Spam Traffic Characterization*

Jung-Yoon Kim[1] and Hyoung-Kee Choi[2]
[1] School of Information and Communications Engineering, Sungkyunkwan University
Chunchun-dong 300, Suwon, South Korea
[2] School of Information and Communications Engineering, Sungkyunkwan University
Chunchun-dong 300, Suwo, South Korea
E-mail: [1]steal83@ece.skku.ac.kr, [2]hkchoi@ece.skku.ac.kr

Abstract: In recent years, email traffic has increased in the Internet. However, as the usage of emails has increased dramatically, the amount of spam traffic has increased tremendously too. There are many filters for spam detection and prevention, but most of them are not sufficiently efficient or accurate for use in practical filtering systems. To enhance the efficiency and accuracy of spam filtering, we first need to characterize and analyze the spam traffic. To accomplish this, we first collected spam traffic by capturing the email traffic on the backbone network of Sungkyunkwan University over a period of 24 hours and then classifying it into several categories using various approach levels in order to discover the characteristics and patterns within it. This study forms the basis for research into efficient and accurate spam filtering techniques.

1. Introduction

In recent years, access to the Internet has been available all over the world. As the extent of Internet access has increased, email has become widely used. Email is widely used not only in business, but also for personal communication. Nowadays the email is a typical means of communication along with cellular-phones and messengers. However, as the usage of email has increased dramatically, the number of spam mails has increased tremendously too. “Spam” means the use of any electronic communications medium to send unsolicited messages in bulk form[1]. A spam mail is a spam message which is sent via email. It is estimated that spam constitutes over 85 percent of all email traffic on the Internet[2]. There has been much research aimed at detecting and preventing spam mails, but there have been few studies on their characterization. For efficient spam detection, the characterization of spam mails first needs to be studied. In this way, it would be possible to improve the efficiency of spam detection. To accomplish this, we captured the email traffic which was sent and received between Sungkyunkwan University in South Korea and the Internet over a 24 hour period and classified it into emails with worms attached, spam mails, and ham mails. We analyzed only the outgoing emails which were transferred from Sungkyunkwan University network to the Internet. The outgoing emails reflect the aspect of the email sender's activities and specific characteristics and are thus suitable to characterize the spam traffic. We characterized and analyzed the outgoing spam traffic on the network level, the application level, and the user level. By using various approaching levels, we provided the basis for research into efficient and accurate spam filtering techniques.

* "This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2008-C1090-0801-0028)

2. Related Work

In the past, several studies were conducted to detect and analyze spam mails. Jung and Sit[3] analyzed the network traffic which is generated by using DNS queries for searching MTAs(Message Transfer Agents). Many spammers are able to send spam mails using an open-relaying MTA, and many DNS queries are sent to DNS servers to obtain the IP address of the MTA. So the DNS queries which are sent to DNS servers provide a basis for the fact that the MTA is used to send spam mails. If the MTA is frequently used to send spam mails, it is put on a blacklist as a spamming MTA. Then if any host sends a query to a DNS server to obtain the IP address of this MTA, the query will be blocked as a spamming DNS query. According to this paper, only 0.39% of all DNS queries were spamming DNS queries in 2000, but 13.96% of all DNS queries were sent for the purpose of obtaining the IP address of spamming MTAs. Gomes et al. [4] characterized a spam traffic using various parameters. They analyzed the load intensity, which means the number of distinct recipients and senders per hour, the email arrival process, email size and the number of recipients per email by characterizing the spam traffic. Lastly, they analyzed the popularity of email senders, recipients and the locality which means that email senders and recipients that have recently been referenced are more likely to be referenced again in the near future[5]. Kim et al. proposed a URL-based spam filter which instead analyzes URL statistics to dynamically calculate the probabilities of whether emails with specific URLs are spam or legitimate, and then classifies them accordingly[6].

Kim and Choi investigated a spam filter to categorize community spam into multiple sub-classes automatically[7]. They adopted the three popular machine-learning algorithms, the Bayesian, the neural network, and the support vector machine (SVM), to educate their filter and evaluate the accuracy of the three algorithms.

3. Traffic Measurement

We recorded the email traffic on December 14, 2005, on the backbone network of Sungkyunkwan University in South Korea. The traffic, which was made up of 8.5GB of outgoing email traffic and 4.5GB of incoming email traffic, was collected over a period of 24 hours using Tcpdump[8]. Tcpdump was configured to record the traffic associated

