EMICS: E-mail based Malware Infected IP Collection System

Taejin Lee[1], Jin Kwak[2]
[1] Department of Computer engineering, Hoseo University, Korea
[e-mail: kinjecs0@gmail.com]
[2] Department of Cyber Security, College of Information Technology, Ajou University, Korea
[e-mail: jkwak.security@gmail.com]
*Corresponding author: Jin Kwak

Received July 29, 2017; revised December 13, 2017; accepted January 24, 2018; published June 30, 2018

Abstract

Cyber attacks are increasing continuously. On average about one million malicious codes appear every day, and attacks are expanding gradually to IT convergence services (e.g. vehicles and television) and social infrastructure (nuclear energy, power, water, etc.), as well as cyberspace. Analysis of large-scale cyber incidents has revealed that most attacks are started by PCs infected with malicious code. This paper proposes a method of detecting an attack IP automatically by analyzing the characteristics of the e-mail transfer path, which cannot be manipulated by the attacker. In particular, we developed a system based on the proposed model, and operated it for more than four months, and then detected 1,750,000 attack IPs by analyzing 22,570,000 spam e-mails in a commercial environment. A detected attack IP can be used to remove spam e-mails by linking it with the cyber removal system, or to block spam e-mails by linking it with the RBL(Real-time Blocking List) system. In addition, the developed system is expected to play a positive role in preventing cyber attacks, as it can detect a large number of attack IPs when linked with the portal site.

Keywords: email, spam, botnet, malware, threat intelligence

1. Introduction

At present, cyber attack is increasing at a fast rate. The Symantec analysis report recently revealed that on average around one million malicious codes appear every day, as malicious codes are continuously distributed by websites that distribute malicious codes and e-mails[27]. The risks are high because the infection of a PC by a malicious code occurs without the PC user's awareness. Also, there are limitations in making an efficient response to malicious codes, as they use intelligent and hidden techniques (e.g. the obfuscation technique). Cyber attack is not limited only to cyber space, as vital infrastructures such as nuclear power plants and water, gas and other power facilities are no longer safe either. Analysis of how such cyber attack occur has revealed that a user's PC on the internal network is infected by a malicious code, and that the configuration of the internal network and the connection information of major systems are monitored to control those systems and achieve the intended purpose of the attack. To prevent these attacks, it is important to detect and take actions against botnets. **Table 1** shows botnet detection techniques.

Table 1. Botnet Detection Techniques

Division	Technical Field
Passive Approach	- Packet Inspection, Analysis of Flow Records, - DNS-based Approaches, Analysis of Spam Records - Analysis of Log files, HoneyPots
Active Approach	- Sinkholing, Infiltration, DNS Cache Snooping - Tracking of Fast-Flux networks - IRC-based Measurement and Detection - Enumeration of peer-to-peer networks
Others	- Reversing, C&C forensics and Abuse Desks

This paper proposes a method of detecting a PC infected with a malicious code automatically by analyzing the e-mails. In particular, we developed a system based on the proposed model and operated the system for more than four months, and then detected 1,750,000 attack IPs by analyzing 22, 570,000 spam e-mails in a commercial environment. It is expected that a detected attack IP can be utilized effectively to remove the malicious code from the infected PCs and block spam e-mails by linking it with the cyber removal system and the RBL system.

The contents of this paper are as follows. Section 2 analyzes the trends and characteristics of existing studies designed to detect spam e-mails and attack IPs from e-mail. Section 3 presents detection logic to analyze the characteristics of each type of spam e-mail. Section 4 presents the results of the analysis of spam e-mails in a commercial environment. The last section presents the conclusion and describes future research.

2. Related Work

Several studies have been conducted regarding the relationship between spam e-mails and botnet. ZMarkoff predicted a long time ago that attacks that exploit zombie PCs would emerge as serious attacks [1]. Zhuang analyzed the characteristics of the botnet group by analyzing the same spam e-mail in the Hotmail web service [2]; John developed a Botlab prototype to

