Statistical Analysis in Logs of DNS Traffic and E-mail Server

Yasuo Musashi,[†] Ryuichi Matsuba,[†] and Kenichi Sugitani[†]

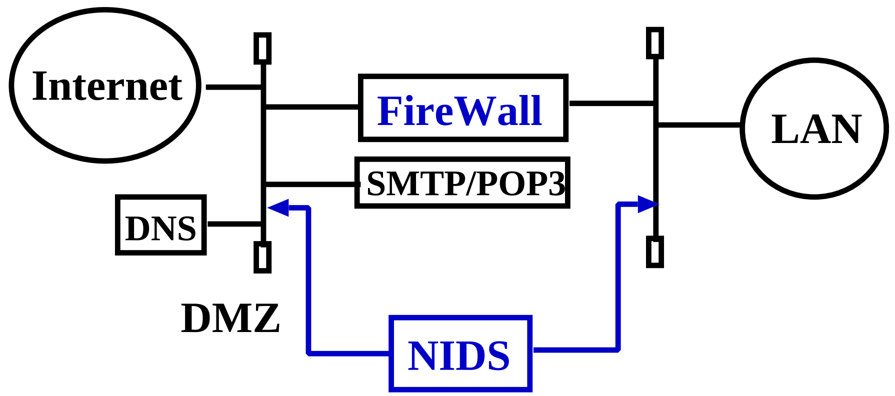
[†] *Center for Multimedia and Information Technologies, Kumamoto University,*
Kumamoto 860-8555 Japan, E-mail: musashi@cc.kumamoto-u.ac.jp

Network- and Host-based Intrusion Detection Systems

(1) Network-based Intrusion Detection System

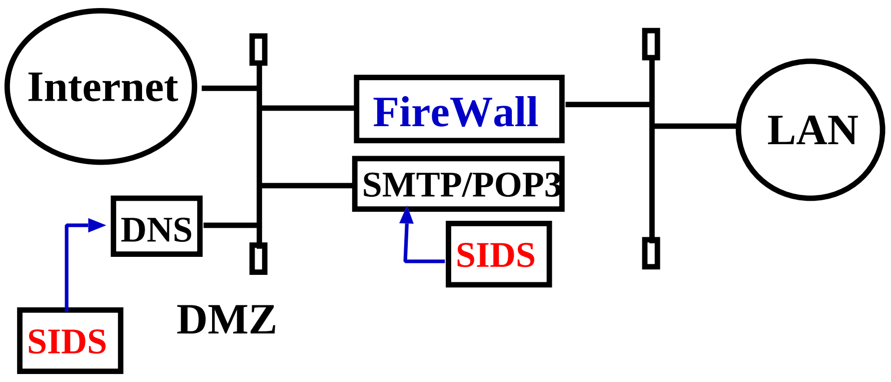
(2) Host-based Intrusion Detection System

(1) Network-based Intrusion Detection System



Detection by monitoring packets

(2) Host-based Intrusion Detection System



Detection by watching Logs and Files Falsfication

Misuse and Anomaly Intrusion Detection Models

(1) **Misuse Intrusion Detection Model.**
Detection by pattern-matching of remote attacks with a signature database, which needs a lot of resources.

(2) **Anomaly Intrusion Detection Model.**
Detection by monitoring anomaly in the network protocols, such as HTTP, SMTP, POP3, FTP, DNS, and in syslogs, which can detect not only with a small amount of resources but also without a signature database.

Anomaly intrusion detection system should be quickly to develop as a new IDS in the next generation.

What kinds of protocols should we select?

Sato *et al.* recently suggested the intrusion detecting method based on observing systemcalls of important daemons in the network server (see *IPSJ Journal* Vol.43 pp.3316(2002)).

Intrusion Detection Using DNS Query Access

The DNS service is the most important network services on the Internet.

SMTP/POP3(Mail),FTP,HTTP,... ⇒ gethostbyaddr(),gethostbyname(),...

We need to protect the DNS server, firmly.

In our previous works, traffic between DNS and E-mail servers is represented as follows:

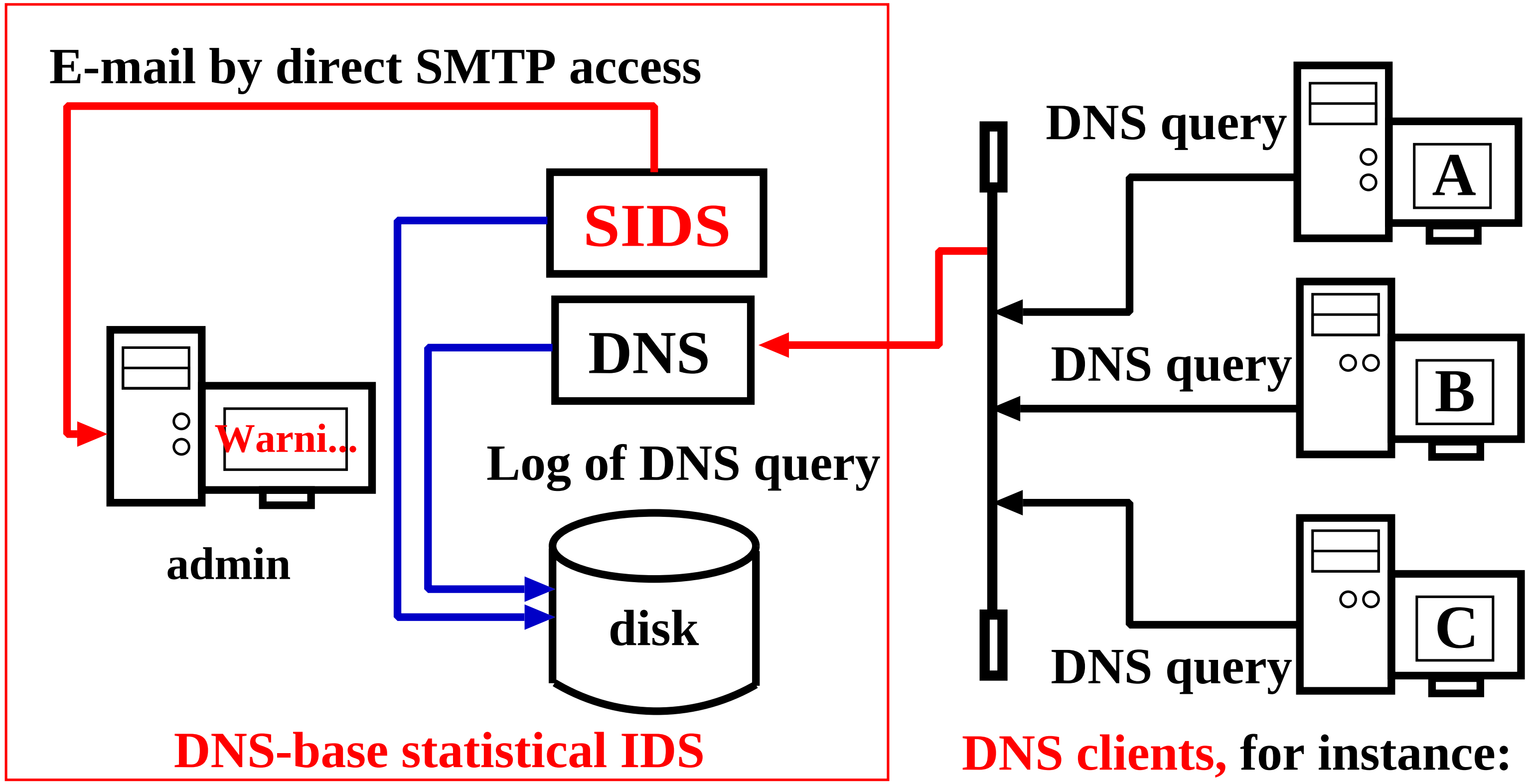
$$D_{\mathrm{q}} = m_{\mathrm{S}} N_{\mathrm{S}} + m_{\mathrm{P}} N_{\mathrm{P}} \tag{1}$$

$$m_{\mathrm{S}} = 2 + 4n(1-q) \tag{2}$$

$$m_{\mathrm{P}} = 1 \tag{3}$$

D_{q} = **the DNS traffic between DNS and E-mail servers,**
N_{S} **and** N_{P} = **the numbers of SMTP and POP3 accesses,**
m_{S} **and** m_{P} = **linear coefficients.**
Musashi *et al.* *IPSJ SIG Notes*, *CSEC19-4*, pp19-24(2002);
***J. Academic Comput. Networking*, No. 6, pp.21-28(2002).**

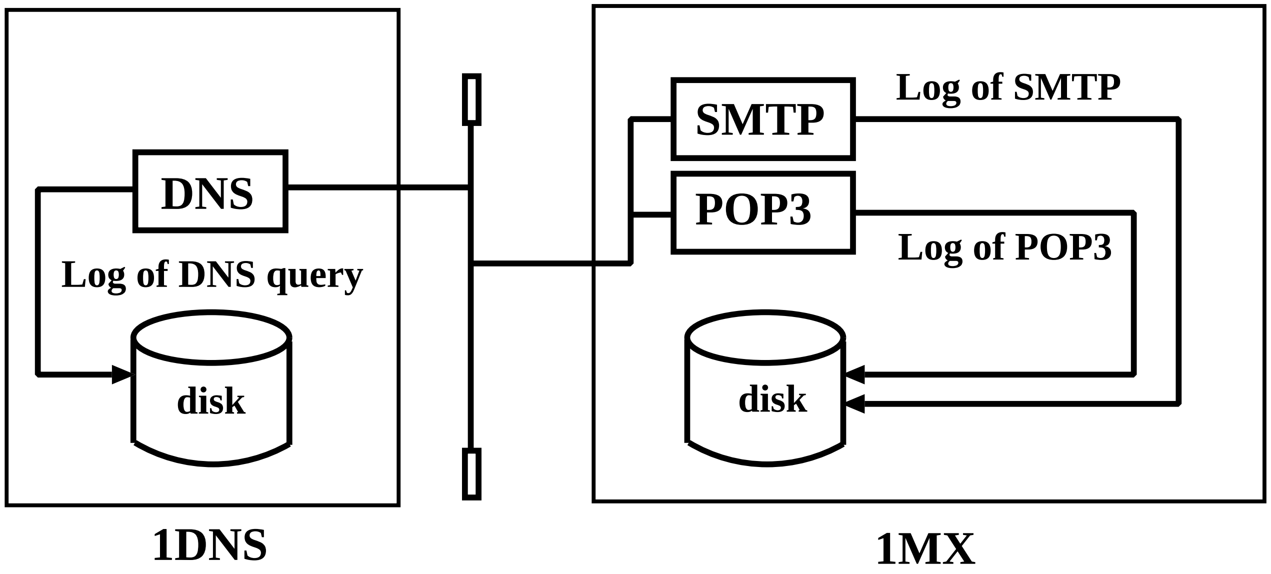
Statistical Intrusion Detection by DNS query Access



DNS clients, **for instance:**
(A) E-mail server (indirectly)
(B) PC-based fire wall (directly)
(C) Trojan horse-infected PC term.
(D) Mass Mailing Worm-infected...

