Vol.2011-CSEC-53 No.1
Vol.2011-IOT-13 No.1
2011/5/12

Detection of DNS Cache Poisoning Attack in DNS Standard Resolution Traffic

Yasuo Musashi,[†] Kazuya Takemori,[††] Shinichiro Kubota,[†] and Kenichi Sugitani[†]

We statistically investigated the total A 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, 2010. The obtained results are: (1) We found five DNS Cache Poisoning (DNSCP) attacks in observation of rapid decrease in the unique source IP address based entropy of the DNS query packet traffic and significant increase in the unique DNS query keyword based one. (2) Also, we found five DNSCP attacks in the score changes for detection method using the calculated restricted Damerau-Levenshtein distance (restricted edit distance) between the observed query keyword and the last one by employing both threshold ranges through 1 to 40. Therefore, it is possible that the restricted Damerau-Levenshtein distance based detection technology can detect the DNSCP attacks.

1. Introduciton

It is of considerable importance to raise 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 phishing and spam mails, or to execute distributed denial of service attacks [1-4].

Recently, the phishing mail almost includes URLs to make the victim users connect to the fraud sites in which the DNS cache poisoning (DNSCP) attack is one of the technologies to create the online fraud sites. Kaminsky attack [5] is most recently developed DNSCP attack technology and the attacker sends a lot of unique DNS queries like the A resource record (RR) based DNS query request packets to the DNS cache server to raise the probability of the DNSCP attack (see Figure 1). From this point, it is required to develop the Kaminsky attack detection system.

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) attack is high [6]. Similarly, we can detect the Kaminsky attack by calculating the entropy changes in the A RR based DNS query request packet traffic. Also, we developed the Euclidian distance based detection technology, since the DNS queries for the HS attack include IP addresses so that we can calculate the Euclidian distance between the current IP address vector (IP_i) and the previous IP address one (IP_{i-1}) [7, 8]. However, we have to employ edit distance like a Levensthein distance [9], or a Hamming distance [10], to calculate the similarity between two strings in the DNS queries of the Kaminsky attack traffic.

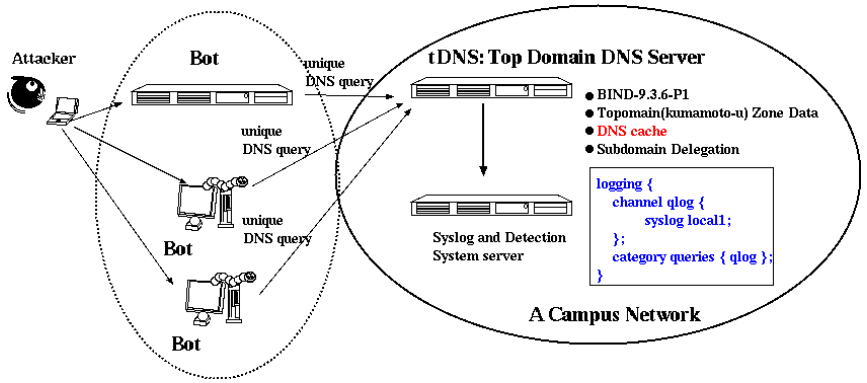

Figure 1 A schematic diagram of a network observed in the present study.

In this paper, (1) we carried out entropy and restricted Damerau-Levenshtein distance based analyses on the inbound A resource record (RR) based DNS query request packet traffic from the Internet through January 1st to December 31st, 2010, and (2) we assessed the both results for entropy and restricted Damerau-Levenshtein distance [9, 11] based analyses on the fully qualified domain names (FQDNs) as the query keywords in the A RR based DNS query packet traffic.

2. Observation

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 Kaminsky attack 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) 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 82575EB Gigabit Ethernet

[†] Center for Multimedia and Information Technologies (CMIT), Kumamoto University
[††] NRI Secure Technologies, Ltd.