monitor and analyze spam-sending botnets continuously [3]; and Zhao developed a BotGraph to detect large amounts of spam botnets [4]. Based on these researches, Xie proposed the AutoRE Framework, which can generate a signature, and detect a spam e-mail based on the botnet [5]. Thomas researched a method of determining an active account by analyzing Twitter spam e-mails, but it is not quite relevant to the detection of PCs infected with a malicious code[6]. Ramachandra proposed a detection method based on spammers' network behavior, but there can be various types of false negative [7]. Berkhin, Becchetti, et al. proposed a method of detecting a group of PCs infected with a malicious code, using the characteristics of spam e-mails, such as the similarity of the URL included in the main text, e-mail sending time, and e-mail sending intervals. This approach can involve false negatives because large quantities of the same e-mail can be sent [8, 9, 10] by the normal e-mail sender. In addition, Lin, Akinyelu, et al. proposed a spam bot detection method based on machine learning, such as the support vector machine [11, 12, 13]. Duan, Qaroush et al. researched the spam characteristics and extracted the general characteristics of spammers and various statistics[14,15,16,17]. Among these things, the network path is good feature as well as the mail contents for detecting the spam mail. Sanchez researched the forgery of spam delivery paths[18] and hu, Yong proposed the non-contents based spam filtering framework[19]. And, Wang and Duan analyzed the outgoing message of the spam mail[20,21]. To detect a PC infected with a malicious code in spam e-mail, Jeong introduced the concept of the IP-pollution level and the group pollution level. He adopted a method of detecting PCs infected with a malicious code by converting suspicious symptoms by IP and group into scores, where a PC is regarded as an infected PC if the score exceeds a certain level. This method should obtain a result based on the statistical characteristics of many e-mails, but it has the shortcoming of producing many false negatives above all, because detection is not based on accurate factors. Furthermore, this method makes a real-time response difficult and entails a significant number of arithmetic operation processes, because statistical data by IP should be created every hour and every day, and arithmetic operation should be performed by a combination of such data, in order to estimate the score [22,23]. And, Lin proposed reputation measurement against malicious feedback[24]. Lee implemented a method that is quite simple, but which is equipped with real-time characteristics, by concentrating on factors that can identify a spam e-mail using a single EML only, by removing the ambiguous concept of "score". Lee broadly proposed three detection methods (e.g. Direct-to-MX). Detection based on the open relay vulnerability needs to be improved because it requires a not inconsiderable amount of arithmetic operation, while its effectiveness is not very significant [25]. This paper removes the detection method based on the open relay vulnerability, as it is not very effective, among the detection methods proposed by Lee. The case of one received header was also excluded from the scope of detection, as there are various types that cannot be defined as an attack IP (e.g. webmail advertising). Conversely, the detection method based on timer order and intervals between received headers was added, as it is frequently used by attackers these days. Emphasis was placed on developing and operating a system that considers actual operation so that analysis results could be obtained in a commercial environment (large quantities of spam e-mails).

3. Proposed Scheme

3.1 System Overview

This system proposed in this paper receives an EML file as the input, and automatically

determines whether the pertinent EML has been sent by a PC infected with a malicious code or not, and then presents the grounds for such determination. In particular, we developed a system based on the proposed model, and tested and verified it for more than four months in a commercial environment. Section 4 describes the test results in full. **Fig. 1** shows the configuration diagram of the developed system.

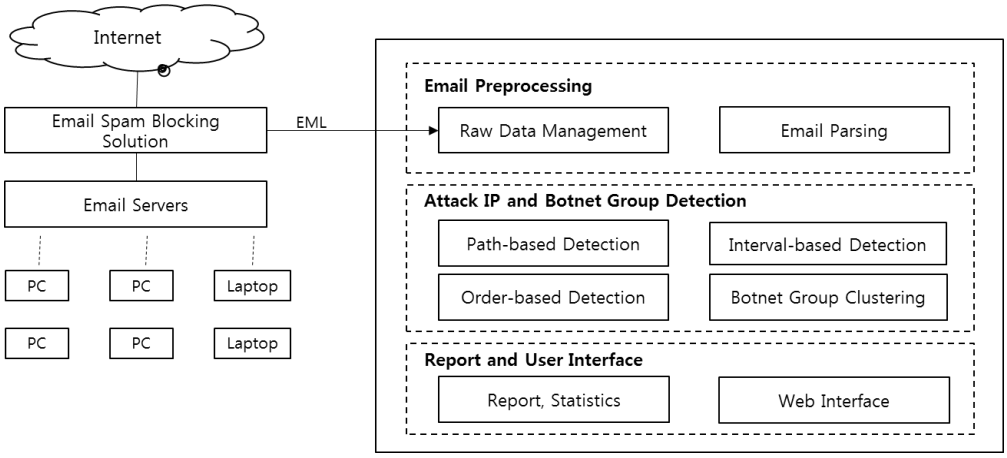

Fig. 1. System Overview

The proposed system is largely composed of the following three parts: E-mail Pre-processing, which stores and manages the original of the incoming EML file, and parses and manages the major data needed to detect the EML sent from the attack IP; Attack IP and Botnet Group Clustering, which detects the spam e-mail, attack IP, and botnet group, and manages the detection results, by applying spam e-mail detection logic; and Report ad User Interface, which manages the status of unique attack IP detection by period, various statistical information, and web interface through separate batch jobs, in order to improve performance based on the detection results.

