hosts like virtual machines on the public cloud. 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 [1].

We also investigated the IP address change in the PTR RR based DNS query request packet traffic through January 8th and 21st, 2009, and the results are shown in Figure 2. In Figure 2A, at January 8th, 2009, 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 attack in the PTR RR based DNS query request packet traffic (consecutive model). In Figure 2B, at January 21st, 2009, we can see it that the IP address as DNS query keyword is discontinuously or randomly changed (random model).

From these results, we need to take into consideration on the consecutive and the random IP address query keyword based models in order to develop an HS attack detection system.

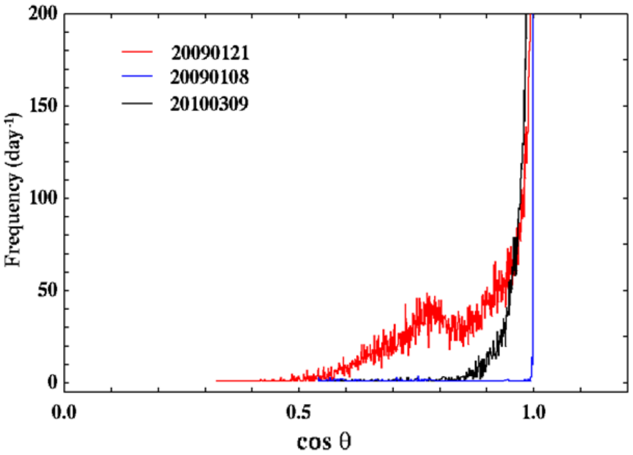


Figure 3. Frequency distributions of the cosine distance at January 10th, 21st, and March 9th, 2010 (day^{-1} unit)..

C. *Euclidean- and Cosine-Distances of query IP addresses in DNS Query Request Packets*

The Euclidean distance, ed(qIP_i, qIP_{i-1}), is calculated, as

$$\mathrm{ed}(\mathrm{qIP}_i, \mathrm{qIP}_{i-1}) = \sqrt{\sum_{j=1}^{4} (x_{i,j} - x_{i-1,j})^2} \qquad (1)$$

where both qIP_i and qIP_{i-1} are the current query IP address i and the last query IP address i-1, respectively, 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 $(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})^T$ can be represented as $(192.0, 168.0, 1.0, 1.0)^T$.

If the HS attack model follows the consecutive DNS query keyword based model, the detection is decided by thresholds ed_{min}=1.0 and ed_{max}=5.0 [5], as

$$\mathrm{ed}_{\mathrm{min}}(=1.0) \le \mathrm{ed}(\mathrm{qIP}_i, \mathrm{qIP}_{i-1}) \le \mathrm{ed}_{\mathrm{max}}(=5.0) \qquad (2)$$

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 $0 \le y_i \le 255$, and $(x_i\text{-}x_{i-1})^2$ or $(y_i\text{-}y_{i-1})^2$ takes a range from 0 to 255^2 i.e. the range of ed(qIP_i, qIP_{i-1}), should be from 0.0 to $\sqrt{255^2+255^2}$ (~ 360.6). If ed(qIP_i, qIP_{i-1}) follows the Gaussian distribution, the probability for the Euclidian 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. ed_{min} and ed_{max} should take values of 150.2 (~ 180.3-31.1) and 210.4 (~ 180.3+31.1) [5].

The cosine distance cd(qIP_i, qIP_{i-1}) is obtained, as

$$\cos\theta = \mathrm{cd}(\mathrm{qIP}_i, \mathrm{qIP}_{i-1}) = \frac{\mathrm{qIP}_{i-1}^{T} \bullet \mathrm{qIP}_i}{\left|\mathrm{qIP}_{i-1}^{T}\right| \left|\mathrm{qIP}_i\right|} \qquad (3)$$