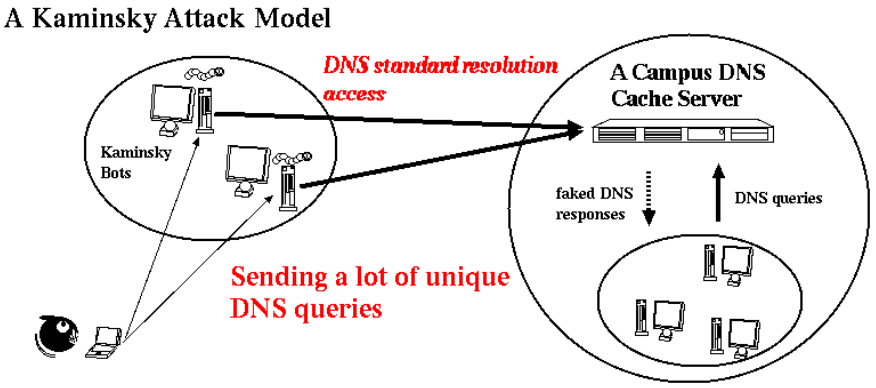

Figure 2 A Kaminsky DNS cache poisoning attack model.

Controller.

In the tDNS server, the BIND-9.3.6-P1 program package has been employed as a DNS server daemon [12]. 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 message 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-resource record (RR)) 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 value H(X), a

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

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 request packet traffic from the Internet, and the probability P(i) is defined, as

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

where i and j (i, j ∈ X) represent the unique source IP address or the unique DNS query keyword in the DNS query request packet, and the frequency freq(i) are estimated with the script program, as reported in our previous work [6].

```
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl100.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl100.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl100.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl101.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl101.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl101.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl102.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl102.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl103.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl103.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl103.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl104.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl104.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl104.***.com IN A +
Jan 26 00:12:17 kun named[20611]: client 222.**.**.***#44642: query: sl102.***.com IN A +
```

Figure 3 Changes in the fully qualified domain names (FQDNs) as the DNS query keywords in the total A-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server at January 26th, 2010.

2.3 Kaminsky Attack Model

We define here a Kaminsky attack model (See Figure 2).

— *A Kaminsky attack model* — the Kaminsky attack 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 hijacked network servers. Since these IP hosts send a lot of the DNS standard name resolution (the A RR based DNS query) request packets to the tDNS server, the source IP address- and the unique DNS query-keyword based entropies decrease and increase, simultaneously.

Here, we should also define thresholds for detecting the Kaminsky attack, as setting to 40,000 packets (day^{-1}) for the frequencies of the top ten unique source IP addresses or the DNS query keywords.

We also investigated the DNS query keyword change in the A RR based DNS query request packet traffic at January 26th, 2010, and the results are shown in Figure 3. In Figure 3, we can view scenery that the fully qualified domain names (FQDNs) as DNS query keyword are consecutively incremented. Therefore, it has a possibility that this consecutive increment of the FQDNs can be useful to detect the Kaminsky attack in the A RR based DNS query request packet traffic.

From these results, we need to take into consideration on the consecutive query keyword based model in order to develop a Kaminsky attack detection system *i.e.* we also suggest hereafter the restricted Damerau-Levenshtein (edit) distance [13] based detection system of the Kaminsky attack.

```
1  #!/bin/tcsh -f
2  set Threshold=10
3  # Step 1 Learning to produce a low-dimensional
4  cat /var/log/querylog | clgrep -v -cclients.conf | \
5  cngrep -i -v -Dnoise | grep "IN A +" | \
6  sdis 0.0 0.0 | dlevens 1 40 | tr '#' ' ' | \
7  awk '{print $7}' | sort -r | uniq -c | sort -r | \
8  awk '{printf("%s\t%s\n",$2,$1);}' | \
9  qdos Threshold >tmpfile
10 # Step 2 Detection
11 cat /var/log/querlog | clgrep -ctmpfile | \
12 grep "IN A +" >KAdet.log
13 # Step 3 Scoring
14 cat KAdet.log | wc -l >>KAdetScore.txt
15 exit 0
```

Figure 4 Suggested Kaminsky Attack Detection Algorithm and Script Code.

2.4 Estimation of restricted Damerau-Levehnshtein Distance between FQDNs as DNS Query Keywords

The Levenshtein distance, LD (X, Y), is calculated, as

$$LD[x,y] = \min(LD[x-1][y]+1, LD[x][y-1]+1, LD[x-1][y-1]+\text{cost}) \quad (3)$$

where both x and y are lengths of the strings X and Y, and the X and the Y are strings of the current fully qualified domain name (FQDN) i and the last FQDN i-1 of the DNS query keywords, respectively. For instance, if the FQDNs are X = "a001.example.com" and Y = "a002.example.com", the Levenshtein distance LD (X,Y) is calculated to be 1, since the Levenshtein distance counts the number of edit operations like "insertion," "deletion," and "substitution." Furthermore, the restricted Damerau-Levenshtein distance takes into consideration the operation "transposition" in order to suppress the overestimation.

