原著論文

Statistical Analysis in Log Files of Electronic-Mail Server and Domain Name System Server. SPAM Mail Generates Many DNS Query Packets

Yasuo Musashi, Ryuichi Matsuba,† and Kenichi Sugitani‡

Center for Multimedia and Information Technologies, Kumamoto University, Kumamoto 860-8555 Japan,
E-mail: musashi@cc.kumamoto-u.ac.jp,
†*E-mail: matsuba@cc.kumamoto-u.ac.jp,*
‡*E-mail: sugitani@cc.kumamoto-u.ac.jp*

Abstract: The system log (syslog) files of the E-mail and the DNS cache servers in Kumamoto University were statistically investigated when receiving a lot of spam mails. The DNS query traffic between the E-mail and the DNS cache servers increases when many traces of spam and/or junk mails are found in syslog file of the E-mail server. The DNS query traffic decreases when preventing access between the E-mail server and the spam/junk transferring SMTP clients. This is because the DNS query between the DNS and E-mail servers are mainly driven by the SMTP access in the E-mail server. Therefore, we can detect abnormality of the E-mail server by monitoring the DNS query traffic from the E-mail server to the DNS server and get access-controlling list by analysis of the SMTP syslog files.

Keywords: statistical analysis, spam mail, junk mail, DNS access, SMTP access, POP3 access

1. Introduction

It is of considerable importance to keep DNS and E-mail servers in good working order in computer center like universities, governmental offices, enterprises, and so on. It is known that a DNS server provides a host domain name (A record), an IP address (PTR record), and mail exchange (MX record) to DNS clients like E-mail server (SMTP/POP3) and/or WWW browsing network applications. If the DNS server stops, all of network applications freeze or crash. From this point, we need to protect DNS server, firmly.

Intrusion detection system (IDS) is one of attractive solutions to keep security of the network servers[1-15]. There are two ways of detection of abnormality of the network servers; one is a pattern-matching with a signature file (Misuse Intrusion Detection; MID)[4,6], which employs a database of the remote attacking pattern (to detect abnormality of the network servers), and the other is directly detection of abnormality of the network servers (Anomaly Intrusion Detection; AID)[4-12]. The former frequently needs to update the signature file because of quick development of cracking technologies. On the other hand, the latter does not always need to update the signature files. In order to develop a new useful statistical AID-based IDS against future remote attack on the network servers, it is of considerable importance to get detailed profile/information for traffic of network packets like DNS query packets between a DNS server and a DNS client.

We previously reported that the number of DNS query packets, $D_{\rm q}$, are predominantly generated from an E-mail server[16],

$$D_{\rm q} = m_{\rm S}N_{\rm S} + m_{\rm P}N_{\rm P} \quad (1)$$

$$m_{\rm S} = 2 + kn(1-q) \quad (2)$$

$$m_{\rm P} = 1$$

where $N_{\rm S}$, $N_{\rm P}$, and n represent the numbers of SMTP access, POP3 access, and different domain hosts, respectively, and $m_{\rm S}$ and $m_{\rm P}$ are linear coefficients. The k value is defined as 2 or 4[17]. Here, a mail-receiving rate is $q = N_{\rm S}(r)/(N_{\rm S}(r) + N_{\rm S}(t))$, in which r and t show the received and the transfered E-mails, respectively. These results show that the DNS access from the E-mail server is mainly driven by the SMTP access.

In the present paper, we statistically investigated traffic of the DNS query access between the DNS server (**1DNS**)[18] and the E-mail server (**1MX**)[19]. Observations were performed days when receiving many SMTP accesses like a denial of service (DoS)

attack. We compare both logs of SMTP and POP3 accesses with that of DNS query access and show how the SMTP access affects the DNS query access. The present work is our research series of correlation analysis between DNS query packets and SMTP access[16].