DNS name resolution:
(1) canonical resolution
(2) reverse resolution
(3) mail exchange (MX) resolution

This Work (1)

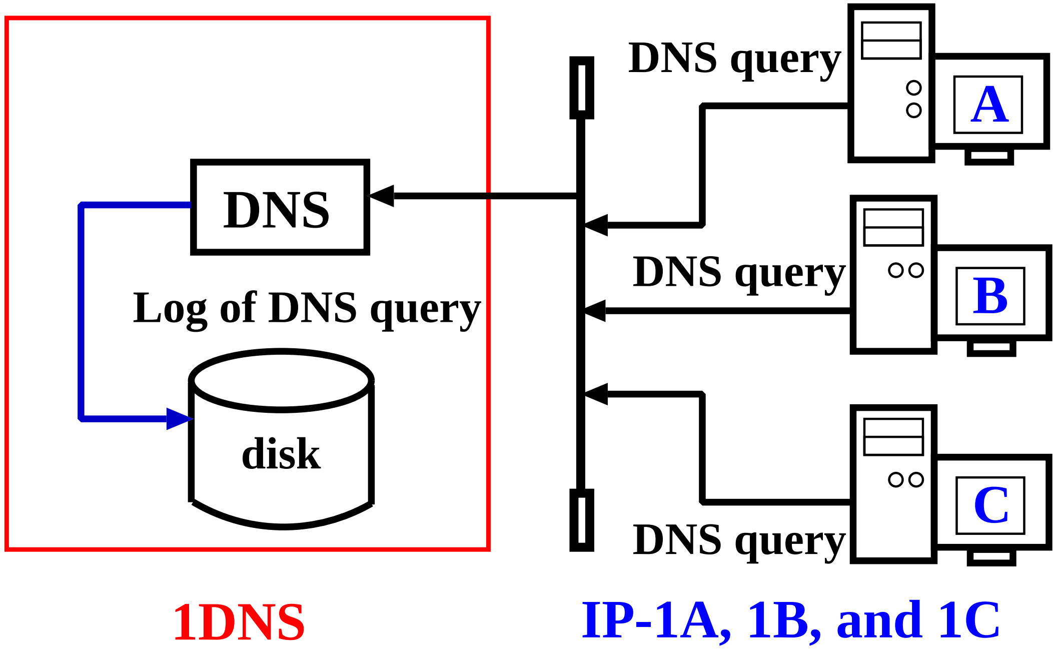


(1) Statistical investigation on DNS query traffic between the DNS server (1DNS) and the E-mail server (1MX) when detecting Frethem. K.

(2) Comparing both logs of SMTP and POP3 daemons to show how DNS traffic are influenced by Frethem. K.

(3) Showing anomaly detection methods of the DNS clients that are likely to be related with the network incidents.

This Work (2)



(1) Statistical analysis on traffic of the DNS query (D_{q}) packets between the DNS server (1DNS) and the E-mail server (1A),

(2) Statistical analysis on D_{q} traffic between the DNS server (1DNS) and the hijacked PC-based fire wall system (1B).

(3) Statistical analysis on D_{q} traffic between the DNS server (1DNS) and the trojan horse virus (THV)-infected PC terminal (1C).

Equations

$$\boldsymbol{D}_{\mathrm{q}} = \boldsymbol{R}_{\mathrm{S}} + \boldsymbol{R}_{\mathrm{P}} + \boldsymbol{R}_{\mathrm{F}} + \cdots \tag{4}$$

$$\boldsymbol{R}_i = \boldsymbol{m}_i \boldsymbol{N}_i \tag{5}$$

$\boldsymbol{D}_{\mathrm{q}}$ = the DNS query traffic between 1DNS and 1MX. $\boldsymbol{i}$ = a network application protocol, such as SMTP, POP3, FTP, ..., $\boldsymbol{R}_i$ = the network application protocol-based DNS query traffic, $\boldsymbol{N}_i$ = the traffic of $\boldsymbol{i}$, $\boldsymbol{m}_i$ = a linear coefficient for $\boldsymbol{i}$, and $\boldsymbol{R}_{\mathrm{S}} + \boldsymbol{R}_{\mathrm{P}} \gg \boldsymbol{R}_{\mathrm{F}} + \cdots$ (1MX)

$$\begin{aligned}
\boldsymbol{D}_{\mathrm{q}} &= \boldsymbol{m}_{\mathrm{S}} \boldsymbol{N}_{\mathrm{S}} + \boldsymbol{m}_{\mathrm{P}} \boldsymbol{N}_{\mathrm{P}} \quad (6) \\
\boldsymbol{m}_{\mathrm{S}} &= 2 + 4\boldsymbol{n}(1 - \boldsymbol{q}) \quad (7) \\
\boldsymbol{m}_{\mathrm{P}} &= 1 \quad (8)
\end{aligned}$$

Used Server Daemon Programs

- 1DNS: The DNS server and the DNS packet recorder.
  BIND-9.2.1 and iplog-1.2
- 1MX:The SMTP and POP3 servers.
  ISC sendmail-8.9.3 and Qualcomm qpopper-4.0

Estimation of Traffic

(1) $\boldsymbol{D}_{\mathrm{q}}$:

```
% grep domain /var/log/messages.1 | wc
```

(2) $\boldsymbol{N}_{\mathrm{from}}(=\boldsymbol{N}_{\mathrm{S}}),\ \ \boldsymbol{N}_{\mathrm{to}}(=\boldsymbol{N}_{\mathrm{SS}}+\boldsymbol{N}_{\mathrm{SD}})$:

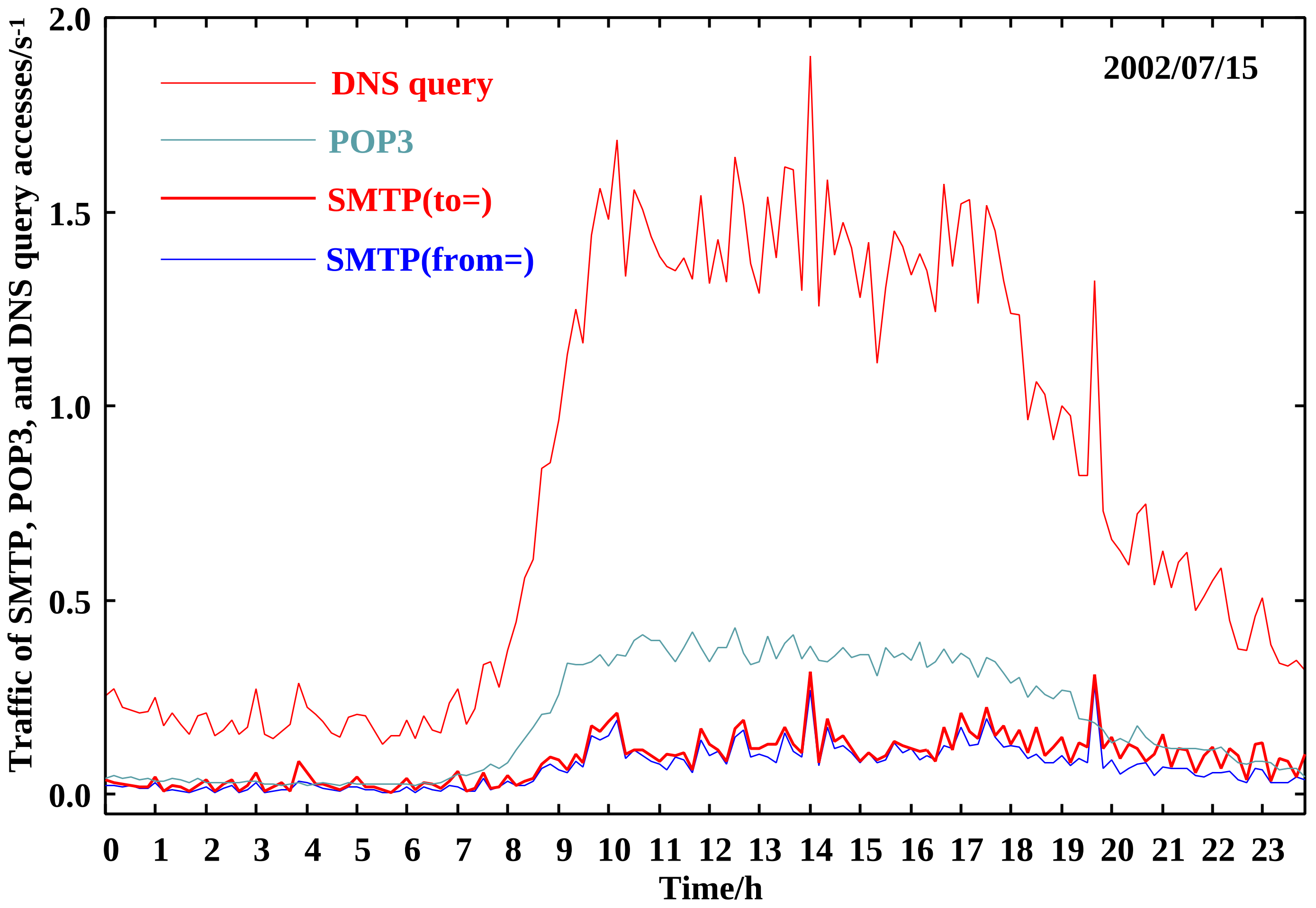
```
% grep "from=" /var/log/syslog.0 | wc
% grep "to=" /var/log/syslog.0 | wc
% grep "to=" /var/log/syslog.0 | grep "stat=Sent" | wc
% grep "to=" /var/log/syslog.0 | grep "stat=Deferred" | wc
```