2.5 Estimation of Euclidean Distances of Source IP addresses

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

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

where both IP_i and IP_{i-1} are the current IP address i and the last IP address i-1 of the source IP addresses, 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 $(192.0, 168.0, 1.0, 1.0)^T$. The consecutive detection of the source IP addresses is decided by thresholds d_{min}=0.0 and d_{max}=0.0, as

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

2.6 Detection Algorithm for Kaminsky Attack

We suggest the following detection algorithm of the Kaminsky attack and we show a prototype program in Figure 4:

── **Step 1** *Learning to produce a low-dimensional*── In this step, the **clgrep**, **cngrep**, and **grep** commands extract inbound A RR based DNS query request packet messages from the DNS query log file (*/var/log/querylog*) with discarding case-insensitively keywords *local* and *kumamoto-u*, the **sdis** command prints out a syslog message if the Euclidean distance between the two source IP addresses is calculated to be zero, the **dleven** command prints out the syslog message if the restricted Damerau-Levenshtein distance **LD(FQDN$_i$, FQDN$_{i-1}$)** takes a range of 1-40 (as discussed in the Section 3.2), and the **awk**, **sort**, **uniq**, and **qdos** commands (lines 7 to 9 in Figure 4) compute the frequencies of the restricted Damerau-Levenshtein distance **LD(FQDN$_i$, FQDN$_{i-1}$)** 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** and **grep** commands extract the Kaminsky attack related messages in the DNS query log file (*/var/log/querylog*), using the training data (*tmpfile*) and they generate only a Kaminsky attack related DNS query log file (*KAdet.log*).

── **Step 3** *Scoring* ──In the final step, the **wc** command calculates the score for the detection of the Kaminsky attack in the file *KAdet.log*, and it writes out the detection score into a score file (*KAdetScore.txt*) in an appending manner.

3. Results and Discussion

3.1 Entropy based Kaminsky Attack Detection Model

We demonstrate the calculated unique source IP address- and unique DNS query keyword-based entropies for the A 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, 2010, as shown in Figure 5.

In Figure 5, we can find five significant peaks and these peaks (1)-(5) correspond to (1) January 25th, (2) 26th, (3) 27th, (4) 28th, and (5) 29th, 2010, respectively, in which all the

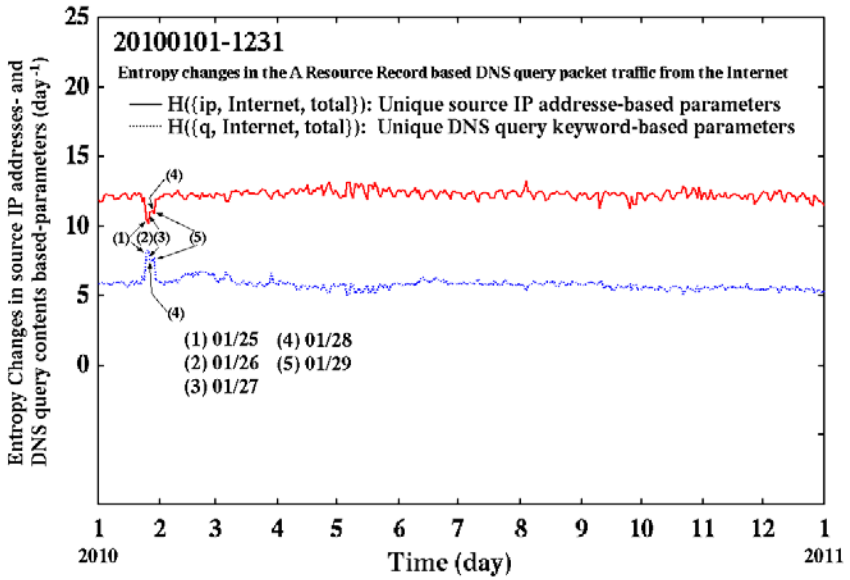

Figure 5 Entropy changes in the total A resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2010. The solid (red) and dotted (blue) lines show the unique source IP addresses and unique DNS query keywords based entropies, respectively (day^{-1} unit).

peaks show significant decrease and increase in the unique source IP address- and the unique DNS query keyword based entropies, respectively. This feature indicates that all the peaks (1)-(5) can be assigned to the Kaminsky attack.