3.2 Malware Infected IP Detection

Most spam e-mails are generated by PCs infected with a malicious code. Indeed, according to the Kaspersky Report, 80-95% of all spam e-mails appear to be sent by PCs infected with a malicious code [26,27]. Analysis of the major patterns of spam e-mail sending has shown that PCs infected with a malicious code have an internal SMTP(Simple Mail Transfer Protocol) function, and receive the spam content provided by the attacker and directly send it to the incoming e-mail server. As spam e-mails do not pass through the outgoing e-mail server, they cannot be monitored easily at a single point, which creates an environment that is more advantageous to the sending of spam e-mails. At this point, the trusted sender's e-mail address (e.g. a global company) is used to deceive the recipient into believing that the spam e-mail has been sent by a reliable sender. The attacker also uses a method of adding a random received header when sending a spam e-mail to avoid detection. Received headers are added one by one after an e-mail has been sent until it arrives at the destination. A received header added on the way to the destination cannot be manipulated, because it is beyond the scope of the attacker's manipulation. As more than two received headers generally appear if the attacker sends a spam e-mail to the incoming e-mail server, the spam e-mail can be easily detected using the number of received headers. Therefore, the attacker also uses a method of putting in the randomly manipulated header in advance. **Table 2** shows an example of the representative spam e-mail

sending pattern.

Table 2. Forged Header-based Spam E-mail Example

MAIL_FROM: Jordioxxx@xxx.com ORG_RCPT_TO: zzlbjxxx@tongxxx.net RCPT_TO: zzlbjxxx@tongxxx.net X-SPAM-TYPE: SPAM X-HELO: ehlo xxx.com X-RECEIVED-IP: 5.141.xxx.xxx Received: from 5.141.xxx.xxx at Sun, 20 Dec 2015 23:25:44 +0900 by mail.com with ESMTP CrediShield X-MAIL-FROM: Jordioxxx@xxx.com Received: from unknown (65.206.xxx.xxx) by mtu67.syds.xxx.net with SMTP; Sun, 20 Dec 2015 06:36:54 -0800 Received: from unknown (101.224.xxx.xxx) by smtp18.xxx.com with ASMTP; Sun, 20 Dec 2015 06:24:13 -0800

Table 1 shows major headers related to the attack pattern only. At a glance, it seems that the spam e-mail was sent from 101.224.xxx.xxx and received at 23:00, December 20, 2015, via smtp18.xxx and mtu67.xxx. However, smtp18.xxx and 65.206.xxx.xxx, and mtu67.xxx and 5.141.xxx.xxx mismatch. As a result, we can see that the two received headers (below) were manipulated and added by the attacker discretionally. The actual e-mail-sending IP is 5.141.xxx.xxx, and the e-mail was directly sent to the incoming e-mail server (zzlbjxxx@tongxxx.net) without passing through the outgoing e-mail server. The spam e-mail pattern described above can be detected from the perspectives of a transfer path, and the attack IP can be detected using the time information while adding a random header. The time sequence should be correct throughout the entire process (arriving at the destination after passing through the transit points), and the time interval between the sending and receiving of an e-mail should be within a certain period of time. If the attacker manipulates a certain header, a situation can occur that cannot happen in the normal e-mail process. Therefore, we can detect a spam e-mail using this method. **Table 3** shows spam the e-mail detection and attack IP detection logic.

Table 3. Attack IP and Spam E-mail Detection Mechanism

n : the number of the received headers $From_i$: the value of the $from\ field$ of the received header(i) By_i : the value of the $by\ field$ of the received header(i) $From_0$: the value of the $from$ of the e-mail body repeat i from n to 2 begin if $From_i$ and By_{i-1} do not match begin *we can decide that this e-mail is generated by the attacker.* *we can detect that Attack IP is* $From_i$. *and all the received headers between header(0) ~ header(i-1) are randomly generated by the attacker.* terminate end end if $From_1$ and $From_0$ do not match, then *we can decide that this e-mail is generated by the attacker.*

We can decide that Attack IP is $From_1$,
and that the e-mail body is randomly forged by the attacker,
or
we can decide this e-mail is not spam mail and normal IP is $From_1$.
end

Separately, the order of the time sequence should be correct in correlation to sending time values in the received header, because the attacker can check the e-mail transfer path in advance and send a spam e-mail according to the path. Also, if the time difference among the sending time values is less than a certain level, the e-mail can be deemed to be a forged or altered e-mail.

4. Experimental Results

4.1 Attack IP Detection Analysis

The detection logic proposed in Section 3 was implemented in a system, and the attack IP detection result was analyzed for 22,570,000 spam e-mails in a commercial environment. Spam e-mails in a commercial environment were collected using the spam trap system, and data collected over a period of four months (from September 1 to December 31, 2015) were used. The spam trap system of the commercial environment generates a large number of e-mail accounts and posts them in the hidden area of the Internet to allow bot crawlers to collect the e-mail address, so that spam e-mails can be collected. **Table 4** shows the status of determining spam e-mails among the entire EML collected over the four-month period (by week), and of determining the attack IPs among all of the sending IPs.