2. Computations and Methods

2.1 Used Server Daemon Programs and Estimation of D_q, N_S N_c, N_f, and N_P

In **1DNS**, the BIND-9.2.2 program package have been employed as a DNS server daemon[20]. The DNS query packets and their contents have been recorded by the query logging option (see man named.conf), as follows:

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

In **1MX**, the program packages of Postfix-2.0.6[22] and Qualcomm qpopper-4.0.5[23] were installed as server daemons of SMTP and POP3, respectively. The logs of SMTP and POP3 accesses have been recorded in the syslog file. All of the syslog files are daily updated by the crond system.

The D_q, N_c, N_f, and N_P values are obtained, as follows: (1) The D_q value is given by the number of lines of /var/log/qlog/querylog in **1DNS** (grep and wc commands)[24]. (2) The N_c value is as the same as N_S value, which is the number of "connect from" lines of /var/log/syslog in (**1MX**) (grep and wc commands). (3) The N_f and N_P values were provided by the numbers of "from=" and "popper" lines of /var/log/syslog in **1MX**, respectively (grep and wc commands).

In sendmail-8.9.3[25], the syslog file consists mainly of both "from=" and "to=" lines. This means that the N_S value is defined as

$$N_S = N_f$$

However, the syslog of postfix-2.0.6 puts several messages such as "connect from" lines, "from=" lines, "to=" lines, and the other lines. Especially, the "connect from" line includes an IP address of a SMTP client and this line also means the beginning of SMTP access. Therefore, the N_c value is equal to the N_S value, as

$$N_S = N_c$$

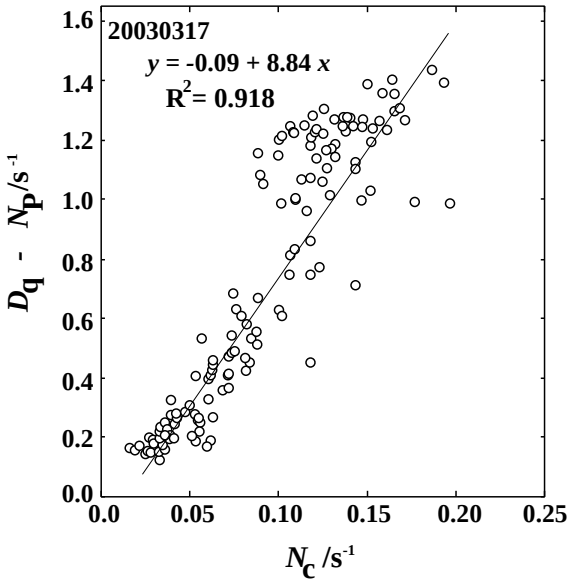

Figure 1. $D_q - N_P$ vs N_c plot (March 17th, 2003). The circle point shows a sampling data by ten minutes in the day (s^{-1} unit). Correlation coefficient (R^2) is 0.918

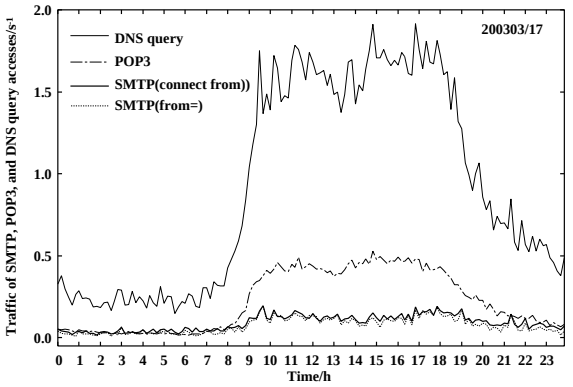

Figure 2. Traffic of the SMTP, POP3, and DNS query accesses in March 17th, 2003. The top curve shows DNS query access, the second curve shows POP3 access, and the bottom two curves are indicate SMTP accesses where the thick real and broken curves demonstrate numbers of "connect from" and "from=" lines, respectively (s^{-1} unit).

Figure 1 illustrates regression analysis between $D_q - N_P$ versus N_c, where D_q, N_P and N_c values with sampling by ten minutes of the syslog files (**1DNS** and **1MX**) are used. The data are March 17th, 2003. In Figure 1, the correlation coefficients (R^2) is 0.918. Hence, eq (1) is rewritten as, $D_q = 8.8N_c + N_P$. A linear function of SMTP access m_S (see eq (1)) is 8.8 so that one SMTP access requires DNS query access in 9 times greater extent than POP3 access in **1MX**. In Figure 2, the D_q curve changes in almost the same manner as N_c one.