3.2 Damerau-Levenshtein Distances based Kaminsky Attack Detection Model

We show the calculated frequency distributions of the restricted Damerau-Levenshtein distance for the five peaks (1)-(5), as shown in Figure 6. In Figure 6, the each frequency distribution has a peak and all the peaks take a range of 1-40. The detection of the Kaminsky attack is decided by thresholds dl_{min}=1 and dl_{max}=40, as

$$dl_{min} \le LD(FQDN_i, FQDN_{i-1}) \le dl_{max} \quad (4)$$

3.3 Evaluation

We illustrate the calculated score of the Kaminsky attack using the restricted Damerau-Levenshtein distance based detection model ($1 \le$ **LD(FQDN$_i$, FQDN$_{i-1}$)** ≤ 40) between the current $FQDN_i$ and the last $FQDN_{i-1}$, as the DNS query keywords in the A

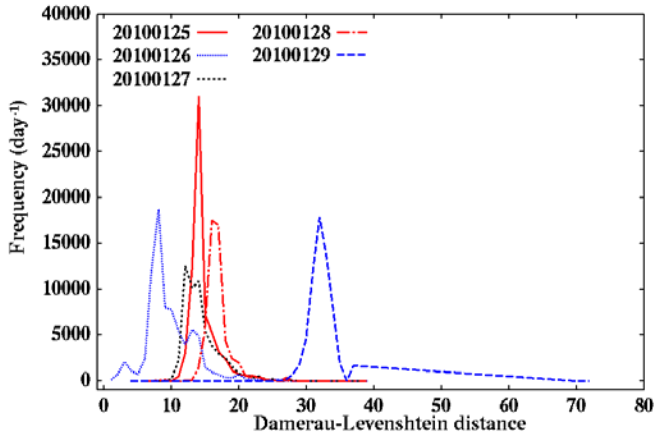

Figure 6 Frequency distributions of the restricted Damerau-Levenshtein distance at January 25th, 26th, 27th, 28th, and 29th, 2010 (day^{-1} unit).