Table 4. Spam Mail and Attack IP Detection

date	EML based Detection			IP based Detection		
	total EML	mal EML	detect rate	total IP	mal IP	detect rate
1week	1,487,004	1,235,864	83.11%	135,204	123,292	91.19%
2week	1,535,912	1,433,264	93.32%	119,289	111,268	93.28%
3week	1,731,952	1,552,057	89.61%	126,281	109,759	86.92%
4week	1,211,564	1,074,325	88.67%	99,573	87,345	87.72%
5week	990,822	873,850	88.19%	99,756	90,924	91.15%
6week	1,098,302	1,009,628	91.93%	78,052	68,735	88.06%
7week	1,959,729	1,779,627	90.81%	156,090	134,972	86.47%
8week	1,898,006	1,703,543	89.75%	158,409	137,796	86.99%
9week	519,595	403,343	77.63%	82,191	71,216	86.65%
10week	699,572	572,466	81.83%	97,525	84,259	86.40%
11week	1,545,935	1,378,107	89.14%	142,933	122,910	85.99%
12week	1,229,098	1,041,616	84.75%	127,371	105,164	82.57%
13week	1,245,704	1,053,523	84.57%	141,168	114,723	81.27%
14week	904,612	730,073	80.71%	118,750	97,100	81.77%
15week	1,342,363	963,715	71.79%	111,773	75,700	67.73%
16week	1,378,086	1,013,907	73.57%	117,623	81,798	69.54%
17week	1,015,489	790,377	77.83%	99,541	84,783	85.17%
18week	779,027	697,697	89.56%	60,899	52,451	86.13%
total	22,572,772	19,306,982	85.53%	2,072,428	1,754,195	84.64%

On average some 1,250,000 spam e-mails were received every week, and 1,070,000 spam e-mails were classified as spam e-mails, showing an average detection rate of 85.53%. In terms of the sending IPs, an 84.64% detection rate was recorded (97,000 out of 115,000 attack IPs were detected on average every week). That is, when we reviewed the test results for the last four months, about 1,750,000 attack IPs were detected in an environment in which 22,570,000 spam e-mails were received. Therefore, we can see that one attack IP was detected for every 12.8 spam e-mails. Previously, three detection methods were used to detect spam e-mails and attack IPs, i.e. path-based detection, order-based detection, and interval-based detection. This section shows the determination results obtained with each detection logic. In particular, as the result of each detection logic may differ depending on the e-mail-sending hop (relay path), the test results were sub-divided according to the number of each hop. **Table 5** shows the detection result of the path-based detection method.

Table 5. Path-based Spam E-mail and Attack IP Detection

hop	EML based Detection			IP based Detection		
	total EML	mal EML	detect rate	total IP	mal IP	detect rate
1	3,261,033	6	0.000%	136,322	1	0.001%
2	18,344,289	17,626,284	96.086%	1,050,095	1,048,629	99.860%
3	447,966	447,166	99.821%	27,062	27,062	100.000%
4	300,457	300,457	100.000%	13,110	13,110	100.000%
5~13	221,105	221,104	100.000%	10,909	10,909	100.000%
	22,574,850	18,595,017	82.371%	1,237,498	1,099,711	88.866%

The path-based detection method detects received headers that are falsely added by an attacker to the spam e-mail sending path. The detection rate was 81.823% and 86.756% based on the spam e-mail and attack IP, respectively. Most of the cases where the number of hops is 1, a PC infected with a malicious code has a built-in SMTP function and sends spam e-mails directly to the incoming e-mail server. This could be translated into an attack IP, but was processed as an undetectable area, because the same pattern appears in a normal advertising e-mail using web mail. As the detected attack IP is used to block the sending of e-mails and repair the infected PC, greater emphasis was placed on removing non-detection than on removing false negatives. About 96.2% of spam e-mails had two hops. The above detection result implies that the attacker manipulated and added one false received header and sent the spam e-mail directly to the incoming e-mail server, without passing through the outgoing e-mail server.

The order-based detection method checks whether the time sequence of the spam e-mail has been reversed or not, while it is transferred from the sending end to the receiving end after being relayed. Likewise, the case of 1 hop was excluded from the scope of detection, and spam e-mail was mostly detected when the number of hops is 2. It was observed a spam e-mail can be detected if a specific time is inserted, even if a manipulated received header has been added. **Table 6** shows the detection results obtained by the order-based detection method.

Table 6. Order-based Spam E-mail and Attack IP Detection

hop	EML based Detection			IP based Detection		
	total EML	mal EML	detect rate	total IP	mal IP	detect rate
1	3,261,033		0.000%	136,322		0.000%
2	18,344,289	17,661,119	96.276%	1,050,095	1,028,691	97.962%
3	447,966	396,723	88.561%	27,062	24,423	90.248%
4	300,457	256,534	85.381%	13,110	11,719	89.390%
5~13	221,105	157,071	71.039%	10,909	8,770	80.392%
	22,574,850	18,471,447	81.823%	1,237,498	1,073,603	86.756%

The interval-based detection method detects a spam e-mail if the time duration on the transfer path exceeds a certain threshold. The overall detection rate of this method is quite low (1.33%), but it has the strength of detecting spam e-mail that cannot be processed by the above-mentioned two methods. **Table 7** shows the detection results of the interval-based detection method.