(3) $\boldsymbol{N}_{\mathrm{P}}$:

```
% grep "popper\[" syslog.0 | wc
```

$$\boldsymbol{N}_{\mathrm{to}}=\boldsymbol{N}_{\mathrm{SS}}+\boldsymbol{N}_{\mathrm{SD}}\geq\boldsymbol{N}_{\mathrm{from}} \tag{9}$$

Traffic SMTP, POP3, and DNS query in 2002/07/15

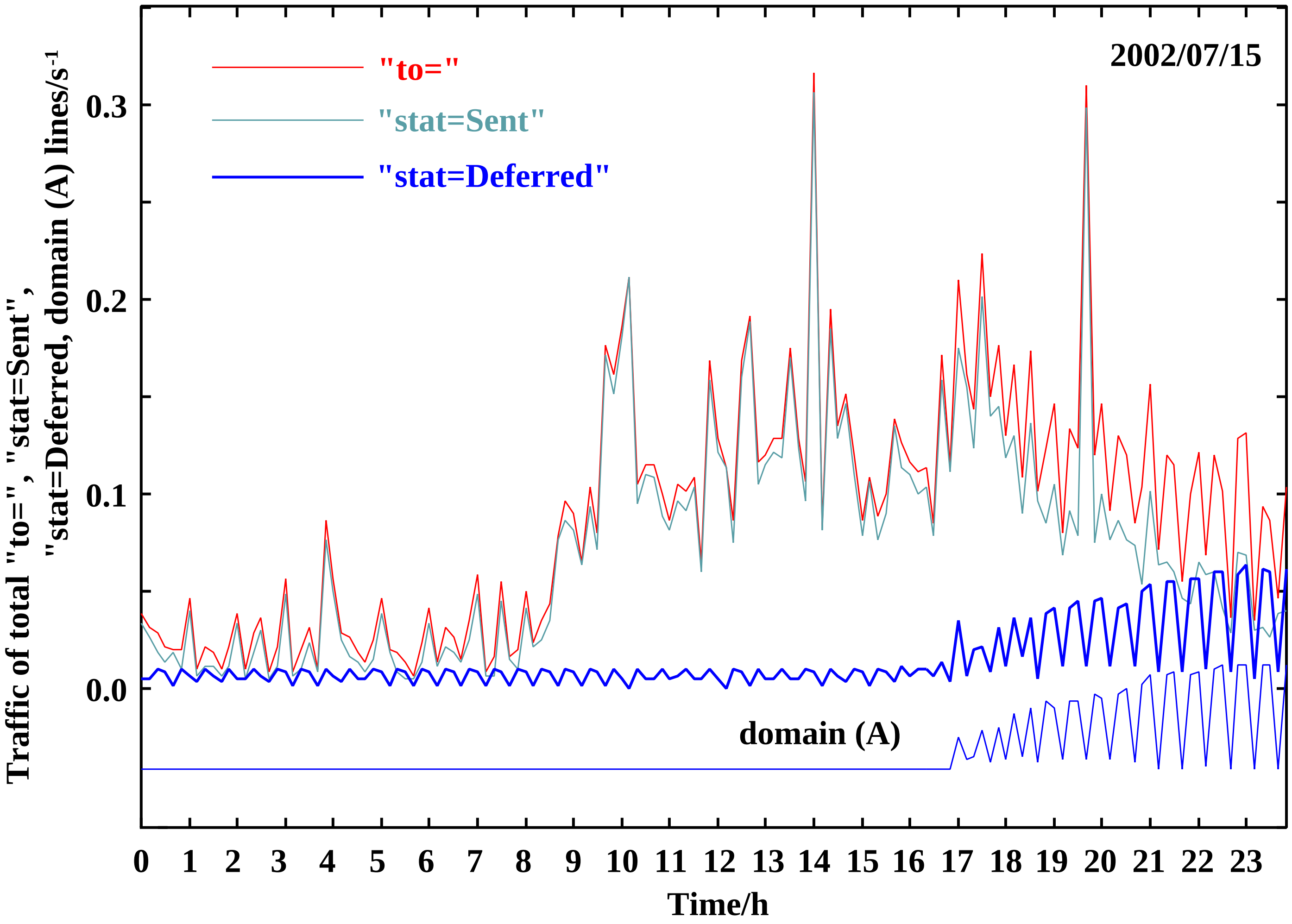


(1) Normally, $N_{\mathrm{to}} \sim N_{\mathrm{from}}$ or $N_{\mathrm{to}} > N_{\mathrm{from}}$.

(2) After 18:00, $D_{\mathrm{q}} \propto N_{\mathrm{to}}$?

(3) Consequently, $N_{\mathrm{to}} \geq N_{\mathrm{from}}$.

N_{to}, N_{SS}, and N_{SD} curves in 2002/07/15



(1) Normally, $N_{\mathrm{SS}} \gg N_{\mathrm{SD}} \rightarrow N_{\mathrm{to}} \sim N_{\mathrm{SS}} \sim N_{\mathrm{from}}$.

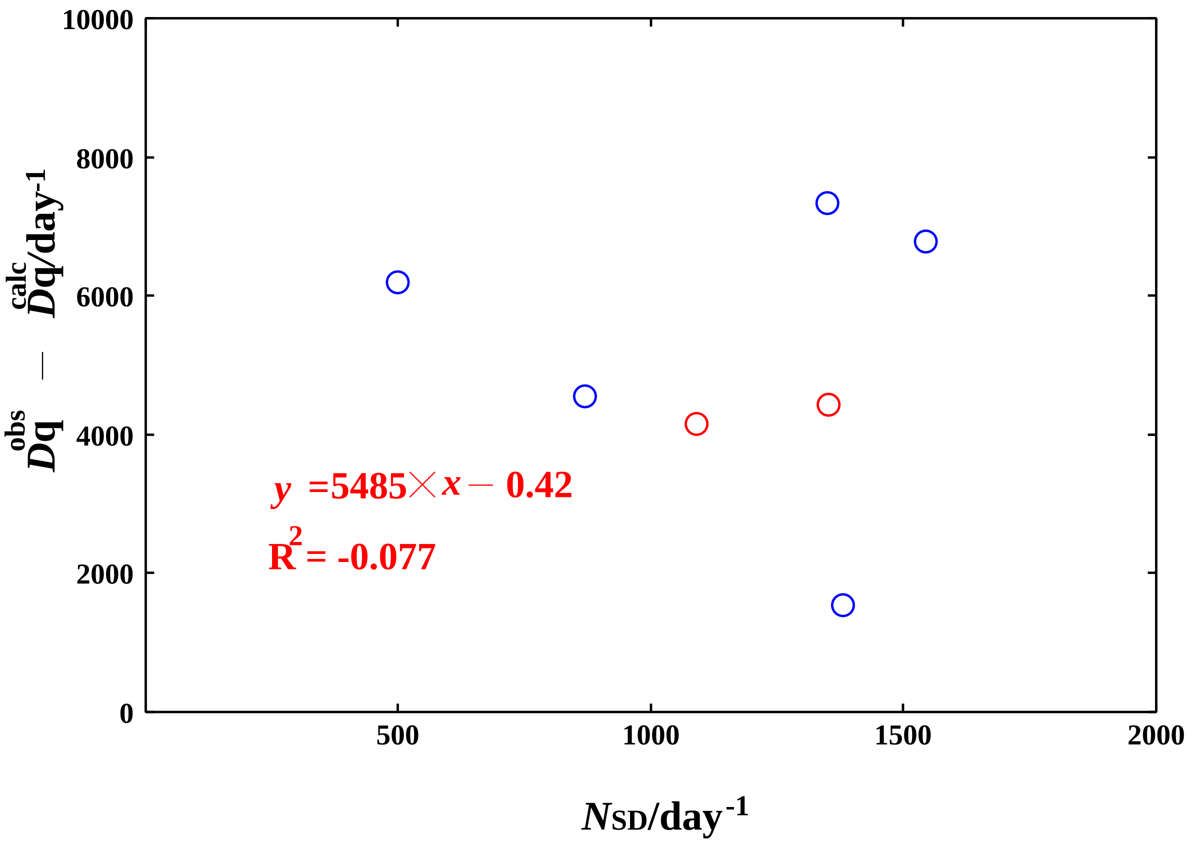
(2) After 18:00, $N_{\mathrm{to}} \sim N_{\mathrm{SS}} + N_{\mathrm{SD}} > N_{\mathrm{from}} \rightarrow N_{\mathrm{to}} \geq N_{\mathrm{from}}$

(3) The N_{to} and N_{SD} curves change in a mostly same manner.