Therefore, the number of "connect from" line in syslog file of E-mail server can be used as the number of SMTP access N_S and these Figures 1 and 2 can be

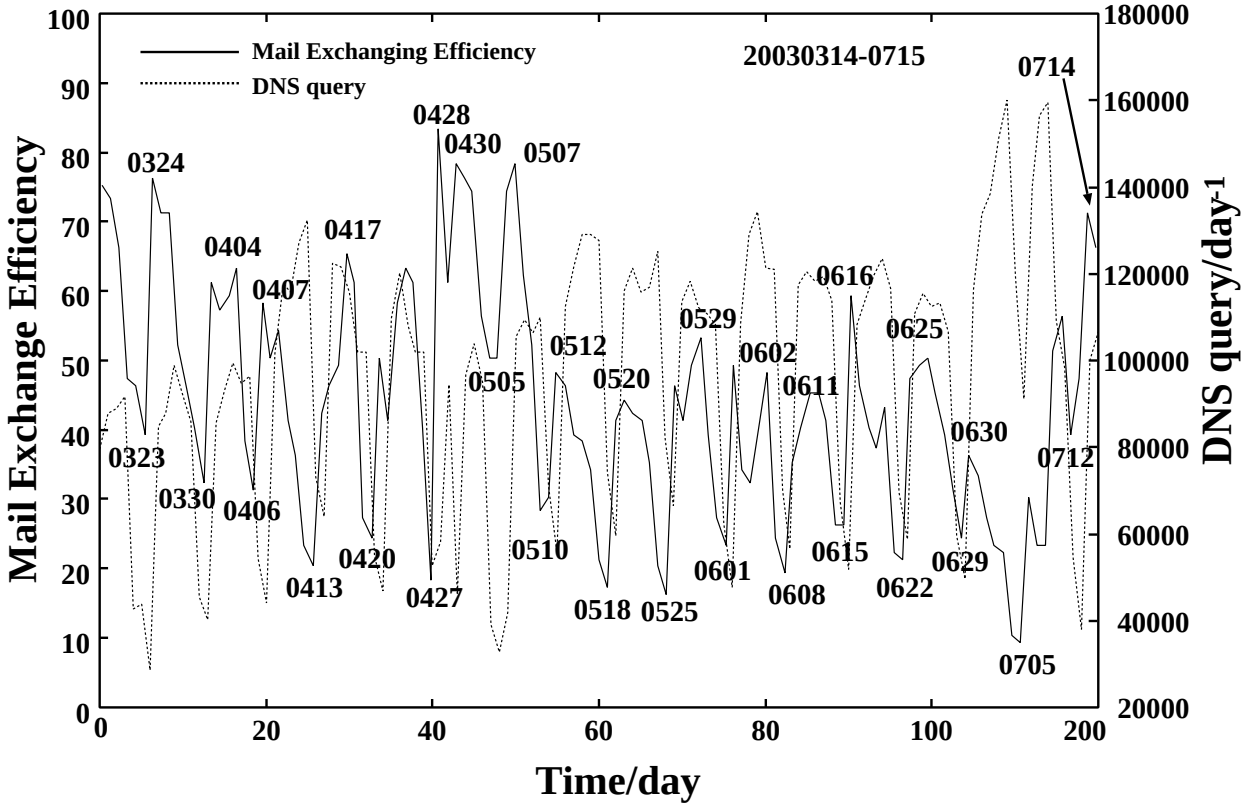

Figure 3. Mail exchange efficiency and DNS query traffic (unit; day^{-1}) between **1DNS** and **1MX** through March 17th to July 15th, 2003. Solid and dotted lines show mail exchange efficiency and DNS query traffic, respectively.

used as statistical profiles for the E-mail server (**1MX**).