TABLE I. WORKLOAD OF THE OUTGOING EMAIL TRAFFIC

Traffic period	2005. 12. 14. 00:00 ~ 2005. 12. 15. 00:00	
Number of hosts	***Inner hosts***	96 hosts
	Outer hosts	67,503 hosts
Number of emails	***Ham mails***	3,994 mails
	Mails with worms attached	460 mails
	Spam mails	705,483 mails
Size of email traffic	***Bytes of ham mails***	4,774,971,871 bytes
	Bytes of mails with worms attached	24,829,748 bytes
	Bytes of spam mails	3,122,998,522 bytes
Number of packets	***Packets of ham mails***	5,282,488 packets
	Packet of mails with worms attached	40,147 packets
	Packets of spam mails	16,055,400 packets
Number of flows	***Ham flows***	3,983 flows
	Flows with worms attached	460 flows
	Spam flows	418,771 flows
Number of incomplete emails	***Number of emails***	1,719,529 mails
	Size of email traffic	2,500,363,979 bytes
	Number of packets	22,436,907 packets

with only port 25 (SMTP). We used port mirroring[9] to collect the email traffic which is sent and received on the backbone network of Sungkyunkwan University.

To divide the email traffic into spam mails, ham mails and mails with worms attached, we set up SpamAssassin 3.2.0[10] and ClamAV 0.90.2[11]. We tuned them for high accuracy and efficient performance by optimizing the rule which is set to classify spam mails and by installing the recent version of the engine which is able to identify new worms and the plugins which support various spam filters. We divided the spam mails into text-based spam mails and image-based spam mails by checking whether an image file was attached or not.

To confirm our dividing method, we extracted a total of 3,000 emails from the traffic, 1,000 emails at 3:00 AM, 1,000 emails at 12:00 PM, and 1,000 emails at 4:00 PM, because the most emails sent in the morning are sent at about 3:00 AM, the emails with the largest traffic size are sent at about 12:00 PM, and most emails sent in the afternoon are sent at about 4:00 PM. We divided these 3,000 emails by hand into text-based spam mails, image-based spam mails, ham mails, and mails with worms attached and we compared them manually with the emails resulting from our dividing method. As a result of the comparison, our dividing method was found to have a level of accuracy for classifying the emails of over 99%.

4. Email Workload

The overview of our email workload is shown in TABLE I. In our trace, the total number of outgoing emails is 709,937. The total number of outgoing spam mails is 705,483, thus the spam mails account for 99.37% of the total emails. Most of the emails were spam mails, because some hosts in Sungkyunkwan University were infected by the spam-bots which send many spam mails to any host.

5. Email Traffic Characterization

In this section, we analyze the spam traffic on the network level, the application level, and the user level. On the network level, the characteristics of the spam traffic are the distribution of the IP addresses of the receivers' domains, the distribution of the email size, the mail inter-arrival time, the proportion of outgoing emails, and the session duration time of the email flows. On the application level, the characteristics of the spam traffic are the number of emails per flow, the number of senders' domains per host, the number of senders' domains per flow, and the number of MTAs. The MTAs are the mail servers which have the responsibility to send emails. Occasionally, some emails are relayed by several MTAs. On the user level, the characteristics of the spam traffic are the words included in the spam mails, and the number of the words per email.

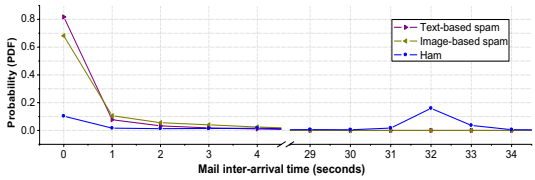

Figure 1. Probabilities of mail inter-arrival time

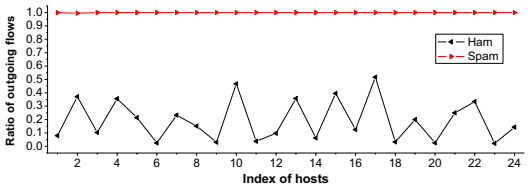

Figure 2. Ratio of flows for outgoing mails

5. 1 On the Network Level

The inter-arrival time of the spam mails is shorter than that of the ham mails, because the spam-bot sends many spam mails rapidly. A spam-bot refers to a host which is infected by a spammer's malware causing it to send spam mails on behalf of the spammer. 79.8% of the spam mails are sent in intervals of less than 1 second and their number is 562,721. 95% of the spam mails are sent in intervals of less than 5 seconds.

The number of ham mails which are sent in intervals of less than 1 second is 412, which represents 10.3% of the ham mails. This refers to the case where an email is sent to two or more receivers. 19.6% of the ham mails are sent in intervals of 32 seconds, because the sender uses a mail transfer application to send emails automatically at regular intervals. Figure 1 shows the PDF (Probability Density Function) of the mail inter-arrival time. In Figure 1, the spam mails have unique characteristics compared with the ham mails.