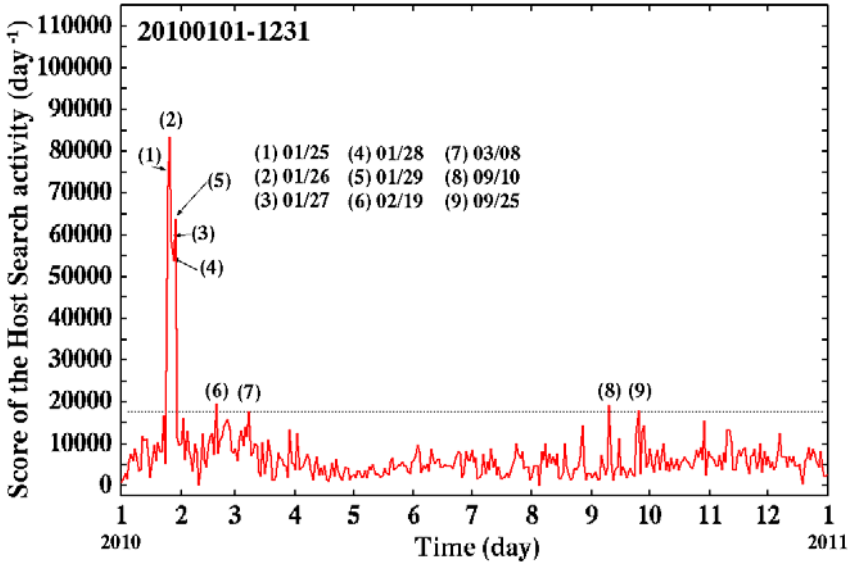

Figure 7 Changes in score of the Kaminsky attack detection in the total A resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to December 31st, 2010 (day^{-1} unit).

```
Feb 19 07:51:18 kun named[9206]: client 6*.***.***.**#59243: query: aaa.KUmamOto-u.AC.Jp IN A +
Feb 19 07:51:21 kun named[9206]: client 6*.***.***.**#58811: query: AaA.KumAmoto-u.ac.Jp IN A +
Feb 19 07:51:25 kun named[9206]: client 6*.***.***.**#45896: query: aaAA.BBbB.KuMAmOtO-u.aC.jP IN A +
Feb 19 07:51:25 kun named[9206]: client 6*.***.***.**#33403: query: aaaAAAa.Bbbb.kumaMoto-U.Ac.Jp IN A +
Feb 19 07:51:55 kun named[9206]: client 6*.***.***.**#42895: query: aaaa.bB.KuMamoTO-U.Ac.jp IN A +
Feb 19 07:51:55 kun named[9206]: client 6*.***.***.**#46946: query: KUn.kumaMOtO-U.aC.jp IN A +
Feb 19 07:51:55 kun named[9206]: client 6*.***.***.**#22191: query: KuN.KUMamOTO-U.AC.JP IN A +
Feb 19 07:51:55 kun named[9206]: client 6*.***.***.**#46824: query: aAAA.bB.KumaMotO-U.ac.jP IN A +
Feb 19 07:51:57 kun named[9206]: client 6*.***.***.**#18359: query: AAAA.bbBB.kUmamoto-U.ac.jp IN A +
Feb 19 07:51:57 kun named[9206]: client 6*.***.***.**#52946: query: aaaaaAAA.BbBb.KuMamOto-U.ac.JP IN A +
Feb 19 07:51:57 kun named[9206]: client 6*.***.***.**#53645: query: MAiL.BBbb.KUMAmoTo-u.AC.JP IN A +
```

Figure 8 Changes in the fully qualified domain names (FQDNs) as the DNS query keywords in the total A-resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server at February 19th, 2010.

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, 2010, as shown in Figure 7.

In Figure 7, we can observe nine significant peaks (1)-(9) being allocated to (1) January 25th, (2) 26th, (3) 27th, (4) 28th, (5) 29th, (6) February 19th, (7) March 8th, (8) September 10th, and (9) 25th, 2010, respectively, in which we can find the peaks (1)-(5) as the same as those in Figure 5.

Also, in Figure 7, we can observe the new peaks (6), (7), (8) and (9) corresponding to February 19th, March 8th, September 10th, and 25th, 2010, respectively, however, we can find no peaks for these days in Figure 5, showing that the restticted Damerau-Levenshtein distance based detection technology can be much false positive. Therefore, we investigated the FQDNs in the A RR based DNS query request packets in the peak (6) at February 19th, 2010, and the results are shown in Figure 8.

In Figure 8, we can observe the FQDNs consisting of the upper- and lower cases letters in an illegal matter. This feature means that the restricted Damerau-Levenshtein distance should be calculated in a case insensitive way.

We show the recalculated score of the improved Kaminsky attack using the restricted Damerau-Levenshtein distance based detection model ($1 \leq \mathbf{LD(FQDN_i, FQDN_{i-1})} \leq 40$) between the current $FQDN_i$ and the last $FQDN_{i-1}$, as the DNS query keywords in the A 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, 2010, in Figure 9.

In Figure 9, we can observe nine significant peaks (1)-(6) being assigned to (1) January 25th, (2) 26th, (3) 27th, (4) 28th, (5) 29th, and (6) September 25th, 2010, in which we can find the peaks (1)-(6) as the same as those in Figure 7.

Expectedly, the peaks at February 19th, March 8th, and September 10th, 2010, are shortening less than those in Figure 7. From this result, it is important that we take into consideration the case-insensitivity of the FQDNs in the restricted Damerau-Levenshtein based detection of the Kaminsky attack.