2.2 Signature of SPAM

We can find a lot of lines that include "reject: RCPT" in syslog files. In the present paper, these lines can be categorized signature patterns for SPAM mail, as follows: (1) The **Relay access denied** line is printed as a message when preventing the third-party-relay. (2) The **Recipient address rejected** line means that the pattern of the destination and/or source E-mail addresses are illegal. (3) The **User unknown** line is divided into two categories; one is an simple error mail by a legal sender and the other is that mail address is generated in a random way by a SPAM mail sender.

3. Results and Discussion

3.1 SPAM/Junk Mails and DNS query access

Normally, one SMTP connection leads to do one exchange of E-mail. We can find, however, that the N_c values frequently become larger than the N_f values in Figure 2. This means that all SMTP connections are not always used for E-mail exchange. Here, we describe the mail-exchange efficiency MEE as

$$\alpha = \frac{N_f}{N_c} \tag{3}$$

In Figure 3, we plot MEE four months (from March 17th to July 15th, 2003). The MEE curve considerably decreases and increases before holidays and at the first weekday, respectively. In other words, the local maximum is every Monday and the local minimum is every Sunday. These results are easily interpreted in terms of the situation in university. This is because all the users of **1MX** work well through weekdays, and several users do on holidays. The MEE values for normal holiday and weekday are calculated to be 0.55-0.86 and 0.30-0.40, respectively (March to May, 2003). Therefore, we employ hereafter the average MEE values of 0.70 and 0.35 as statistical criterions for holiday and weekday, respectively.

As shown in Figure 3, the DNS query curve changes in almost same manner with the MEE curve, however, it frequently deviates from the normal situation. In April 13th, 2003, for instance, the MEE value decreases to 0.21 while the DNS query one increases, significantly. Moreover, this day is Sunday and the daily total number of DNS query access is observed to be 59780 (Usually, about 23000 in a holiday)[16].