The spam-bot sends many spam mails, but it receives no mail because it doesn't have a domain which can be used for receiving emails and is only designed to send emails. Therefore, a high ratio of outgoing mails is a characteristic of spam mails. Although the purpose of this paper is not to analyze spam-bots but spam traffic, as more than 70% of

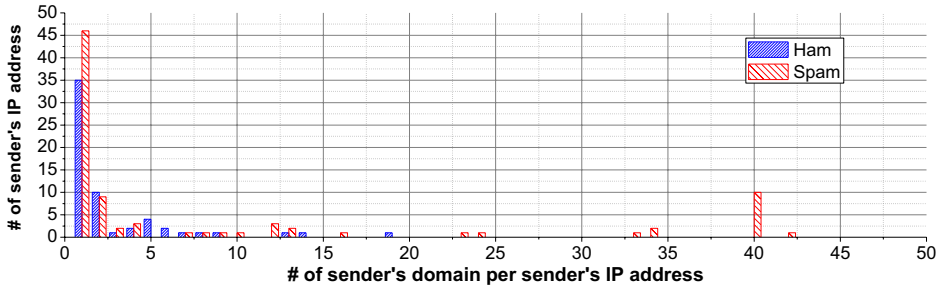

Figure 3. The number of senders' domains per host

TABLE II. THE NUMBER OF SENDERS' DOMAINS PER FLOW

The number of the senders' domains per flow	The number of flows		
	Text-based spam mails	*Image-based spam mails*	*Ham mails*
1	232,987	99,230	3,983
2	7,278	1,428	0
3	5,921	303	0
4	21,168	228	0
5	50,190	25	0
6	0	13	0

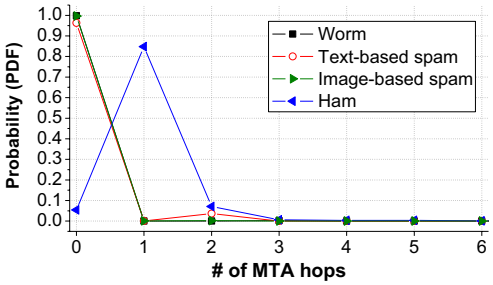

Figure 4. The distribution of the number of relayed MTAs

the spam mails are sent by the spam-bot[12], the characterization of the spam-bot is meaningful in this study. Figure 2 shows the ratio of the flows for the outgoing mails. The flows for the spam mails have a longer session time than the flows for the ham mails, because more of the flows for the spam mails are destined for foreign countries than those of the ham mails.

5. 2 On the Application Level

The emails are able to be sent to numerous receivers' domains per host, whereas the number of senders' domains per host. As the spammer generates the sender's mail address which is spoofed, however, the number of senders' domains per host in the spam mails is greater than that in the ham mails. Figure 3 depicts the number of senders' domains per host. In Figure 3, the number of senders' domains in the ham mails is always less than 20, whereas that in the spam mails is sometimes more than 40.

When an email is sent to several receivers, the sender's domain is generally fixed, because one host has one domain. Hence, in a legitimate flow, the number of sender's domains cannot be two or more. If a flow has two or more sender's domains, the emails in the flow are spam mails. TABLE II shows the number of senders' domains per flow.

We analyzed the number of relayed MTAs which transfer the emails to the receivers. The number of relayed MTAs in the spam mails is different from that in the ham mails, because the spam-bots which have the function to send emails don't use the MTAs. Figure 4 depicts the number of MTAs which relay the emails. In Figure 4, most of the spam mails are not relayed by the MTA. As supposed, they are sent by the spam-bots which transfer the spam mails using their own mail transferring function rather than the MTAs.

5. 3 On the User Level

The spammers send short spam mails which have only core words. Figure 5 shows the probabilities of the number of words in the text-based spam mails, image-based spam mails, and ham mails. As shown in Figure 5, 15% of the text-based spam mails have no words in the contents. In this case, the spam messages are included only in the subject or these are incomplete emails. 34.2% of the total number of text-based spam mails contain between 180 and 190 words, because the same mails are sent to many receivers. The image-based spam mails have images attached to them, which contain the spam messages. 68.8% of them contain 1,300~1,500 words. 11% of the ham mails have 911~920 words, because the same mails are continuously sent by one host from 12 AM to 9 AM. In our trace, the number of the ham mails is small, which makes this a special case. The ham mails have a flatter distribution than the spam mails shown in Figure 5.

The spammers use rare words in the spam mails to avoid spam filters. As typical spam filters detect spam mails by identifying the spam words, if these spam words aren't included in the spam mails, they are not detected by the word-based spam filters. For example, a spammer who intends to avoid spam filters uses the word 'Vragra' instead of 'Viagra'. The word-based spam filters cannot filter the word 'Vragra', so the spam mail with the word is sent to the receiver. This means that the number of rare words in the spam mails is a crucial characteristic of spam mails.