Why is the N_{to} curve similar to the D_{q} one (1)

$$\begin{aligned} \boldsymbol{R}_{\mathbf{SD}} &= \boldsymbol{m}_{\mathbf{SD}}\boldsymbol{N}_{\mathbf{SD}} \qquad (10) \\ \boldsymbol{D}_{\mathbf{q}}^{\mathbf{calc}} &= \mathbf{8.6}\boldsymbol{N}_{\mathbf{S}} + \boldsymbol{N}_{\mathbf{P}} \qquad (11) \\ \boldsymbol{D}_{\mathbf{q}}^{\mathbf{obs}} - \boldsymbol{D}_{\mathbf{q}}^{\mathbf{calc}} &= \boldsymbol{m}_{\mathbf{SD}}\boldsymbol{N}_{\mathbf{SD}} \qquad (12) \end{aligned}$$

Why is the N_{to} curve similar to the D_{q} one (2)



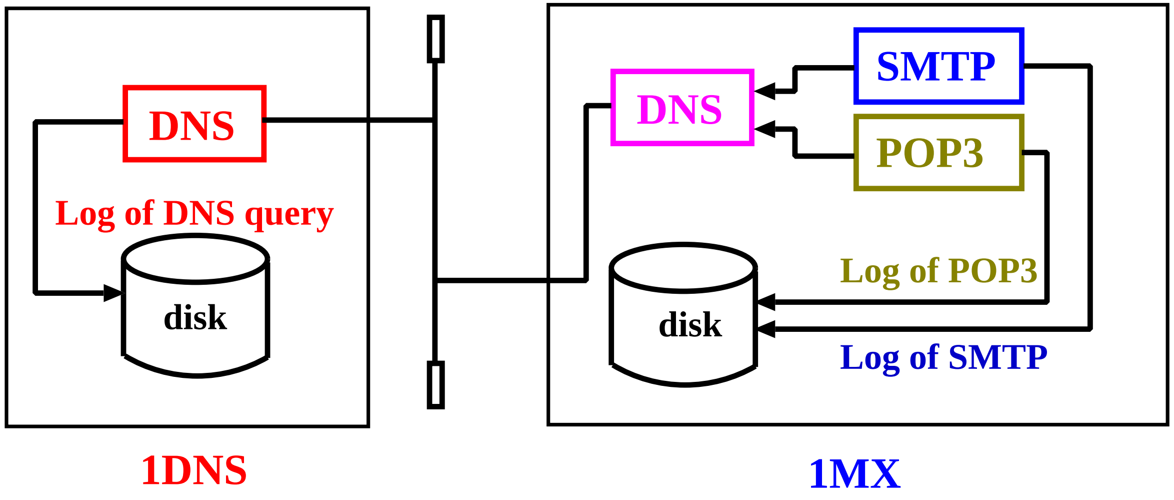
No correlation was found in the D_{q} and N_{SD}.

(1) E-mail users would repeat to send the deferred E-maill.

(2) The SMTP relay may retry to send the deferred E-mail at stated periods as the N_{SD} curve gradually fluctuates.

Cache Effects on DNS traffic from E-mail servers

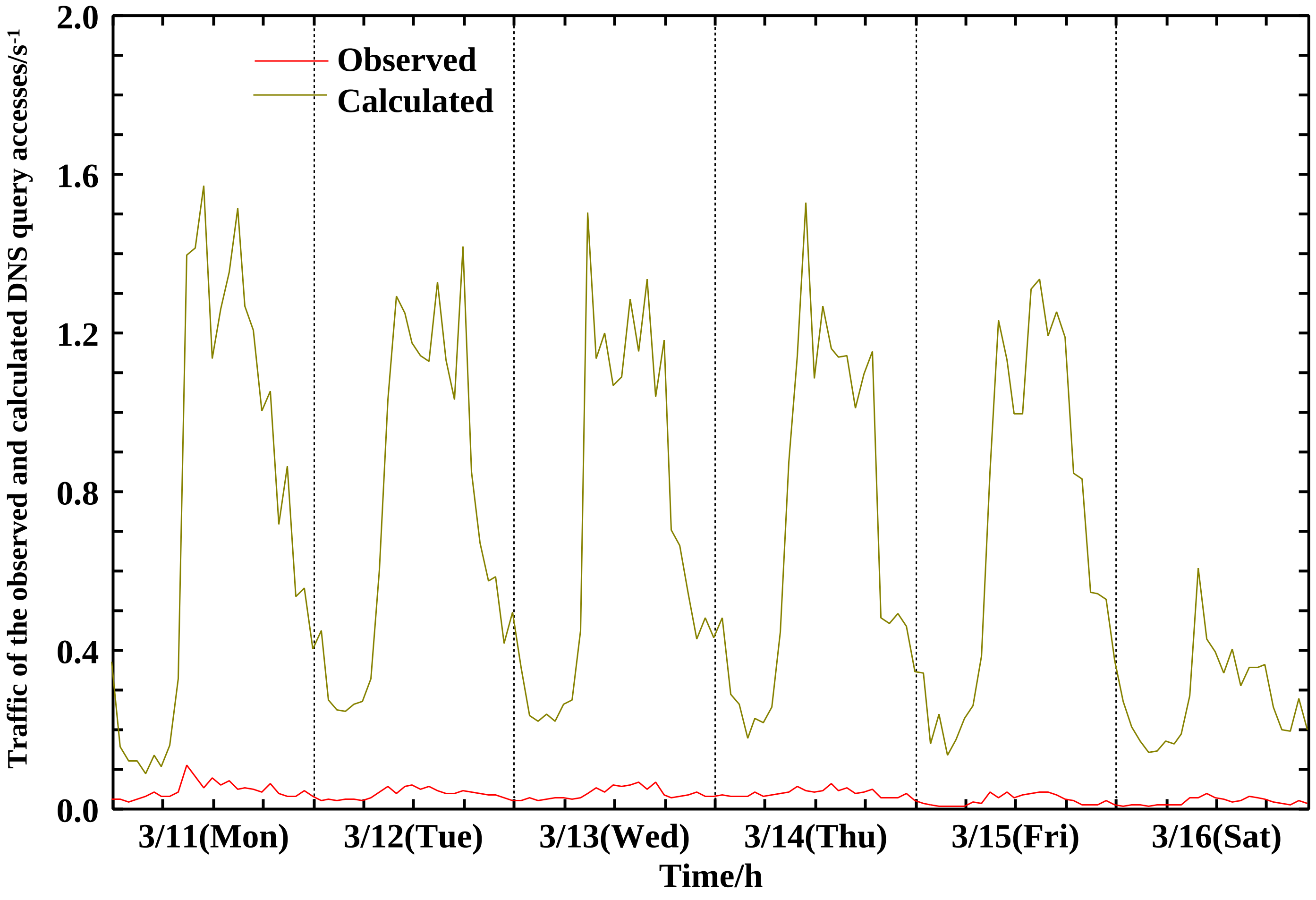
We present the DNS cache effects of the DNS query access between 1DNS and 1MX with the equation ($D_{\mathrm{q}} = 8.6N_{\mathrm{SMTP}} + N_{\mathrm{POP3}}$).



Used server daemon programs

- 1DNS: The DNS server and the DNS packet recorder.
  BIND-9.1.3 and iplog-1.2
- 1MX:The SMTP and POP3 servers.
  ISC sendmail-8.9.3 and Qualcomm qpopper-4.0

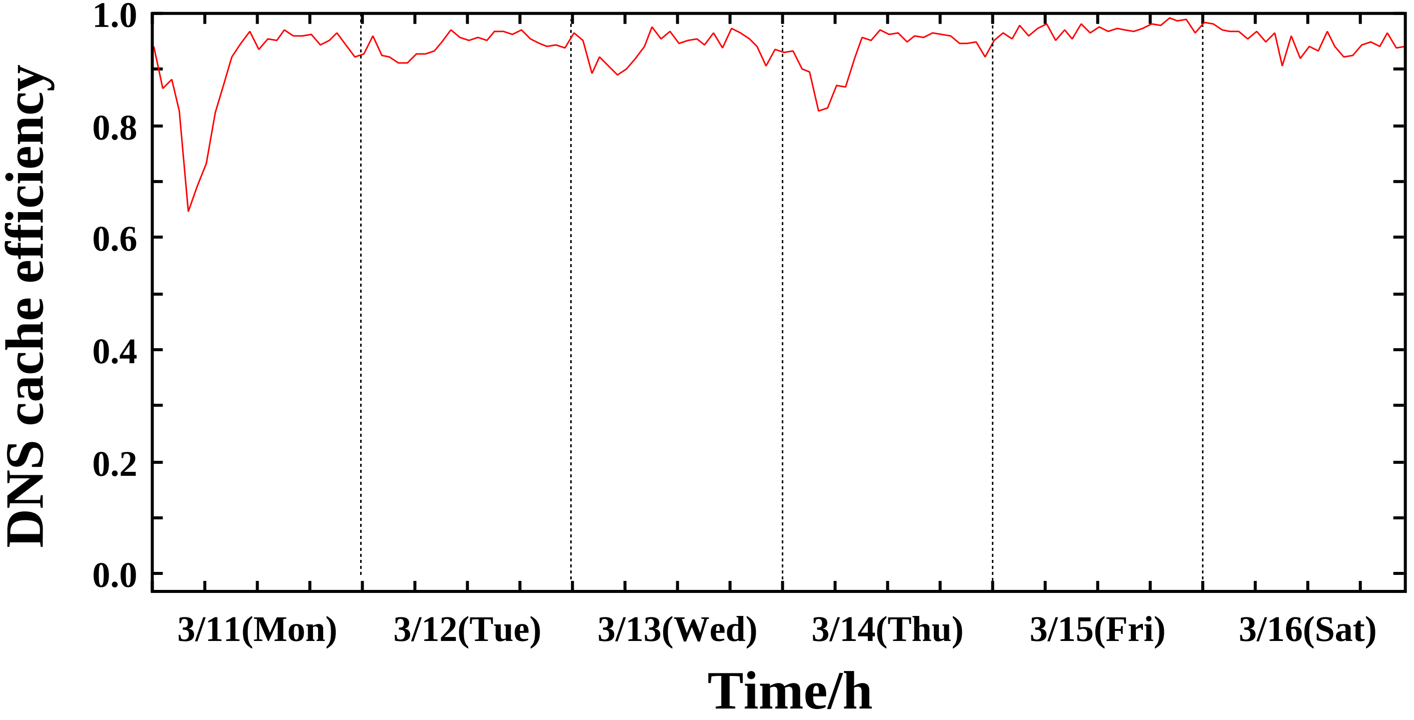
Observed and calculated DNS traffics in 20020311-0316



The observed traffic is considerably much smaller than the calculated one.