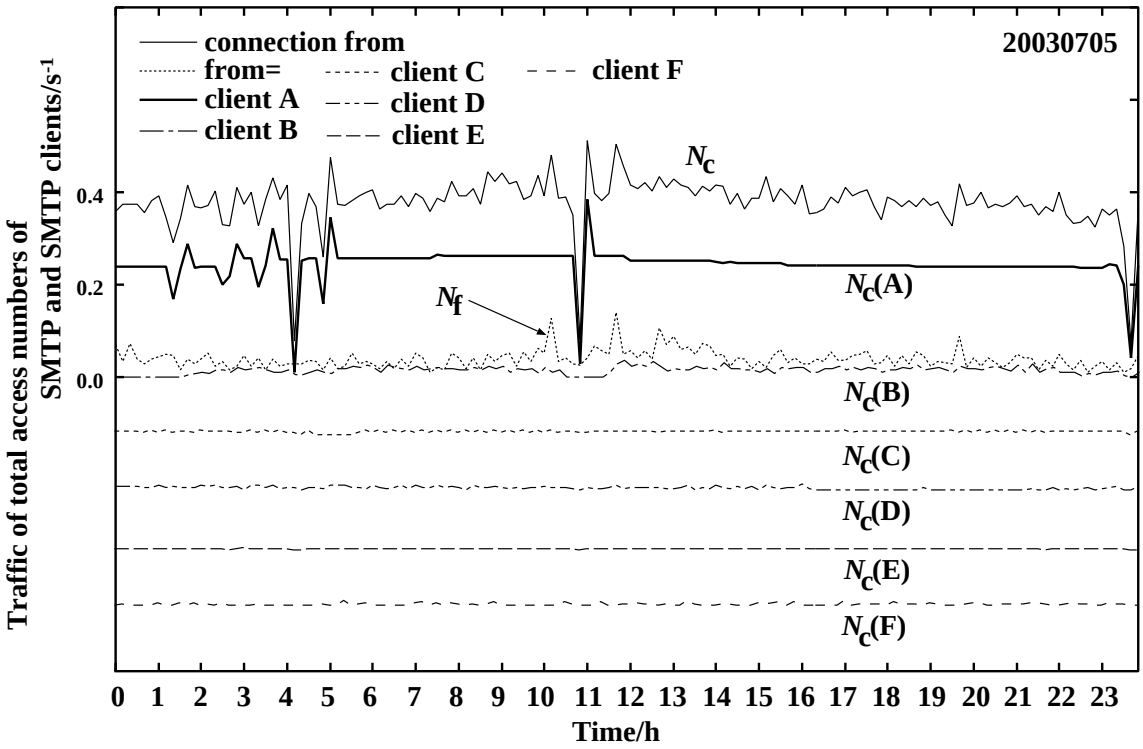

Figure 4. Traffic of total SMTP access and the top SMTP access of the SMTP clients in July 5th, 2003. Both first and third lines indicate the number of the total SMTP access ($N_{\rm c}$) and the number of "from=" line ($N_{\rm f}$) and the others are the top SMTP accesses of each SMTP clients A-F (s^{-1} unit).

This value is about 2.5 times larger than that in the normal holiday. In the syslog files at this day, the 13523 rejected RCPT lines are observed, which consist of 13505 **User unknown** lines, 16 **Relay access denied** lines, and 2 **Recipient address rejected** lines, respectively. These results are interpreted in terms of recent situations of E-mail servers: One is that the third party relay and illegally prepared E-mail address are usually omitted in the recent E-mail servers by a default configuration. The other is that several E-mail accounts are removed before this day. In the **User unknown** lines, the removed E-mail accounts are included and unknown IP addresses of the uncertain SMTP clients are taken to be almost the same as the known IP addresses.

Surprisingly, in July 5th, 2003, although the MEE value is calculated to be 0.11, the daily total number of DNS query access is given to be 119397 and this value is 5.0 times abnormally larger than the averaged value *i.e.* the E-mail server **1MX** works only 11 % normally in this day. In the syslog file of **1MX** at the day, the 32028 rejected RCPT lines are observed, which consist of 32021 **User unknown** lines and 7 **Relay access denied** lines, respectively. The **User unknown** lines consist of a specific E-mail account which takes to be 21132 lines (65 %). These E-mail and IP addresses do not change through the day so that the E-mail server **1MX** is probably attacked by SMTP client that is related to either the SPAM sender or Denial of Service (DoS) attacker.

As a result, the mail exchange efficiency (MEE) provides important information of abnormality of an E-mail server: If the MEE value and the DNS query access simultaneously decrease, the E-mail server should be in a normal situation. However, if the MEE value decreases while the DNS query access considerably increases, the E-mail server is probably in an abnormal situation that the E-mail server would be attacked by the SPAM/Junk mail senders.

3.2 Detecting Strange SMTP Clients

The number of the SMTP access ($N_{\rm c}$) is represented by the sum of the SMTP access numbers of all the SMTP clients, $N_{\rm c}(i)$, as follows

$$N_{\rm c} = \sum_i N_{\rm c}(i)$$

It is worthwhile to compare the $N_{\rm c}$ curve of **1MX** with several $N_{\rm c}(i)$ curves. Figure 4 demonstrates $N_{\rm c}$, $N_{\rm c}$(A), $N_{\rm c}$(B), $N_{\rm c}$(C), $N_{\rm c}$(D), $N_{\rm c}$(E), and $N_{\rm c}$(F), curves through the day, July 5th, 2003, in which the clients A to F are six top clients of SMTP access in the day[26].

In Figure 4, we can obtain the following clear results: (1) Although $N_{\rm c}$(A) curve takes about a value