where the cosine distance takes a range from 0.413 (cd_{min}) to 1.000 (cd_{max}), since cd_{min} is estimated from cosine distance between vectors $(133, 95, 0, 0)^T$ and $(133, 95, 255, 255)^T$.

```
 1  #!/bin/tcsh -f
 2  set Threshold=10
 3  # Step 1 Reduction of the Noise
 4  cat /var/log/querylog | clgrep -v -cclients.conf  | \
 5  grep "IN PTR" | arpa | \
 6  awk '{print $9}' | sort -r | uniq -c | sort -r | \
 7  awk '{printf("%s\t%s\n",$2,$1);}' | \
 8  qdos 1000 >noise.conf
 9  # Step 2 Learning to produce a low-diemnsianl
10  cat /var/log/querylog | clgrep -v -cclients.conf | \
11  grep "IN PTR" | arpa | \
12  cngrep -v -Dnoise.conf | \
13  sdis 0.0 0.0 | \
14  qdis 1.0 5.0 150.2 210.4 | \
15  tr '#' ' ' | awk '{print $7}' | sort -r | uniq -c | sort -r | \
16  awk '{printf("%s\t%s\n",$2,$1);}' |  qdos $Threshold | \
17  awk '{print $1}' >tmpfile
18  # Step 3 Detection
19  cat /var/log/querylog | clgrep -ctmpfile | \
20  grep "IN PTR" | arpa >HSdet.log
21  # Step 4 Scoring
22  cat HSdet.log | wc -l >>HSdetScore.txt
23  exit 0
```

Figure 4. Host Search Attack Detection Algorithm for Euclidian distance.

```
13  sdis 0.0 0.0 | \
14  qdis -yz 0.73 0.83 0.9 1.0 | \
```

Figure 5. Host Search Attack Detection Algorithm for Cosine Distance.

In Figure 3, we show the calculated frequency distribution of the cosine distances at January 8th (consecutive model), 21st (random model; normal distribution), 2009, and March 9th, 2010 (random model; exponential distribution). The frequency distribution at January 21st, 2009, has a significant peak at a range between 0.73 and 0.83 as well as quick increasing from 0.9 to 1.0, indicating that the thresholds cd_{min} and cd_{max} should take ranges between 0.73-0.83 and 0.9-1.0.

$$cd_{min}(= 0.73 \text{ or } 0.9) \le cd(qIP_i, qIP_{i-1}) \le cd_{max}(= 0.9 \text{ or } 1.0) \quad (4)$$

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 4):

—Step 1 *Reduction of the Noise*— In this step, the **clgrep** and **grep** commands extract the 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 top **awk** commands and the two **sort** and one **uniq** commands calculate and print out the query IP addresses and their frequencies, and the **qdos** command prints out the query IP addresses and the frequencies into the *noise.conf* file when the frequencies are greater than 1,000 day^{-1}.

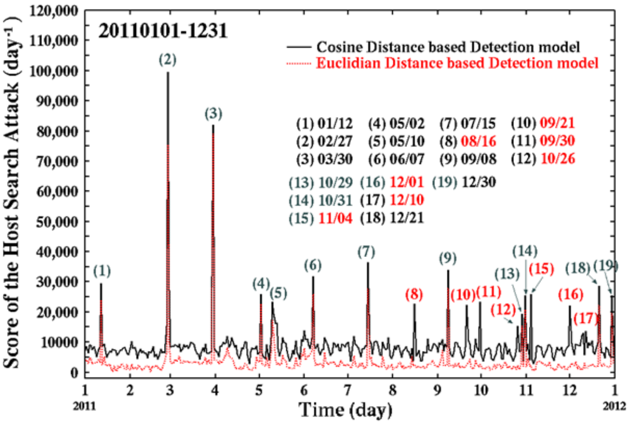


Figure 6. Changes in score of the host search (HS) attack detection in the total PTR resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2011 (day^{-1} unit). The solid- and the dotted-curves show scores for the cosine distance- and the Euclidian distance-based detection models, respectively.

—Step 2 *Learning to produce a low-dimensional*— In this step, the **clgrep**, **grep**, and **arpa** commands take the same functions as ones in Step 1, the **cngrep** command discards the IP addresses listed in the *noise.conf* file from the syslog messages, the **sdis** command prints out a syslog message if the Euclidean distance ed(sIP_i, sIP_{i-1}) between the two source IP addresses is calculated to be zero, the **qdis** command prints out the syslog message if the Euclidean distance ed(qIP_i, qIP_{i-1}) between the two query IP addresses takes ranges of 1.0-2.0 and 150.2-210.4, or the **qdis -yz** command (Figure 5) pints out the syslog messages if the cosine distance cd(qIP_i, qIP_{i-1}) takes ranges of 0.73-0.82 and 0.9-1.0, and the **awk**, **sort**, **uniq**, and **qdos** commands (lines 15 to 17 in Figure 5) compute the frequencies of the detected source IP addresses and if the frequency exceeds a threshold value (*Threshold*=10), they write out the candidate source IP addresses into a *tmpfile* as training data.

