Detecting Mass-Mailing Worm Infected Hosts by Mining DNS Traffic Data

Keisuke Ishibashi, Tsuyoshi Toyono, and
Katsuyasu Toyama
NTT Information Sharing Platform Labs.
Masahiro Ishino, Haruhiko Ohshima
and Ichiro Mizukoshi
NTT Communications Corporation
Presented by Anagha Mudigonda

Background

- Applications rely on the DNS when they access the Internet.
- By monitoring DNS traffic, we can effectively monitor the activity of those applications.
- For e.g. MyDoom.A, which attacks the SCO web site on a specific day, can be found by monitoring DNS traffic because it sends queries to the DNS server to resolve the domain name www.sco.com. to an IP address
- However it is often difficult to find such a clear characteristic query which can be used as a signature of the malware activity in DNS traffic.

Background

- For e.g. Bots are setup such that they can be controlled by an IRC channel, so there is no "characteristic query" that we can rely on.
- After mass-mailing worms have spread through the Internet number of queries sent to Internet Service Provider DNS servers has increased significantly.
- Another example: Netsky propagates by sending virus e-mail to addresses that are found in the infected host.

- Sends DNS queries to find the mail server of the address
- Same queries cannot be expected to be sent from all infected hosts by this type of worm because there is a variety of target addresses found in the infected host.

Problem Statement

- Focus on the activity of mass-mailing worms in DNS traffic, and propose a method to detect hosts infected by those worms with partial prior information about the characteristic queries.

Related Work

- Wong *et al.* analyzed the traffic behavior of mass-mailing worms, including DNS query patterns. They show that there are positive correlations between the number of SMTP flows and volume in DNS traffic, and DNS traffic can be used as signal of mass-mailing worm behavior.

Related Work

- Musashi *et al.* reported on detecting mass-mailing worms using DNS traffic data in campus networks. They assumed that the hosts that send many mail exchange queries are infected by a mass-mailing worm because normal end hosts use a local mail server to send their mail and do not send such queries.

Approach

- Use a Bayesian inference method to calculate suspiciousness of queries for domain names.
- Experimental Data captured in February 2K4(before Netsky appeared) and March 2K5(just after Netsky appeared).
- There were 31,278,205 queries, the number of unique hosts seen in the period was 221,782, and the number of unique query contents was 1,302,204.

Percentage of query types (%) - Table 1

Query type	Ratio (Feb. 2004)	Ratio (Mar. 2005)
A	77.15	46.11
MX	2.10	36.34
PTR	15.16	6.18
NS	0.21	0.05
SOA	0.93	0.56
CNAME	0.03	0.00
AAAA	1.72	1.69
ANY	0.86	0.14
SRV	1.10	0.22
Other	0.75	8.70

Observations from Table 1

- The percentage of MX queries increased from only 2% to about 36%; most of the MX queries are suspected of having been sent from the infected hosts
- Some characteristic query content sent by hosts who send many MX queries :
- The most typical query is the one whose query type is not defined.

- By monitoring the byte string of the queries, they found that this is a format error due to a bug of worms that want to send MX queries.
- They used this type of query as a signature query.
- 4% unique hosts sent this type of query (855).

Proposed Method (based on Bayesian Spam Filter)

- Objective : Detect a worm-infected host by using queries sent by the host.
- For each host h, given the query content sent by a host h as Qh, we want to calculate Pr(host h is infected $|Qh$).
- Classify each host based on whether the host sends the signature query content or not.