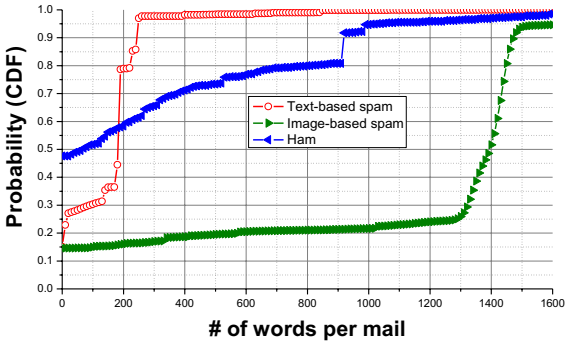

Figure 5. The number of words in all mails

6. Discussion

Many practical spam filtering systems usually fall into the two categories: 1) weight-based systems using various algorithms, and 2) machine-learning based systems using algorithms which learn the characteristics of spam mails automatically. In the weight-based systems, they use many algorithms together, so there is a large overhead. In addition, if a spam is intended to avoid various filtering algorithms, it is not detected by the weight-based systems. In the machine-learning-based systems, they need pre-learning time to be performed well, and many spam mails must be collected to evolve the systems. If a new type spam is sent, however, it is not detected by the systems because it does not pre-learned as before.

To overcome these flaws of the practical systems, we analyzed the spam traffic on the network level, the application level, and the user level. Especially, the filtering systems based on the network level and the application level detect spam mails accurately even though the headers and contents of the spam mails are modified. Furthermore, as the systems do not have to consider the headers and contents of the spam mails, they reduce more overhead than established spam filtering systems.

To deploy our characterization and analysis in practical spam filtering systems, we need to get threshold values of spam parameters from our characterization and analysis. In the practical spam filtering systems with SpamAssassin, the system administrators manually set these thresholds based on their own decisions. If they need to set the optimum thresholds, the theoretical analysis is required. We remain this work for the future work.

7. Conclusion

We analyzed spam mails by characterizing the spam traffic. In the past, several studies were conducted which focused on filters which detect and prevent spam mails and on the characterization of the spam traffic on the application and user levels.

In this paper, we characterized the spam traffic not only on the application and user levels, but also on the network level. In this way, we were able to discover some new characteristics of spam mails. This can form the basis for further research on spam mails and provide information on the characteristics of spam traffic.

In our trace, most of the spam mails were sent rapidly in intervals of less than one second, and they had no relayed MTA. Furthermore, the number of words in contents of the spam mails was biased. These results of the spam traffic characterization can be used for spam detection and prevention system. The application of these results for the system is our future work.

To characterize the spam traffic, we considered only the outgoing spam traffic which was sent from Sungkyunkwan University to the Internet. This allowed us to identify the various patterns of the spam sources and the detailed characteristics of the spam mails.

We classified the email traffic which was captured on the backbone network of Sungkyunkwan University into spam mails, ham mails and mails with worms attached, and the spam mails were further classified into text-based spam mails and image-based spam mails. This classification provided us with the detailed characteristics of the spam mails. Eventually, our study will contribute to research not only on the characterization of spam traffic, but also on filters used for spam detection and prevention.

References

[1] Neville, H., "In Defense of Spam," *IEEE Computer*, vol. 38, no. 4, pp. 86-88, 2005

[2] Barracuda Networks, http://www.barracudanetworks.com/ns/downloads/Barracuda_WP_Spam_Trends.pdf

[3] Jaeyeon, J., Emil, S., "An Empirical Study of Spam Traffic and the Use of DNS Black Lists," *11th Internet Measurement Conference*, pp. 370-375, Taormina, 2004.

[4] Luiz, H. G., Cristiano, C., Jussara, M. A., Virgílio, A., Wagner, M. Jr., "Characterizing a Spam Traffic," *11th Internet Measurement Conference*, pp. 356-369, Taormina, 2004.

[5] Almeida, V., Bestavros, A., Crovella, M., Oliveira, A., "Characterizing reference locality in the www," *In Proceedings of IEEE Conference on Parallel and Distributed Information Systems*, pp. 92-107, Miami Beach, 1996.

[6] Jangbok, K., Kihyun, C., Kyunghee, C., "Spam Filtering With Dynamically Updated URL Statistics," *IEEE Security and Privacy*, vol. 5, pp. 33-39, 2007.

[7] Bumbae, K., Hyoungkee, C., "A Prototype for Community Spam Filter based on Machine Learning Algorithm," *22nd International Technical Conference on Circuits/Systems, Computers and Communications*, vol. 3, pp. 1061-1062, Pusan, 2007.

[8] Tcpdump, http://www.tcpdump.org

[9] Port mirroring, http://www.juniper.net/techpubs/software/junos/junos70/swconfig70-policy/html/sampling-config21.html

[10] SpamAssassin, http://spamassassin.apache.org

[11] Clam AntiVirus, http://www.clamav.net

[12] Zdnet, http://news.zdnet.co.uk/security/0,1000000189,39167561,00.htm