Estimated Cache Efficiency of DNS traffic

$$\mathrm{DCE} = 1 - \frac{D_{\mathrm{q}}^{\mathrm{obs}}}{D_{\mathrm{q}}^{\mathrm{calc}}} \tag{13}$$



The DNS cache for SMTP/POP3 services is considerably effective.

Total DNS query and IP-terminal DNS client accesses

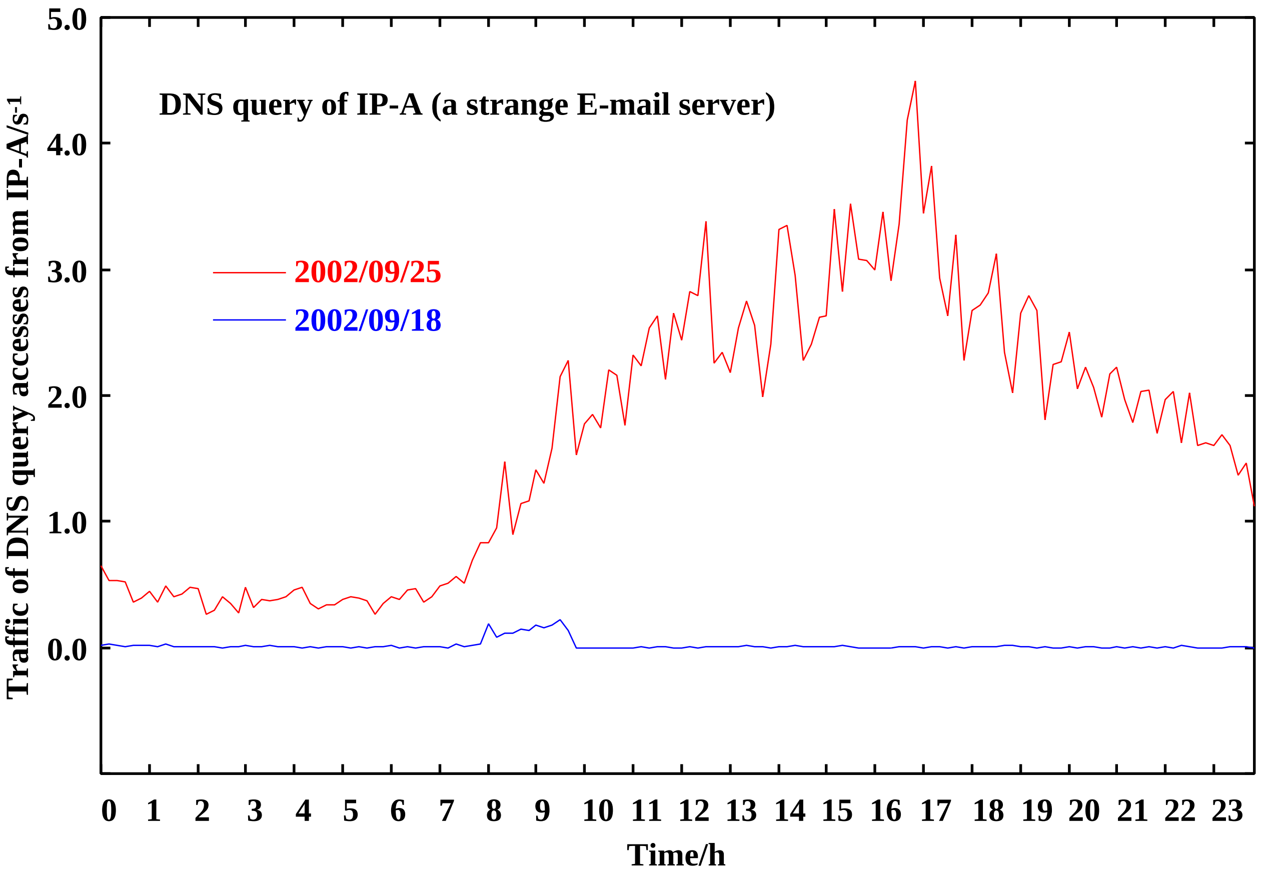
$$\boldsymbol{D}_{\mathrm{q}} = \sum_{i} \boldsymbol{D}_{\mathrm{q}}(\boldsymbol{i}) \tag{14}$$

D_{q} = the total number of the DNS query access to 1DNS.

$D_{\mathrm{q}}(i)$ = the number of the DNS query access by IP terminal i,

where i represents IP terminals A∼C is the top DNS clients of 1DNS.

D_{q} traffic curves of IP-A in normal and abnormal days

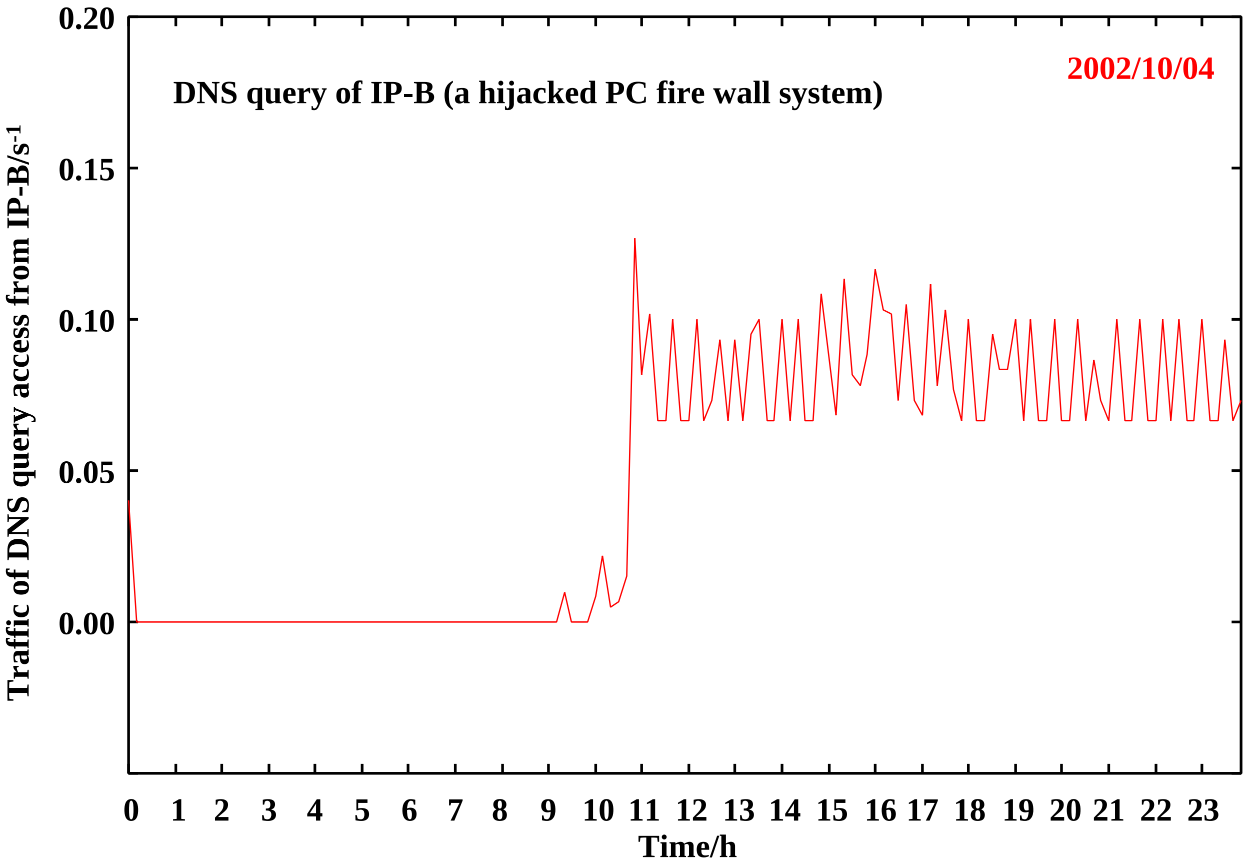


(1) In a normal day (15th), the D_{q} curve exhibits nearly zero.

(2) The D_{q} curve shows a normal curve of the E-mail server.

⇒ The DNS query cache system virtually crashes with the increase of the mass mailing worm(MMW)-infection.