Proposed Method (based on Bayesian Spam Filter)

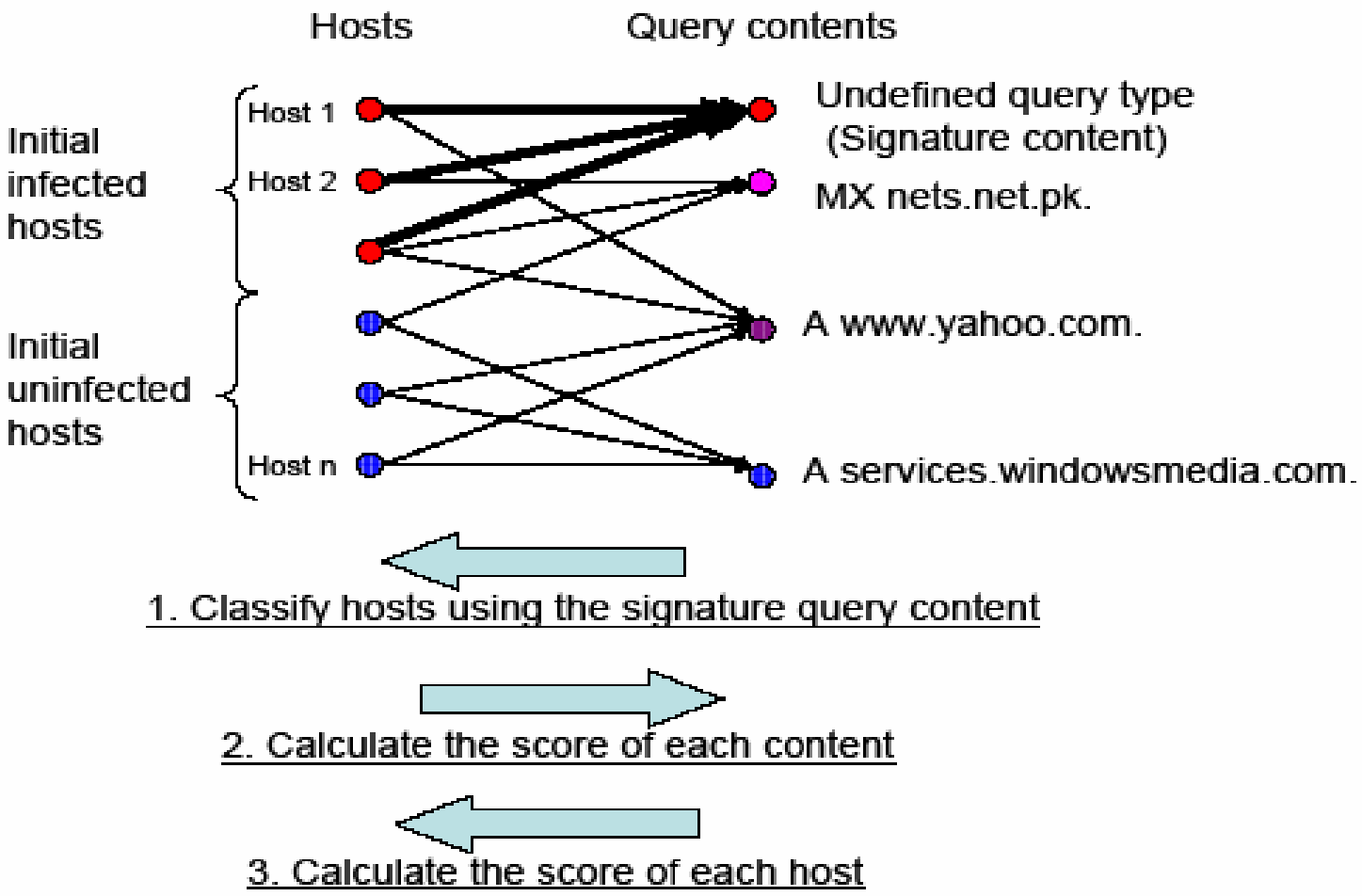

Figure 1: Host and query content pair

- If $H = \{h\}$ is the total number of hosts. Let the set of those hosts that have sent the signature content be I.
- $I := \{h \in H \mid h$ sent signature query content$\}$, and call host h in I the initially infected hosts.

Scoring Query Content

- Now we calculate the score of a query content that expresses the infection probability of a host given that the host sends the query content.
- Two ways: Host Based Scoring and Query Based.