Table 7. Interval-based Spam E-mail and Attack IP Detection

hop	EML based Detection			IP based Detection		
	total EML	mal EML	detect rate	total IP	mal IP	detect rate
1	3,261,033		0.000%	136,322		0.000%
2	18,344,289	190,865	1.040%	1,050,095	14,156	1.348%
3	447,966	32,523	7.260%	27,062	1,331	4.918%
4	300,457	32,476	10.809%	13,110	1,182	9.016%
5~13	221,105	44,443	20.100%	10,909	1,964	18.003%
	22,574,850	300,307	1.330%	1,237,498	18,633	1.506%

The detection status of three logics was analyzed, along with the result obtained from the combination of the three detection logics. A total of 8 cases can occur when combined; **Table 8** shows the number of spam e-mails and IPs for each case.

Table 8. E-mail and IP Distribution according to the Detection Mechanism

Case	Detection Logic Combination			EML		IP	
	Path_based	Order_based	Interval_based	total EML	ratio	total IP	ratio
case-1	F	F	F	3,266,016	14.47%	138,066	10.92%
case-2	F	F	T	8,193	0.04%	4,679	0.37%
case-3	F	T	F	705,624	3.13%	569,837	45.08%
case-4	F	T	T	-	0.00%	-	0.00%
case-5	T	F	F	548,915	2.43%	28,897	2.29%
case-6	T	F	T	280,279	1.24%	13,576	1.07%
case-7	T	T	F	17,753,988	78.64%	508,433	40.22%
case-8	T	T	T	11,835	0.05%	542	0.04%

The statistics presented in **Table 7** show that about 14.47% of the spam e-mails could not be detected by any logic. Each logic contributes to improving the detection rate, as 2.43% could be detected by the path-based method only, whereas 3.13% and 0.04% could be detected by the order-based method only and the interval-based method only, respectively. The amount of

spam e-mails detected by both the path-based and order-based methods came to 78.69%, which accounts for a large majority; that of the path-based and interval-based methods was 1.29%; and that of the order-based and interval-based methods was 0.09%. **Table 9** shows the results of analysis of the attack IP detection status by country. When 1,230,000 unique attack IPs were analyzed, the U.S. accounted for the largest proportion (21.277%), followed by Vietnam (10.272%) and China (8.265%).

Table 9. Attack IP Distribution

nation	distribution	mal IP	Path_based	Order_based	Interval_based
US	**21.772%**	234,248	21,862	231,098	3,203
ETC	**12.988%**	139,743	35,461	136,478	3,013
VN	**10.272%**	110,518	109,179	110,272	132
CN	**8.265%**	88,927	53,969	87,742	1,855
IN	**4.872%**	52,424	49,250	52,126	160
BR	**2.451%**	26,374	19,377	25,059	627
JP	**2.224%**	23,934	3,398	23,529	418
AR	**1.858%**	19,994	17,996	19,857	68
DE	**1.697%**	18,258	6,196	17,098	436
FR	**1.451%**	15,616	2,013	15,227	242
GB	**1.412%**	15,194	2,694	14,773	201

4.2 System Deployment

The results of running the system for more than 22,000,000 spam e-mails collected over a period of four months in a commercial environment were analyzed. The system developed in this paper can be utilized for various purposes. Basically, the system was designed to detect PCs infected with a malicious code. Therefore, the method of removing a malicious code from a PC will be described first. **Fig. 2** shows the flow by which the cyber removal system operated by the KISA interworks with the model proposed in this paper [28].

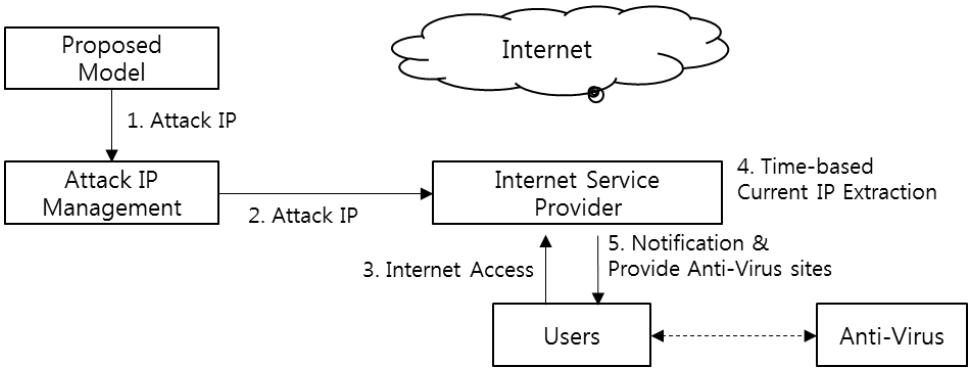

Fig. 2. System Deployment with Malware Infected PC Deletion

The cyber removal system checks the IP used by the user in tandem with the Internet service provider, provided that information regarding when and which IP has been infected with a malicious code is available. If the pertinent IP launches a DDoS attack or causes damages due to malicious code infection, the infection information is notified by a pop-up window when the pertinent user accesses the Internet, and guides the user to the anti-virus product site to take