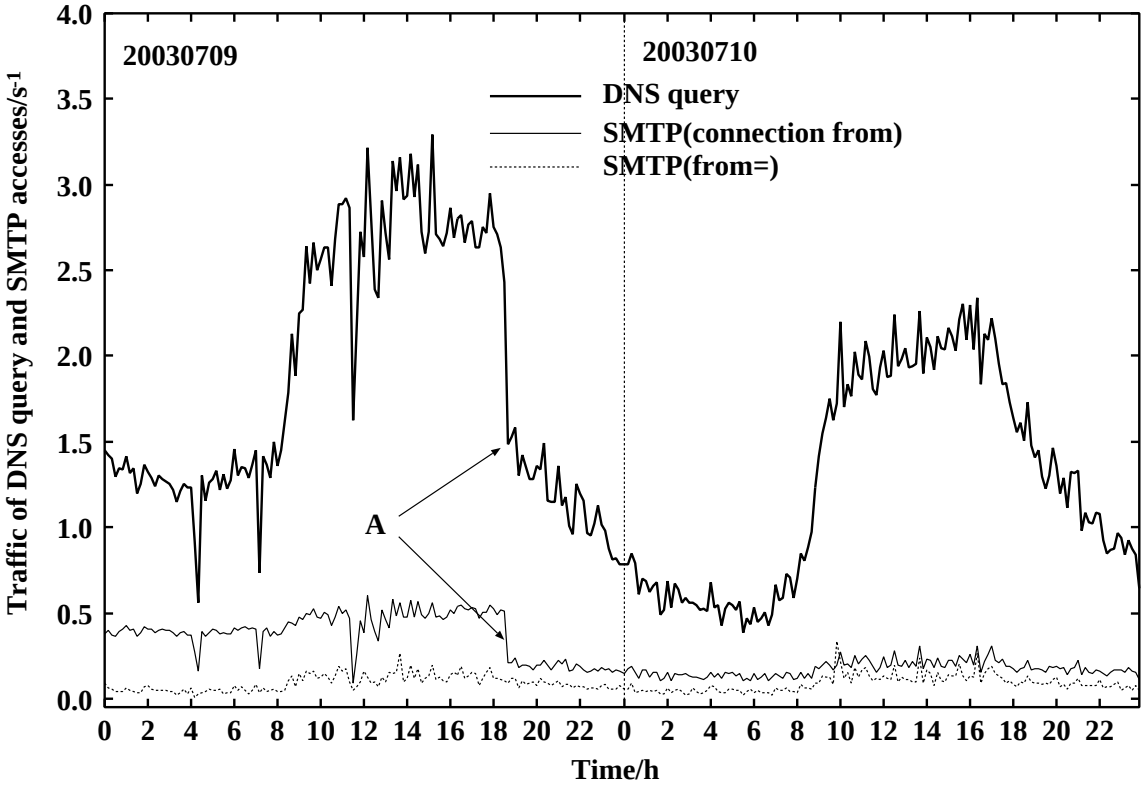

Figure 5. Traffic of the DNS query and SMTP accesses through July 9th to 10th, 2003. The top and the other curves show the DNS query access (**1DNS**) and the numbers of "connection from" and "from=" lines in the syslog file of the E-mail server **1MX**, respectively (s^{-1} unit).

of 0.24 s^{-1}, the five local maximums and the six local minimums of N_c(A) curve take place in almost the same positions as those of total N_c curve. (2) The ripple parts of total N_c curve are, however, very similar to those of the N_f curve without the peak parts in the N_c(A) curve. It is clear that the SMTP access of client A (N_c(A)) contributes significantly to the total SMTP access (N_c) in the E-mail server **1MX** at the day.

3.3 Filtering Effects and DNS Query Access

It is well-known that the access control list (ACL) is one of methods to filter the unfavorable/illegal specific access from the SMTP clients. An ACL table in the postfix-2.0.6[22] is defined in a main configuration file (/etc/postfix/main.cf) and it is a text formated file as follows;

```
smtp.clients.a        discard
smtp.clients.a        reject
```

There are two filtering options that can be available in the ACL file. The "discard" option is to allow all the access from the SMTP clients but does nothing by faking to have received the SPAM/Junk mails, silently. The "reject" option is to deny the specific access and to inform the rejected message by SMTP to SPAM/Junk mail sender (if acceptable).

Firstly, we set "discard" option into the ACL file in order to stop the mass SMTP access of client A at July, 5th, 2003. The feature of this option maybe good for filtering abnormal SMTP access. But the DNS query access from the E-mail server does not change. Next, we try "reject" option to set into the ACL file to stop the access of client A at July 9th 18:30, 2003. We plot D_q and N_c curves through July 9th-10th, 2003, as shown in Figure 5. The N_c curve considerably decreases at July 9th 18:30, 2003 (see Figure 5A). Expectedly, the D_q curve gradually decreases upon going from July 9th 18:30 to July 10th 06:00. Actually, D_q and MEE values of 1.5 s^{-1} and 0.38 at July 9th 06:00 change to values of 0.5 s^{-1} and 0.67, respectively, at July 10th 06:00. Hence, we understand that the "reject" option is effective to decrease not only the N_c value but also D_q value.