Host Based Scoring

- Host-based scoring calculates the score of a query content based on the number of initial infected hosts who send the content.

$$I_H(q) = \frac{\#\{h \in I \mid q \in Q_h\}}{\#I}.$$

- Gives the ratio for each query content *q*. This value is the ratio of content q that is queried by initially infected hosts.
- Because even infected hosts may send queries to popular domain names, the popular query content can also obtain a high *IH*(*q*). Thus *IH*(*q*) is not appropriate to estimate the suspiciousness of the query content *q*.

- We also calculate the likelihood that content is queried by initially uninfected hosts.

$$N_H(q) = \frac{\#\{h \in \bar{I} | q \in Q_h\}}{\#\bar{I}}$$

- Then, score of query

$$S_H(q) = \frac{I_H(q)}{I_H(q) + N_H(q)}$$

- *SH*(*q*) is a rough estimate of the probability that a host sending a query "*q*" is infected by a mass-mailing worm.
- With this calculation, a query sent by only one host is scored as 1 or 0, depending on whether the host is an initially infected host or not which is not realistic.
- We modify the probability using constants *Pinit* and *Ninit* as follows.

$$S'_H(q) := \frac{P_{init} + N_q * S_H(q)}{N_{init} + N_q}$$

- where *Nq* is the number of total queries for the query content, *Pinit is* 0.5, and *Ninit* is 1.
- Initial score of a content is no longer 0 or 1 but *Pinit/Ninit*
- As the number of queries for the content increases, the score *SH*(*q*) is weighted in the calculation of *SH* (*q*).

Query-Based Scoring

- In Query-Based Scoring we also take the number of queries sent by the same host into account. So the score is calculated as

$$I_Q(q) = \frac{\sum_{h \in I} \text{number of queries of “}q\text{” sent by host } h}{\sum_{h \in I} \#Q_h}$$

- Similarly

$$N_Q(n) = \frac{\sum_{h \in I} \text{number of queries of “}q\text{” sent by host } h}{\sum_{h \in I} \#Q_h}$$

Scoring Hosts

- We then calculates the score of a host that indicates the probability that the host is infected based on both the queries sent by the host and the scores of the queries.
- It is reported that using only queries with extreme scores (near 1 or 0) and taking the geometric mean of the scores performs well for Bayesian spam filtering

- We set two thresholds, *TH* and *TL*, for high values and low values, respectively, and calculated the host score as follows

$$m = \#\{q \in Q_h | S'_H(q) > T_H\} \text{ and}$$

$$I(h) = \begin{cases} 1 & \text{if } m = 0 \\ 1 - (\prod_{\{q \in Q_h | S'_H(q) > T_H\}} S'_H(q))^{1/m} & \text{otherwise} \end{cases}$$

- Similarly

$$\text{Let } k = \#\{q \in Q_h | S'_H(q) < T_L\} \text{ and}$$

$$N(h) = \begin{cases} 1 & \text{if } k = 0 \\ 1 - (\prod_{\{q \in Q_h | S'_H(q) < T_U\}} (1 - S'_H(q)))^{1/k} & \text{otherwise} \end{cases}$$

- The degree of belief that the host is infected is calculated as

$$P(h) = \frac{1 + N(h) - I(h)}{2(N(h) + I(h))}$$

- P(h) is the probability that the host is infected and takes a value between 0 and 1.

Experimental Results

Table: Host-based score rank (S'H(q))

Rank	SH(q)	Query Content
1	0.999	Undefined query type
2	0.997	MX nets.net.pk.
3	0.997	MX gto.net.om.
4	0.995	MX sexnet.com.
5	0.994	MX lebanon-online.com.lb.
6	0.993	MX aa2.so-net.ne.jp.
7	0.993	MX domain.com.
8	0.992	MX m.
9	0.992	MX phx.gbl.
10	0.991	A dev.null.
11	0.990	MX -.
12	0.990	MX ocn.ad.jp.
13	0.990	MX hatch.co.jp.
14	0.990	MX ezweb.ne.jp.
15	0.990	MX a.
16	0.990	MX rcpt-impgw.biglobe.ne.jp.
17	0.990	MX nifty.ne.jp.
18	0.990	MX 2.
19	0.990	MX nifty.com.
20	0.990	MX h.