appropriate action. The model proposed in this paper can practically remove spam e-mails from a commercial environment by linking with the cyber removal system, because it can detect an attack IP when operated with various spam e-mail systems. As a PC infected with a malicious code is the starting point of large-scale cyber attack, it is very important to take action against the attack IP before a serious incident can occur. In addition, the research findings of this paper can be effectively utilized to block actual spam e-mails. **Fig. 3** shows the process of blocking spam e-mails by linking the proposed model with the RBL system.

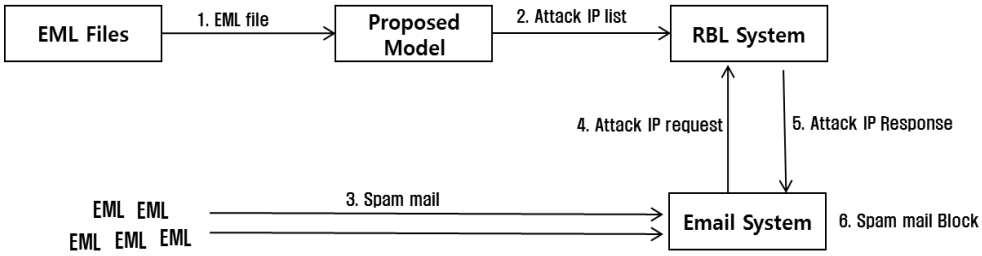

Fig. 3. System Deployment with Spam E-mail Blocking

The RBL (Real-time Block List) system manages attack IPs detected in real time. The e-mail systems within a given commercial environment check whether the e-mail sender is registered in the RBL system before transferring the e-mail to the specified recipient. If registered, the e-mail will be regarded as a spam e-mail and blocked [29,30,31]. Therefore, if the system developed in this paper is interlinked with the RBL system, it can improve the blocking of spam e-mails significantly in many e-mail systems running the RBL system.

The results of running the system for more than 22,000,000 spam e-mails collected over a period of four months in a commercial environment were analyzed. The system developed in this paper can be utilized for various purposes. Basically, the system was designed to detect PCs infected with a malicious code. Therefore, the method of removing a malicious code from a PC will be described first. **Fig. 4** shows the flow by which the cyber removal system operated by the KISA interworks with the model proposed in this paper [28].

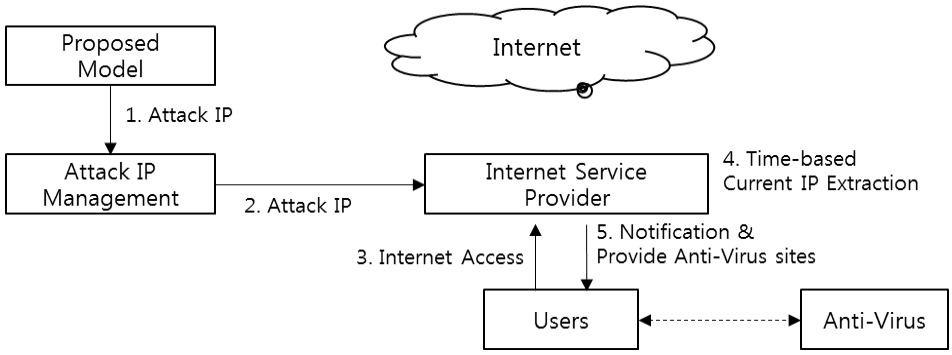

Fig. 4. System Deployment with Malware Infected PC Deletion

The cyber removal system checks the IP used by the user in tandem with the Internet service provider, provided that information regarding when and which IP has been infected with a malicious code is available. If the pertinent IP launches a DDOS attack or causes damages due to malicious code infection, the infection information is notified by a pop-up window when the

pertinent user accesses the Internet, and guides the user to the anti-virus product site to take appropriate action. The model proposed in this paper can practically remove spam e-mails from a commercial environment by linking with the cyber removal system, because it can detect an attack IP when operated with various spam e-mail systems. As a PC infected with a malicious code is the starting point of large-scale cyber attack, it is very important to take action against the attack IP before a serious incident can occur. In addition, the research findings of this paper can be effectively utilized to block actual spam e-mails. **Fig. 5** shows the process of blocking spam e-mails by linking the proposed model with the RBL system.