D_{q} traffic curve of the Hijacked Fire Wall System



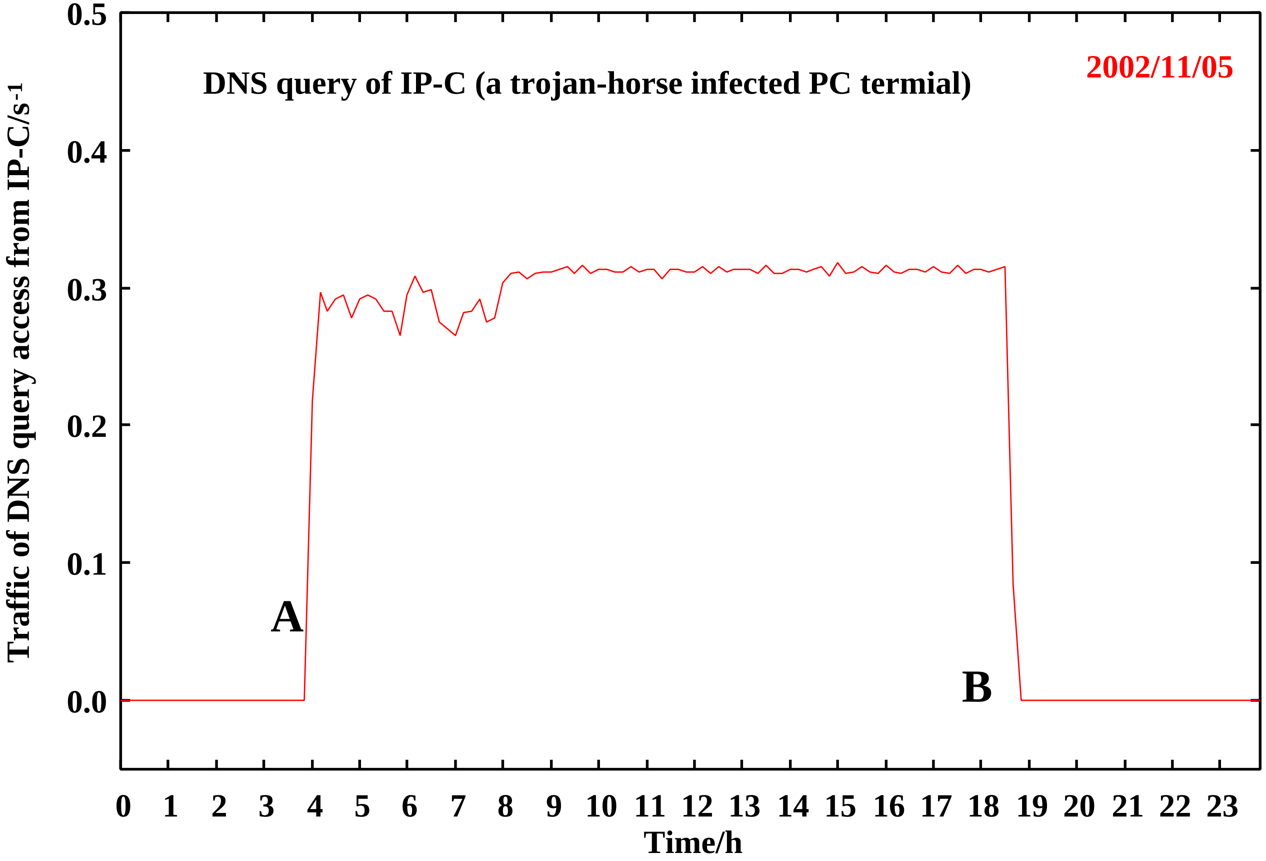
(1) The D_{q} curve shows zero in the early morning.

(2) It rises straightly upon going from 10:30 to 11:00.

(3) The rippled part can be observed after 11:00 and the system was hijacked.

⇒ The rippled curve means an indication of remotely hijacked system.

D_q traffic curve of the Trojan Horse Virus-infected PC

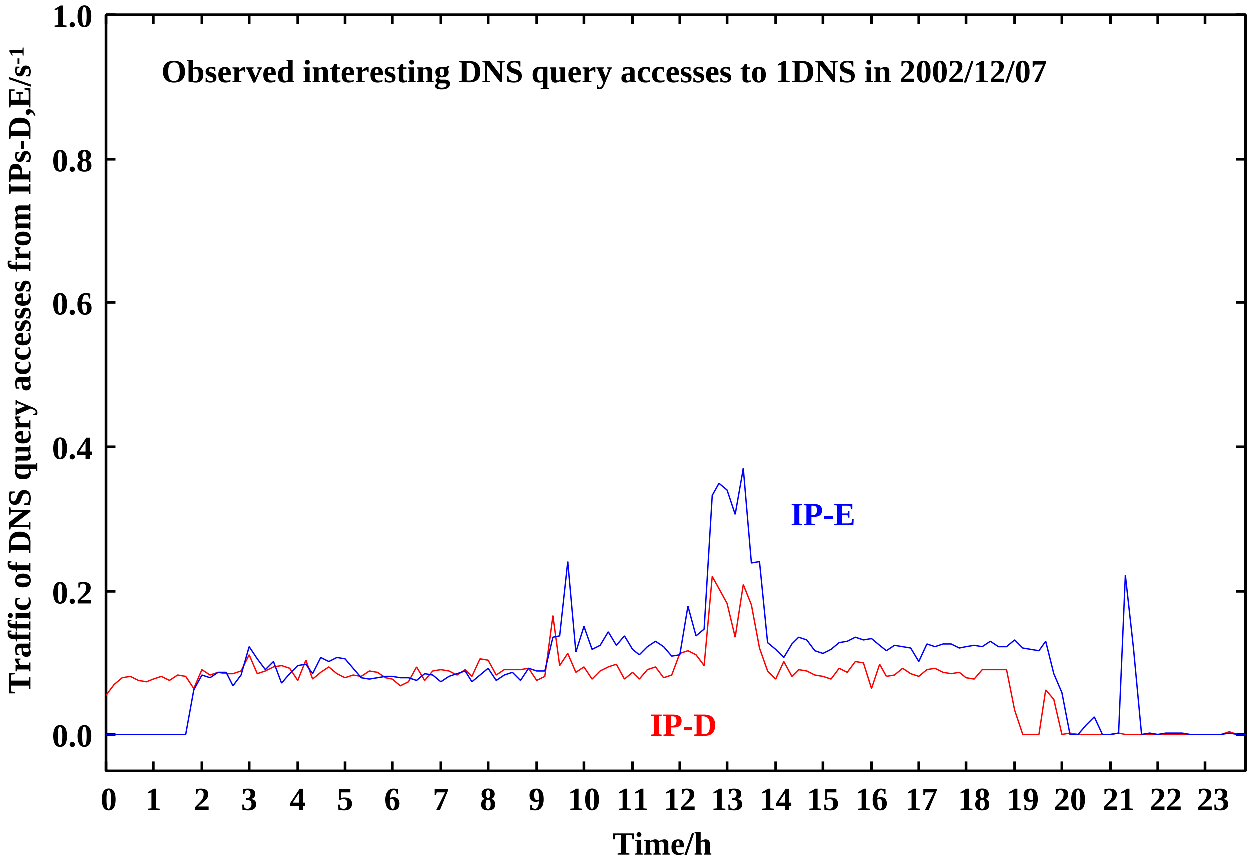


(1) In IP-C, trojan horse virus (THV), *Trojan.IrcBounce*, **is detected.**

(2) In B point, we filtered the IP-C by iptables.

We can detected THV by only observing DNS query access.

D_q traffic from the hi-jacked PC (xscan.exe)

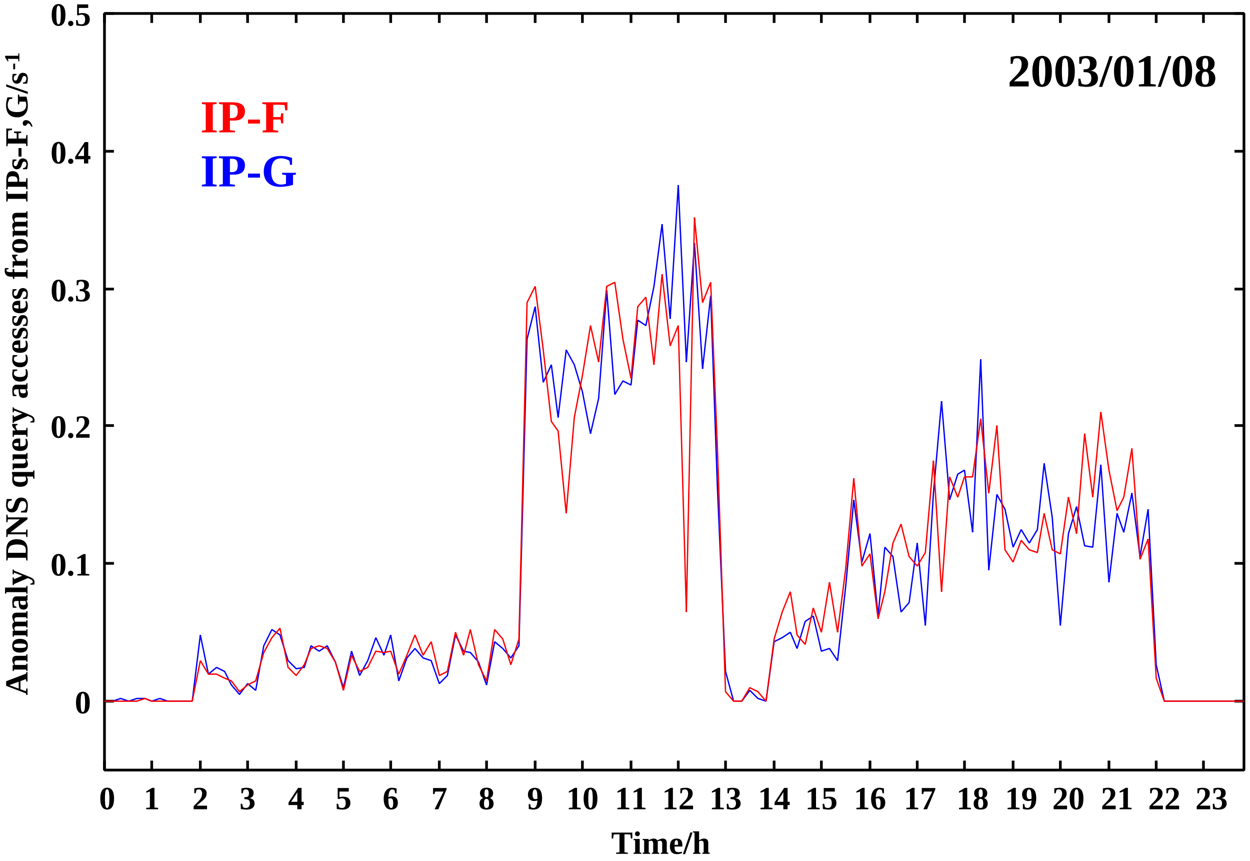


(1) The IP-D PC had been hijacked so that security scanning tools, such as xscan.exe, exec.exe, ..., etc were detected in the IP-D PC.