4. Concluding Remarks

We statistically investigated system log (syslog) files of the DNS server (**1DNS**) and the E-mail server (**1MX**) and we obtained several useful informations for countermeasure against the SPAM/Junk mails. BIND-9.2.2 and Postfix-2.0.6 has been employed in

1DNS and **1MX**, respectively[27].

Conclusions presented in this work are summarized as follows: (1) The number of "connection from" line (N_c) in syslog files of **1MX** program represents the number of SMTP access (N_S). The DNS query traffic (D_q) between **1DNS** and **1MX** is written as $D_q = m_S N_S + N_P$, in which N_P and m_S are the line number of POP3 access in the E-mail server and a linear coefficient of SMTP access, respectively. A linear coefficient m_S is calculated to be about 9 in the normal day (for **1MX**). (2) Usually, the number of "from=" line (N_f) in syslog file takes almost the same and/or a little bit smaller than the N_c one. The mail exchange efficiency (MEE) as $\alpha = N_f/N_c$, is nearly equal to 1. The MEE value considerably decreases when receiving SPAM mails/SMTP-DoS attacks, severely. The MEE value has an interesting possibility that becomes one of the indicators to detect abnormality in the E-mail server. (3) The "reject" option can be useful and effective to stop receiving SPAM mails with the mass SMTP access since both D_q and N_c values decrease and the MEE value increases.

From these results, it is reasonably concluded that we can detect abnormality of the E-mail server by monitoring DNS query access from the E-mail server to the DNS server and get several important informations to update an access-control-list by analysis of the SMTP syslog files. Note that the certainty and efficiency of detecting SPAM mail attacks by statistical analysis decreases to 50 % when only observing a signature **RCPT:User unknown** in syslog files of the E-mail server. This is because the signature includes two possibilities that one is an simple error of E-mail address and the other is a SPAM mail. From this point, we continue further investigation in order to get information of developing an automated system for preparing access-control-list against either SPAM mails or the mass SMTP access (SMTP-DoS attack).

Acknowledgement. All the calculations and investigations were carried out in Center for Multimedia and Information Technologies, Kumamoto University. We specially thank to technical officers, K. Tsuji, M. Shimamoto and T. Kida, and K. Makino who is a system engineer of MQS (Kumamoto) for daily supports and constructive cooperations.

References and Notes

[1] Northcutt, S. and Novak, J., *Network Intrusion Detection*, 2nd ed; New Riders Publishing: Indianapolis (2001).

[2] Sato, I., Okazaki, Y., and Goto, S.: An Improved Intrusion Detecting Method Based on Process Profiling, *IPSJ Journal*, Vol. 43, No.11, pp.3316-3326 (2002).

[3] Jones, D.: Building an E-mail Virus Detection System for Your Network, *LINUX Journal*, No.92, pp.56-65 (2001).

[4] Denning, D. E.: An Intrusion-detection model, *IEEE Trans. Soft. Eng.*, Vol. SE-13, No.2, pp.222-232 (1987).

[5] Cisco Systems: The Science of Intrusion Detection System Attack Identification, http://www.cisco.com/warp/public/cc/pd/sqsw-/sqidsz/prodlit/idssa_wp.htm, 2002.

[6] Laing, B.: How To Guide-Implementing a Network Based Intrusion Detection System, http://www.snort.org/docs/iss-placement.pdf, ISS, 2000.

[7] Mukherjee, B., Todd, L., and Heberlein, K. N.: Network Intrusion Detection, *IEEE Network*, Vol. 8, No.3, pp.26-41 (1994).