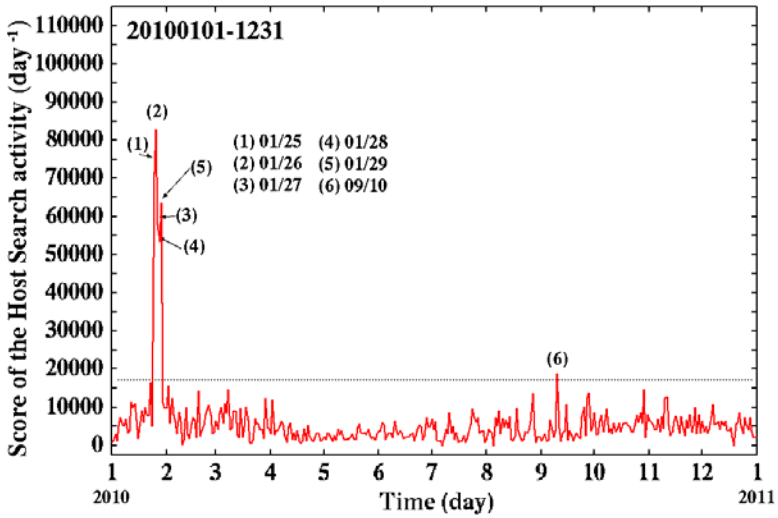

Figure 9 Changes in score of the improved Kaminsky attack detection in the total A resource records (RR) based DNS query request packet traffic from the Internet to the top domain DNS (tDNS) server through January 1st to July 31st, 2010 (day^{-1} unit).

4. Conclusions

We developed and evaluated the restricted Damerau-Levenshtein distance based detection model of the Kaminsky DNS cache poisoning (DNSCP) attack in the total inbound A resource record (RR) based DNS query request packet traffic through January 1st to December 31st, 2010. The following interesting results are found: (1) we observed five peaks for the Kaminsky attacks in the entropy changes in the A RR based DNS query request packet raffic, however, (2) we found the nine peaks in the score changes of the restricted Damerau-Levenshtein based Kaminsky attack detection model. These results show that the restricted Damerau-Levenshtein distance based Kaminsky attack detection model can be much false positive.

Therefore, we investigated the fully qualified domain names (FQDNs) as the DNS query keywords in the A RR based DNS query request packet traffic at February 19th, 2010. We obtained the case-sensitive FQDNs in the DNS query request packets, indicating that if we can take into consideration the case-insensitivity when employing the restricted Damerau-Levenshtein distance based detection model, the false positive probably decreases.

Thus, we added the case-insensitive code to the restricted Damerau-Levenshtein distance based detection engine and evaluated the detection score of the Kaminsky attack in the A RR

based DNS query request packet traffic through January 1st to December 31st, 2010, again.

Finally, we found six significant peaks of in the score changes of the newly improved restricted Damerau-Levenshtein distance based Kaminsky DNSCP attack detection model *i.e.* the three score peaks disappeared or shortened but the other one remained. These results show that the former peaks include false positive and the latter one means precise detection.

Consequently, it is very important to employ the case-insensitivity in the restricted Damerau-Levenshtein distance based Kaminsky attack detection model, since the detection rate strongly depends on the presence of the case insensitive FQDNs.

We continue further investigation and development of the Kaminsky DNS cache poisoning attack detection technology in the near future.

Acknowledgment All the studies were carried out in CMIT of Kumamoto University. We gratefully thank all the CMIT staffs and all the members of Kumamoto Universiy.

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) Kaminsky, D.: It's The End of The Cache As We Know it," 2008, http://kurser.lobner.dk/dDist/DMK_BO2K8.pdf.

6) Ludeña Romaña, D. A., Kubota, S., Sugitani, K. and Musashi, Y.: DNS-based Spam Bots Detection in a University, International Journal of Intelligent Engineering and Systems, Vol. 2, No. 3, pp. 11-18 (2009).

7) Lei, M., Musashi, Y., Ludeña Romaña, D. A., Takemori, K., Kubota, S., and Sugitani, K.: Detection of Host Search Activity in Domain Name Reverse Resolution Traffic, IPSJ Symposium Series (IOTS2009), Vol. 2009, No. 15, pp.91-94 (2009).

8) Musashi, Y., Hequet, F., Ludeña Romaña, D. A., Kubota, S., and Sugitani, K.: 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, pp.1284-1288 (2010).

9) Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady, Vol. 10, No. 8, pp.707-710 (1966).

10) Hamming, R. W.: Error detecting and error correcting codes, Bell System Technical Journal, Vol. 29, No. 2, pp.147-160 (1950).

11) Damerau, F. J.: A technique for computer detection and correction of spelling errors, Communications of the ACM, Vol. 7, No. 3, pp.171-176 (1964).

12) BIND-9.3.6-P1: http://www.isc.org/products/BIND/