- By definition, the signature query gets the highest scores, while other queries have similarly high scores.
- Query content with ranks two, three, and five are domains of mail addresses for customer support service of a company and were found to be used as the sender address of some worms
- Content such as “MX m.” that is clearly expected to be queries sent by worms.

- The query content of “A www.yahoo.com” has a score of 0.53. This means that while the content is queried by many infected hosts, uninfected hosts send queries as well.
- Score of the query becomes neutral, that is, the content gives no information about whether a host is infected or not.
- The result indicates that even if a worm sends a lot of normal query traffic to confuse the algorithm, it may be filtered because that query content will be given normal scores.

Queries ranked by the number of infected hosts that sent the query

Table 3: Queries ranked by $I_H(q)$

Rank	$I_H(q)(\%)$	Query content
1	100.00	Undefined query type
2	31.92	MX nets.net.pk.
3	31.43	MX gto.net.om.
4	28.50	A img.yahoo.co.jp.
5	28.34	A ai.yimg.jp.
6	27.85	MX sexnet.com.
7	26.71	A pa.yahoo.co.jp.
8	24.76	A www.yahoo.co.jp.
9	20.20	A i.yimg.jp.
10	19.71	MX lebanon-online.com.lb.
11	15.96	A ca.c.yimg.jp.
12	15.47	A search.yahoo.co.jp.
13	15.47	A rd.yahoo.co.jp.
14	14.82	A srd.yahoo.co.jp.
15	14.33	A dailynews.yahoo.co.jp.
16	13.52	MX domain.com.
17	12.54	A shopping.yahoo.co.jp.
18	11.73	A headlines.yahoo.co.jp.
19	11.07	A wpad.
20	10.59	MX ezweb.ne.jp.

- Domains with a popular web site are highly ranked among the suspicious content listed in
- We cannot detect suspicious queries by simply monitoring the queries sent by hosts that send the signature query.

- The frequency plots of host-based scores and query based scores are shown

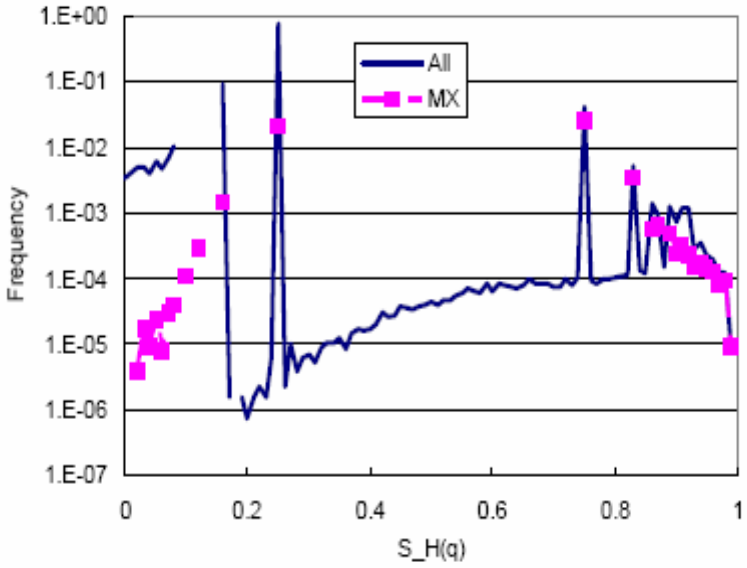

Figure 2: Frequency of query score (host-based)