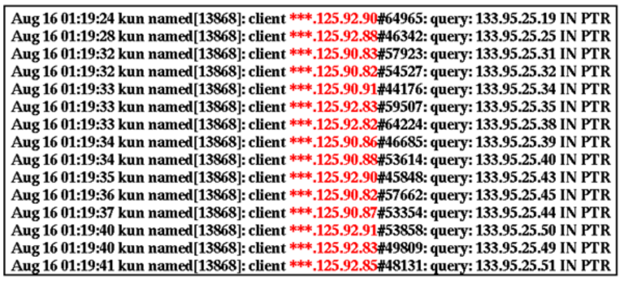
—Step 3 Detection— In the next step, the **clgrep**, **grep**, and **arpa** commands extract the HS attack related messages in the DNS query log file (*/var/log/querylog*), using the training data (*tmpfile*) and they generate only an HS attack related DNS query log file (*HSdet.log*).

—Step 4 Scoring— In the final step, the **wc** command calculates the score for the detection of the HS attack activity in the file *HSdet.log*, and it writes out the detection score into a score file (*HSdetScore.txt*).

III. Results and Discussion

A. Euclidean distance- and cosine distance-based Host Search Detection Model

We illustrate the calculated score of the host search (HS) attack using Euclidean- and cosine-distance based detection models ($1.0 \le ed(qIP_i, qIP_{i-1}) \le 2.0$ or $150.2 \le ed(qIP_i, qIP_{i-1}) \le 210.4$.) or (($0.73 \le cd(qIP_i, qIP_{i-1}) \le 0.83$ or $0.9 \le cd(qIP_i, qIP_{i-1}) \le 1.0$)) between the current query IP address qIP_i and the last query IP address qIP_{i-1}, as the DNS query keywords



```
Aug 16 01:19:24 kun named[13868]: client ***.125.92.90#64965: query: 133.95.25.19 IN PTR
Aug 16 01:19:28 kun named[13868]: client ***.125.92.88#46342: query: 133.95.25.25 IN PTR
Aug 16 01:19:32 kun named[13868]: client ***.125.90.83#57923: query: 133.95.25.31 IN PTR
Aug 16 01:19:32 kun named[13868]: client ***.125.90.82#54527: query: 133.95.25.32 IN PTR
Aug 16 01:19:33 kun named[13868]: client ***.125.90.91#44176: query: 133.95.25.34 IN PTR
Aug 16 01:19:33 kun named[13868]: client ***.125.92.83#59507: query: 133.95.25.35 IN PTR
Aug 16 01:19:33 kun named[13868]: client ***.125.92.82#64224: query: 133.95.25.38 IN PTR
Aug 16 01:19:34 kun named[13868]: client ***.125.90.86#46685: query: 133.95.25.39 IN PTR
Aug 16 01:19:34 kun named[13868]: client ***.125.90.88#53614: query: 133.95.25.40 IN PTR
Aug 16 01:19:35 kun named[13868]: client ***.125.92.90#45848: query: 133.95.25.43 IN PTR
Aug 16 01:19:36 kun named[13868]: client ***.125.90.82#57662: query: 133.95.25.45 IN PTR
Aug 16 01:19:37 kun named[13868]: client ***.125.90.87#53354: query: 133.95.25.44 IN PTR
Aug 16 01:19:40 kun named[13868]: client ***.125.92.91#53858: query: 133.95.25.50 IN PTR
Aug 16 01:19:40 kun named[13868]: client ***.125.92.83#49809: query: 133.95.25.49 IN PTR
Aug 16 01:19:41 kun named[13868]: client ***.125.92.85#48131: query: 133.95.25.51 IN PTR
```

Figure 7. Changes in the source IP address in the total PTR-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server at August 16th, 2011.

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, 2011, as shown in Figure 6.

In Figure 6, we can observe nineteen significant peaks (1)-(19) being allocated to (1) January 12th, (2) February 27th, (3) March 30th, (4) May 2nd, (5) 10th, (6) June 7th, (7) July 15th, (8) August 16th, (9) September 8th, (10) 21st, (11) 30th, (12) October 26th, (13) 29th, (14) 31st, (15) November 4th, (16) December 1st, (17) 10th, (18) 21st, and (19) 30th, 2011, respectively.