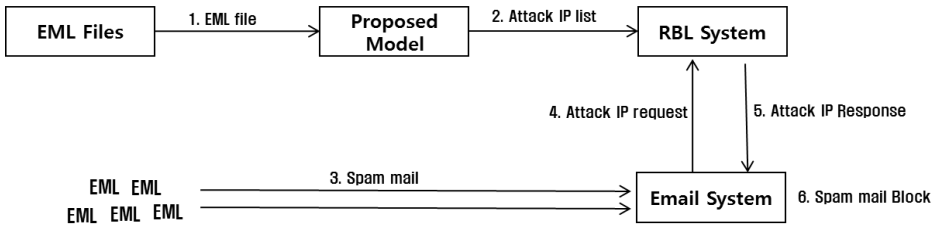

Fig. 5. System Deployment with Spam E-mail Blocking

The RBL (Real-time Block List) system manages attack IPs detected in real time. The e-mail systems within a given commercial environment check whether the e-mail sender is registered in the RBL system before transferring the e-mail to the specified recipient. If registered, the e-mail will be regarded as a spam e-mail and blocked [29,30,31]. Therefore, if the system developed in this paper is interlinked with the RBL system, it can improve the blocking of spam e-mails significantly in many e-mail systems running the RBL system.

5. Discussion

The method proposed in this paper has a few good points as follows. First, there is no possibility of a false negative, because spam e-mails are detected using characteristics that cannot appear in a normal e-mail. Second, it can detect a botnet group infected with the same malicious code and analyze the trends, besides detecting a PC infected with a malicious code only. Third, it can also be used to supplement the limitations of the pattern-based method in blocking spam e-mails. Fourth, the proposed method is very effective and powerful when applying the real-environments. Considering a portal site environment in which some 10 million e-mails are received every day, an average of 7,210,000 spam e-mails are received every day, and 473,065 attack IPs can be detected every day. Even though this estimation is based on simple ratio, significant detection effects can be obtained, because 76,000,000 spam e-mails are distributed on average each day, according to the Symantec Report published in 2015. **Table 10** shows the approximation of attack IP detection in the real environments. However, there is a potential limitation. The proposed model is mainly focused on the relayed path. If the attacker may find the relayed IP and forge the received headers according to the relayed IP as well as the EML body, the proposed model is hard to detect. However, while we analyzed hundreds of spam e-mail, we did not find these cases. As a topic of further study, the development results of this paper will be applied to commercial portal sites to verify the actual operation results. The batch processing technology needed to verify the development technology has been already applied and a technology for developing a scalable data structure and securing stability is required for the process of large-scale data (e.g. more than hundreds of thousands of e-mails).

6. Conclusions

The majority of recent cyber attacks were caused by malicious codes. Malicious codes infect the PCs in an organization through the web or by e-mail, and destroy major servers and disclose personal information by exploiting the infected PCs. This paper focuses on the detection of a given PC infected with a malicious code by analyzing the e-mail transfer path, in order to detect and respond to such cyber attacks in advance.

Table 10. Attack IP Detection Approximation

- Considering a portal site (10 million e-mails / everyday) It means 7,210,000 spam e-mails *(According to the Kaspersky Report published in 2013, spam e-mail : 72.1%)* - One attack IP sends 12.9 spam e-mails/day *(According to the Experiments)* It means 558,914 Attack IP - Our Detection Rate : 84.64% *(According to the Experiments)* **Proposed model can detect 473,065 Attack IP everyday**

As described above, it is expected that an average of 500,000 daily attack IPs will be detected in a system where an average of 10 million e-mails are distributed daily. Since the proposed method is a feature that cannot appear in normal mail, it is difficult for false detection to occur, and even if it occurs, legitimate e-mail through the outgoing mail server is not blocked. In addition, from the viewpoint of operational performance, it can be operated in a large amount of e-mail distribution environment including only static-based operation. In recent years, malicious code infected IP and spam IP are very important for threat intelligence, which is becoming important, and it is expected that good results will be obtained in EDR field when analyzing these technologies in an integrated manner.

Conflict of Interests: The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by Institute for Information & communications Technolo gy Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00683-001, Endpoint forensics and STIX analysis Machine learning based real time new malicious code detection/control system) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2017R1E1A1A01075110).

References

[1] Markoff, John, "Attack of the zombie computers is a growing threat, experts say," *New York Times*, 157, 1-3, 2007. Article (CrossRef Link)
[2] Zhuang, Li, et al, "Characterizing Botnets from Email Spam Records," *LEET*, 8, 1-9, 2008. Article (CrossRef Link)
[3] John, John P., et al, "Studying Spamming Botnets Using Botlab," *NSDI*, Vol. 9, 2009. Article (CrossRef Link)
[4] Zhao, Yao, et al, "BotGraph: Large Scale Spamming Botnet Detection," *NSDI*, Vol. 9, 2009. Article (CrossRef Link)