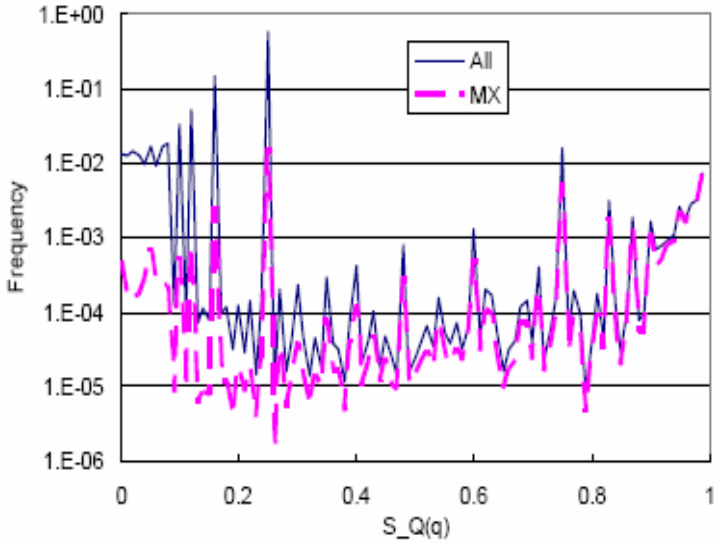

Figure 3: Frequency plot of query score (query-based)

- There are several spikes, especially on scores 0.75 and 0.25. If there is only one initial infected (uninfected) host that sent the query, then the content is scored as 0.75 (0.25)
- Shape of spikes for host-based and query-based are similar, proving that both methods are equivalent.

- Calculated host scores using *TU* to 0.95 and *TL* to 0.05.

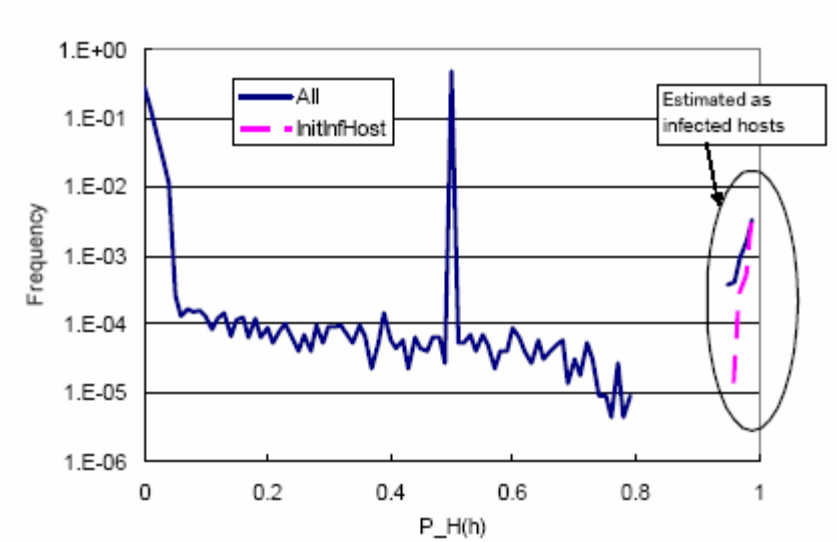

Figure 4: Frequency plot of host score

- There is a high spike at 0.5
- These are hosts that do not send a query whose score is larger than *TU* or smaller than *TL* *i.e.* neutral hosts.
- Change the values of *TU* and *TL*, leads to a trade-off between false positives and false negatives.

- Frequency of host scores that have sent signature queries. (pink line) The difference between the two lines in the high-score area indicates the number of hosts that can be detected by this method, but not detected by the signature method.

My thots

- Won't work for a worm that uses the local mail server e.g. the iloveyou virus
- Will lead to a DOS on the dns server when the network is under attack.
- TH and TL values are directly taken from those that are supposed to be best for spam. In spam we can tolerate false negatives but not false positives. Here it is the opposite.