In the score curve for the Euclidian distance based detection model, we can find no peaks corresponding to the peaks (8), (10), (11), (12), (15), (16), and (17) in the score curve for the cosine distance based detection model. This result shows that the cosine distance based detection technology is much precisely than the Euclidian distance based one or much false positive. Thus, we also investigated the source IP address- and query IP address-changes in the PTR RR based DNS query request packet traffic through August 16th, 2011, and the results are shown in Figure 7.

In Figure 7, we can view scenery that the source IP addresses change periodically and the query IP addresses are incremented like in a consecutive manner, showing that the cosine distance based detection model can be useful for detecting the source IP address distributed host search (HS) attack like a distributed denial of service (DDoS) attack.

B. Frequency Distribution of the Euclidian distance in Source IP addresses

We calculated frequency distributions of the Euclidian distance for the five peaks (8), (10), (11), (12), and (15), as shown in Figure 8. In Figure 8, the each frequency distribution has a significant peak and all the peaks take a range of 1.0-5.0. This feature indicates that the HS attacks in the peaks (8), (10)-(12), and (15), are surely an IP address distributed HS attack like a conventional DDoS attack.

Also, it can be concluded that the HS attacker has changed their strategy in an IP distributed manner after August 16th, 2011.

IV. Conclusions

We developed and evaluated the Euclidean- and cosine-distance based detection models of the source IP address

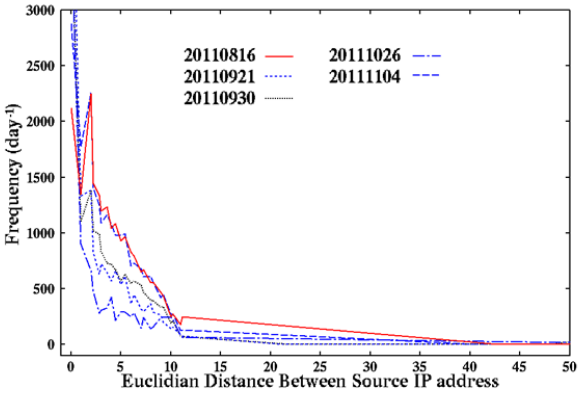


Figure 8. Frequency distributions of the Euclidian distance between the source IP addresses at August 16th, September 21st, 30th, October 26th, and November 4th, 2011 (day^{-1} unit).

distributed host search (HS) attack in the total inbound PTR resource record (RR) based DNS query request packet traffic through January 1st to December 31st, 2011. The following interesting results are found: (1) we observed nineteen peaks for host search (HS) attacks in the score changes of the cosine distance based HS attack detection model, however, (2) we found the only twelve peaks in the score changes of the conventional Euclidian based HS attack detection model, (3) we observed the source IP addresses were not fixed but changed periodically in the newly found peaks, and (4) we investigated frequency distribution of the source IP addresses in the newly found peaks. These results show that the cosine distance based detection model can detect the source IP address distributed HS attack and it has a possibility that the conventional Euclidian distance based detection model also can detect more precisely when taking into consideration the distributed source IP addresses.

Acknowledgment

All the studies were carried out in Centre for Multimedia and Information Technologies (CMIT) of Kumamoto University. We gratefully thank all the CMIT staffs and all the members of Kumamoto University.

References

[1] K. Takemori, D. A. Ludeña R., S. Kubota, K. Sugitani, Y. Musashi: Detection of NS Resource Record DNS Resolution Traffic, Host Search, and SSH Dictionary Attack Activities, *International Journal of Intelligent Engineering and Systems*, Vol. 2, No. 4, 2009, pp.35-42.

[2] R. Bejtlich: Understanding the advanced persistent threat, *Information Security magazine online, 2010,* http://searchsecurity.techtarget.com/magazineContent/Understanding-the-advanced-persistent-threat

[3] Y. Musashi, F. Hequet, S. Kubota, and K. Sugitani: Detection of Host Search Activity in PTR Resource Record Based DNS Query Packet Traffic, *Proceedings for the Sixth International Conference on Information and Automation (ICIA2010),* Harbin, Heilongjiang, China, 2010, pp.1284-1288.

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