[5] Xie, Yinglian, et al, "Spamming botnets: signatures and characteristics," A*CM SIGCOMM Computer Communication Review*,38.4, 171-182, 2008. Article (CrossRef Link)
[6] Thomas, Kurt, et al, "Suspended accounts in retrospect: an analysis of twitter spam," in *Proc. of Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM*, 2011. Article (CrossRef Link)
[7] Ramachandran, Anirudh, and Nick Feamster, "Understanding the network-level behavior of spammers," in *Proc. of ACM SIGCOMM Computer Communication Review*, Vol. 36, No. 4, ACM, 2006. Article (CrossRef Link)
[8] Berkhin, Pavel, Zoltan Istvan Gyongyi, and Jan Pedersen, "Link-based spam detection," *U.S. Patent*, No. 7, 533,092. 12 May 2009. Article (CrossRef Link)
[9] Becchetti, Luca, et al, "Link analysis for web spam detection," *ACM Transactions.* Article (CrossRef Link)
[10] Han, K. S., Y. H. Shin, and E. G. Im, "A study of spam-spread malware analysis and countermeasure framework," *Journal of Security Engineering*, 7.4, 363-383, 2010. Article (CrossRef Link)
[11] Lin, Kuan-Cheng, Sih-Yang Chen, and Jason C. Hung, "Botnet detection using support vector machines with artificial fish swarm algorithm," *Journal of Applied Mathematics 2014*, 2014. Article (CrossRef Link)
[12] Akinyelu, Andronicus A., and Aderemi O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics 2014*, 2014. Article (CrossRef Link)
[13] Chiang, Ken, and Levi Lloyd, "A Case Study of the Rustock Rootkit and Spam Bot," *HotBots*, 7, 10-10, 2007. Article (CrossRef Link)
[14] Duan, Zhenhai, Kartik Gopalan, and Xin Yuan, "Behavioral Characteristics of Spammers and Their Network Reachability Properties," *ICC*, Vol. 7, 2007. Article (CrossRef Link)
[15] Qaroush, Aziz, Ismail M. Khater, and Mahdi Washaha, "Identifying spam e-mail based-on statistical header features and sender behavior," in *Proc. of Proceedings of the CUBE International Information Technology Conference. ACM*, 2012. Article (CrossRef Link)
[16] Al-Jarrah, Omar, Ismail Khater, and Basheer Al-Duwairi, "Identifying potentially useful email header features for email spam filtering," in *Proc. of The Sixth International Conference on Digital Society (ICDS)*, 2012. Article (CrossRef Link)
[17] A. C. Solutions. January 7, 2011 Statistics and Facts About Spam.Retrieved: July, 2011. Article (CrossRef Link)
[18] Sanchez, Fernando, Zhenhai Duan, and Yingfei Dong, "Understanding forgery properties of spam delivery paths," in *Proc. of Proceedings of 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010. Article (CrossRef Link)
[19] Hu, Yong, et al, "A scalable intelligent non-content-based spam-filtering framework," *Expert Systems with Applications*,37.12, 8557-8565, 2010. Article (CrossRef Link)
[20] Duan, Zhenhai, et al. "Detecting spam zombies by monitoring outgoing messages," *IEEE Transactions on dependable and secure computing*,9.2, 198-210, 2012. Article (CrossRef Link)
[21] Wang, Chih-Chien, and Sheng-Yi Chen, "Using header session messages to anti-spamming," *Computers & Security*, 26.5, 381-390, 2007. Article (CrossRef Link)
[22] Jeong, Hyun-Cheol, et al, "Study for tracing zombie pcs and botnet using an email spam trap," *Journal of the Korea Institute of Information Security and Cryptology*, 21.3, 101-115, 2011. Article (CrossRef Link)
[23] Jeong, HyunCheol, et al, "Detection of Zombie PCs Based on Email Spam Analysis," *KSII Transactions on Internet & Information Systems*, 6.5, 2012. Article (CrossRef Link)
[24] Huang, Lin, et al, "Using reputation measurement to defend mobile social networks against malicious feedback ratings," *The Journal of Supercomputing*, 71.6, 2190-2203, 2015. Article (CrossRef Link)
[25] Lee, et al, "Detection of malware propagation in sensor Node and botnet group clustering based on e-mail spam analysis," *International Journal of Distributed Sensor Networks 2015*, 15, 2015. Article (CrossRef Link)

[26] Kaspersky, http://www.kaspersky.com Article (CrossRef Link)
[27] Symantec. http://www.symantec.com Article (CrossRef Link)
[28] KISA. http://www.kisa.or.kr Article (CrossRef Link)
[29] KISA RBL. https://www.kisarbl.or.kr/ Article (CrossRef Link)
[30] https://www.spamhaus.org/ Article (CrossRef Link)
[31] Article (CrossRef Link)

Taejin Lee is a professor at Computer Engineering in Hoseo University. He received the Ph.D. from Ajou university. His research interests include malware analysis, network security and endpoint detection response.

Jin Kwak is a professor at Dept. of Cyber Security in Ajou University, Korea. He received the Ph.D. degree from SKKU, Korea. His research interests include Cryptographic protocols, Applied Security Mechanisms for Cloud and BigData System and so on.