[8] Barbará, D., Wu, S., and Jajodia, S.: Experience with EMERALD to DATE", Proceedings 1st USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, California, April 1999, pp.73-80, http://www.csl.sri.com/neumann/det99.html

[9] Neumann, P. and Porras, P.: Detecting Novel Network Intrusions using Bayes Estimators", First SIAM International Conference on Data Mining, 2001, http://www.siam.org/meetings/sdm01/pdf/sdm-01_29.pdf

[10] Warrender, C., Forrest, S., and Pearlmutter, B.: Detecting Intrusions Using System Calls: Alternative Data Models, *Proc. IEEE Symposium on Security and Privacy*, No.1, pp.133-145 (1999).

[11] Hofmeyr, S. A., Somayaji, A., and Forrest, S.: Intrusion Detection Using Sequences of System Calls, *Computer Security*, Vol. 6, No.1, pp.151-180 (1998).

[12] Ptacek, T. H. and Newsham, T. N.: Insertion, Evasion, and Denial os Service: Eluding Network Detection, January, 1998, http://www.robertgraham.com/mirror/Ptacek-Newsham-Evasion-98.html

[13] Anderson, D., Lunt, T. F., Javitz, H., Tamaru, A., and Valdes, A.: Detecting unusual program behavior using statistical component of the Next-generation Intrusion Detection Expert System (NIDES), *Computer Science Laboratory SRI-CSL-95-06*,1995.

[14] Symantec: ManHunt, http://enterprisesecurity.symantec.com/products/products.cfm?ProductID=156&EID=0

[15] Yamamori, K.: An Improvement of Network Security Using an Intrusion Detection Software, *Journal for Academic Computing and Networking*, No.4, pp.3-13 (2000).

[16] (a) Musashi, Y., Matsuba, R., and Sugitani, K.: Statistical Analysis in Logs of DNS Traffic and E-mail Server, *IPSJ SIG Notes, Computer Security 20th*, Vol. 2003, No.18, pp.185-189 (2003). (b) Musashi, Y., Sugitani, K., and Matsuba, R.: Traffic Analysis on Mass Mailing Worm and DNS/SMTP, *IPSJ SIG Notes, Computer Security 19th*, Vol. 2002, No.122, pp.19-24 (2002). (c) Musashi, Y., Matsuba, R., and Sugitani, K.: Traffic Analysis on a Domain Name System Server. SMTP Access Generates Many Name-Resolving Packets to a Greater Extent than Does POP3 Access, *Journal for Academic Computing and Networking*, No.6, pp.21-28 (2002).

[17] The k values are defined as 4 when using an old E-mail server program package like sendmail-8.9.3 and 2 when using a new E-mail server one like postfix-2.0.6.

[18] **1DNS** is the DNS cache server for **1MX** [19] which is our mail server of the generic domain name of the Kumamoto University (kumamoto-u), which is run by our center. The OS is Linux OS (kernel-2.4.21), and the AMD Athlon 1.1 GHz.

[19] **1MX** is our mail server of the generic domain name of the Kumamoto University (kumamoto-u). The OS is Solaris 2.6 (Ultra-SPARC 300MHz, Sun Microsystems Inc.).

[20] http://www.isc.org/products/BIND/

[21] Bauer, M.: syslog Configuration, *LINUX Journal*, No.92, pp.32-39 (2001).

[22] http://www.postfix.org/

[23] http://www.eudora.com/qpopper/

[24] The query log facility (local1) has been assigned into /etc/syslog.conf, as follows:

```
*.info;local1.none /var/log/messages
local1.*           /var/log/qlog/querylog
```

[25] http://www.sendmail.org/

[26] A is the clients A of the top SMTP client, B is the client of the secondary top SMTP client, C is the client of the third top SMTP client, D is the client of the fourth top SMTP client, E is the client of the fifth top SMTP client, and F is the client of the sixth top SMTP client, for **1MX** at July 5th, 2003.

[27] The other DNS and E-mail server programs are available in almost the same manner as the BIND-9.2.2 and Postfix-2.0.6 programs. For instance, in the case of tinydns (http://cr.yp.to/djbdns.html) and sendmail (http://www.sendmail.org/), iplog-2.2.3 program (http://ojnk.sourceforge.net/) can be available to get the same information as D_{q} and N_{c} values (The N_{f} value can be obtained from the syslog file of sendmail).