(2) Interestingly, the D_q(D) curve resembles well the D_q(E) curve.

(3) Regrettably, the IP-D PC attacked several network sites of outside the university through December 20th-23th, 2002.

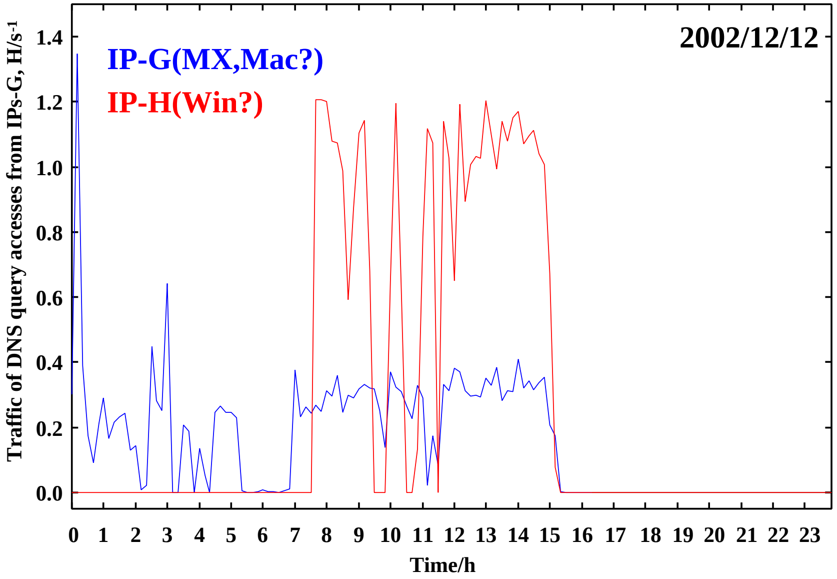
Abnormal D_{q} traffic from the outside of our university

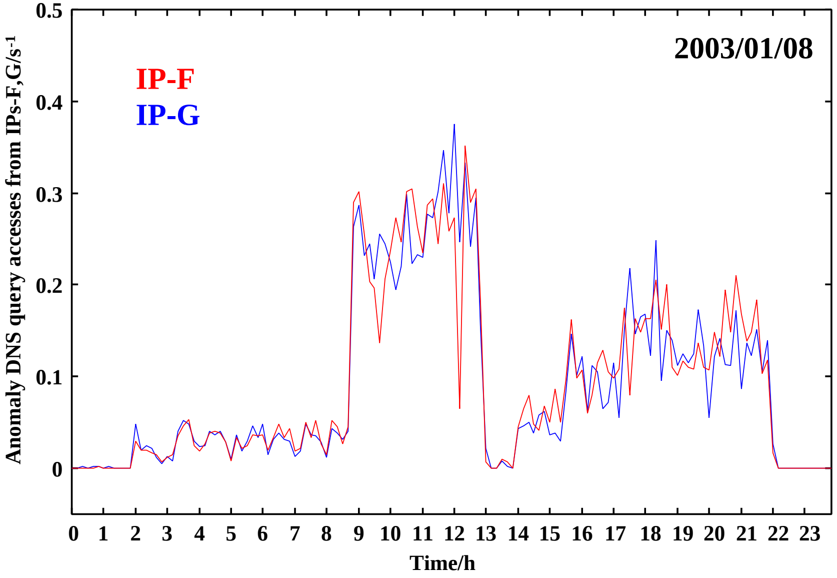


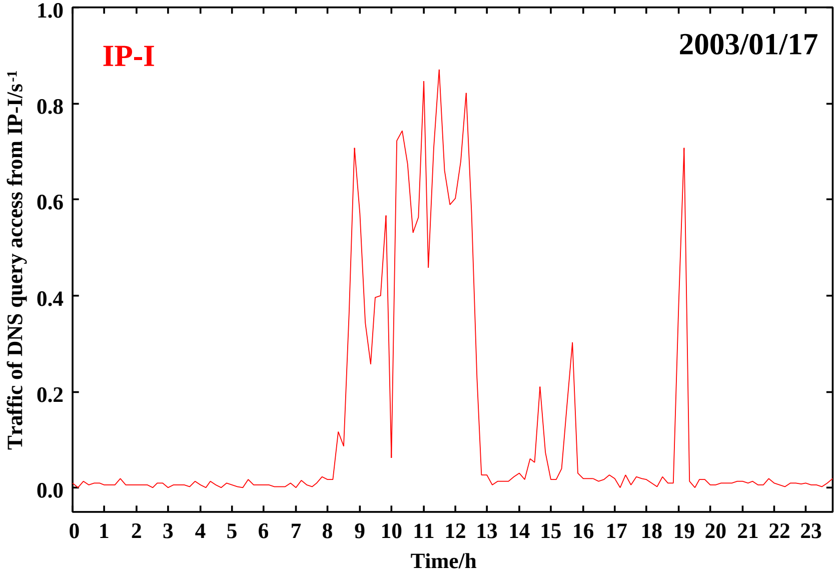
(1) The D_{q}(D) curve is considerably similar to the D_{q}(E) one.

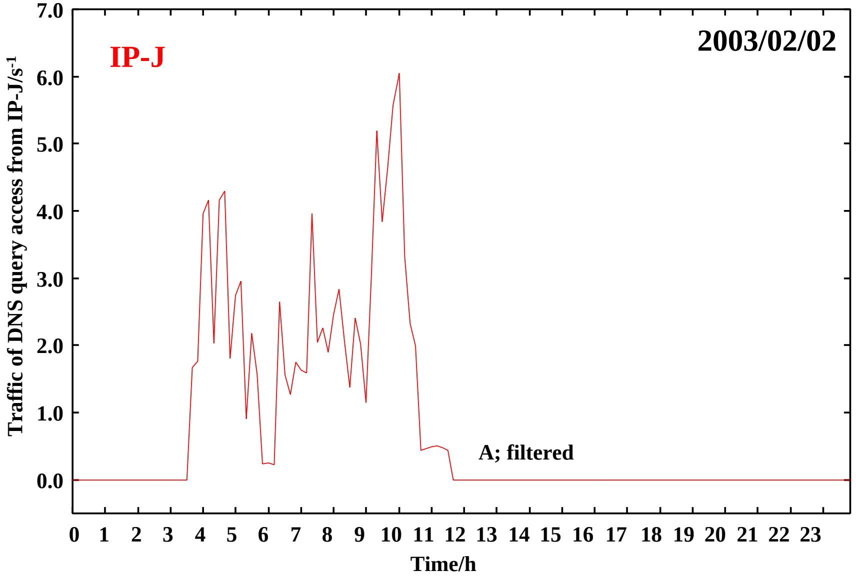
(2) Is the DNS server under a DDoS attack?

Abnormal D_q traffic from the inside/outside of our university



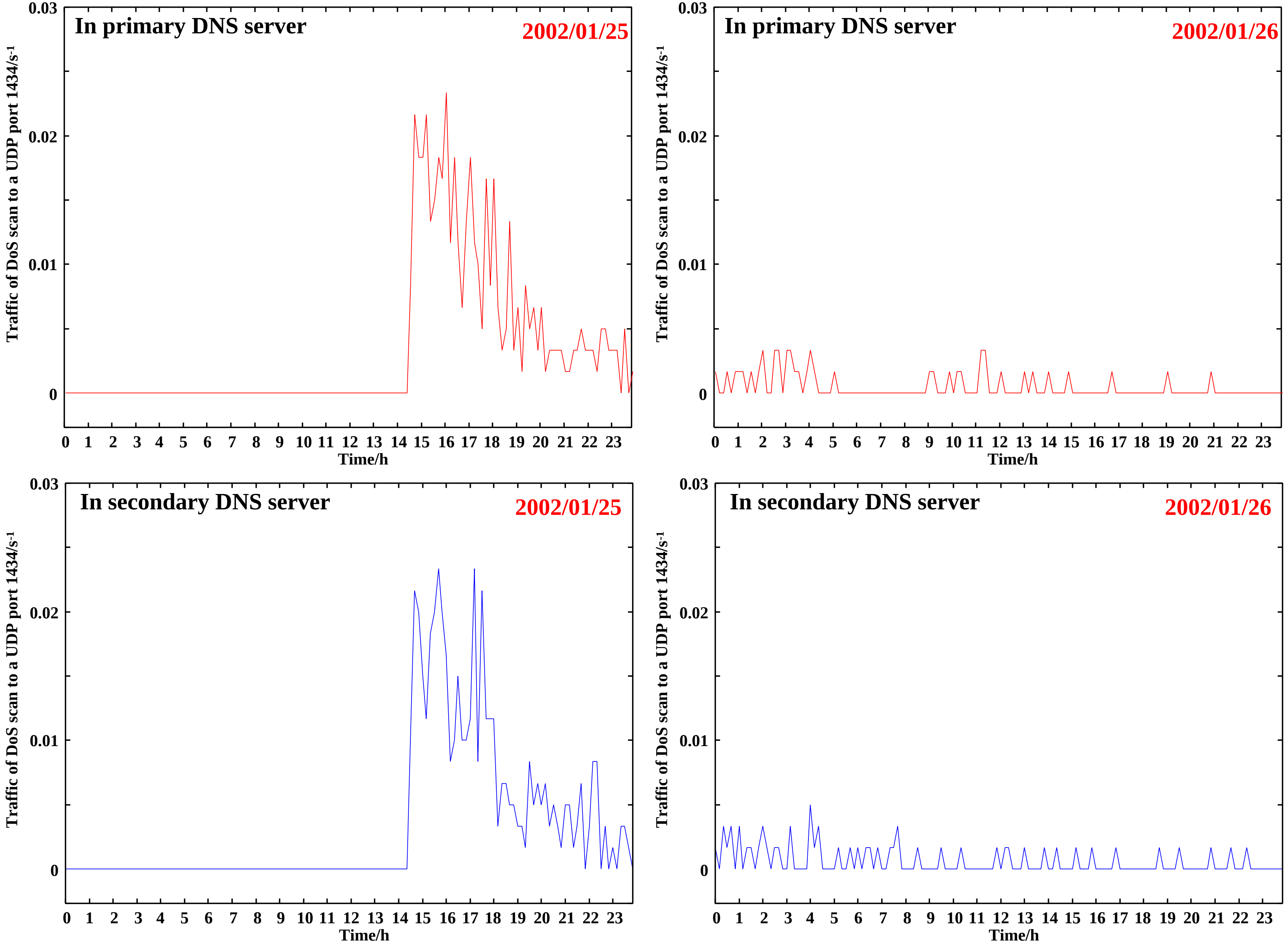




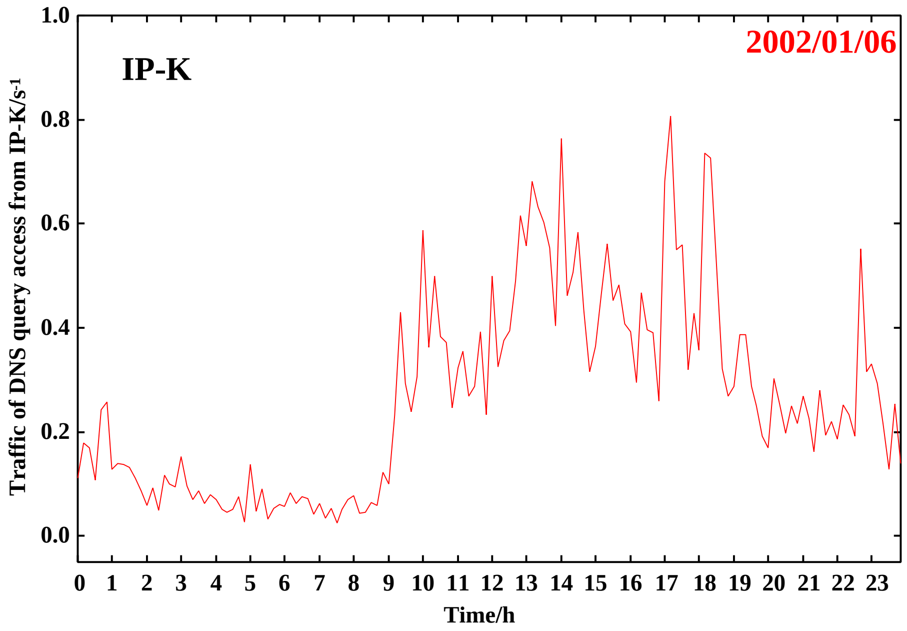


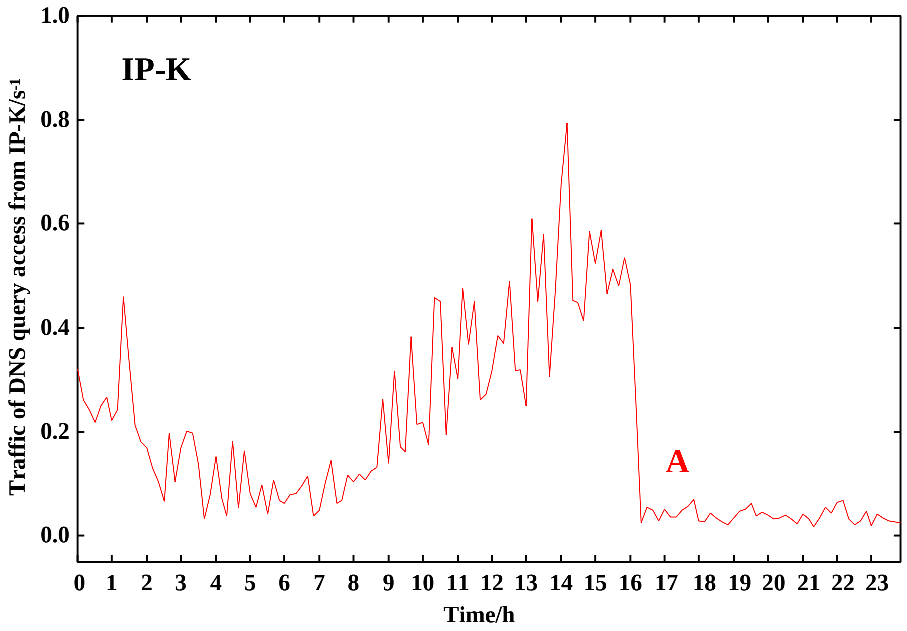
These D_q(D) curves are similar to each other.

Traffic of W32/Slammer SQL Worm



Traffic of the DNS server for a subdomain





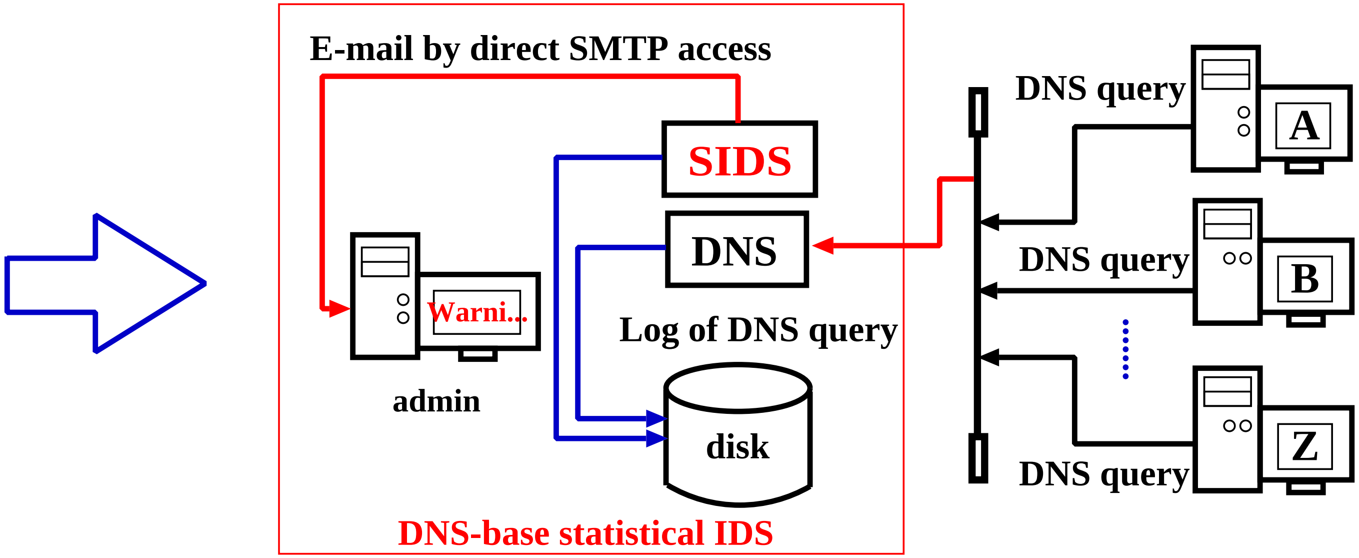
(1) After 16:00, Jan 8th, 2003, the D_{q} curve becomes to be nearly zero.

(2) We applied to an administrator of the subdomain in order to remove "forwarders;" line for /etc/named.conf.

Conclusions

(1) The DNS query traffic, D_{q}, are represented as, $D_{\mathrm{q}} = m_{\mathrm{S}}N_{\mathrm{S}} + m_{\mathrm{P}}N_{\mathrm{P}}$, where N_{S} and N_{P} represent the numbers of the SMTP and POP3 accesses, respectively. The linear coefficients m_{S} and m_{P} are given to be $m_{\mathrm{S}} = 2 + 4n(1 - q)$ and 1.0, where q is a mail-receiving rate and n is a number of different domain hosts, and the N_{S} values should be estimated by only "from=" lines ⇒ *Useful information for estimation and design of an E-mail server.*

(2) In the DNS query traffic curve, a rippled/flat curves emerge when a PC terminal is infected with virus/worm, especially mass mailing worm. ⇒ *Virus/Worm can be detected by only observing DNS query traffic or we can predict the next network security incidents.*



E-mail by direct SMTP access

SIDS

DNS

Log of DNS query

Warni...

admin

disk

DNS-base statistical IDS

DNS query

A

DNS query

B

DNS query

Z