Research Article

Detection of Malware Propagation in Sensor Node and Botnet Group Clustering Based on E-mail Spam Analysis

Taejin Lee,1 Hesun Cho,1 Haeryong Park,1 and Jin Kwak2

1KISA, IT Venture Tower 135, Jungdae-ro, Songpa-gu, Seoul 138–950, Republic of Korea
2Division of Information and Computer Engineering, College of Information Technology, Ajou University, Republic of Korea

Correspondence should be addressed to Jin Kwak; jkwak.security@gmail.com

Received 8 December 2014; Accepted 17 February 2015

Copyright © 2015 Taejin Lee et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cyber incidents are increasing continuously. More than 200,000 new malicious codes appear, with more than 30,000 malicious codes distributed each day on average. These cyber attacks are expanding gradually to the social infrastructure (nuclear energy, power, water, etc.) and smart sensor networks. This paper proposes a method of detecting malware propagation in sensor Node and botnet clustering automatically by analyzing e-mails. More than 80% of spam e-mails are generated by the Node infected with malicious code, using various methods to avoid filtering such as direct-to-MX, fake Received header, and open relay vulnerability. This paper proposes a scheme that detects those types accurately, including a clustering method that targets the URL included in the e-mail body, e-mail subject, attached file, and hosting server, to detect the botnet group infected with the same malicious code. The proposed method recorded about 85% zombie IP detection rate when spam e-mails distributed in a commercial environment were analyzed. When applied to the portal site that delivers 10 million e-mails, the proposed technology is expected to detect at least 150,000 zombie Nodes each day. If advanced measures are taken against the detected zombie Nodes, the spread of cyber attack damages can apparently be reduced.

1. Introduction

Cyber infringement incidents in cyber space are increasing rapidly. The AV-TEST analysis results show that over 200,000 new malicious codes appear each day on average in 2013. Bluecoat reported that over 30,000 websites distributing malicious code appear every day. Since the website that distributes malicious code uses the drive-by-download technique, which causes malicious code infection without the user’s knowledge, its risks are high. According to Gartner, more than 60% of the malicious codes used for the infringement attacks made in 2013 used unknown attack techniques. As a result, it is difficult to cope with the attack using antivirus technology that uses pattern-based detection. Those infringement attacks are not limited to online attacks only. Recently, with IT convergence services like IoT increasing by a great deal, there are reports of vehicle, cleaner, and refrigerator getting infected by malicious code. Even worse, infrastructure such as nuclear energy, water, gas, and power is no longer safe.

The US government defines cyber space as the fifth battlefield (followed by land, ocean, air, and astrospace), and it is making efforts in this area (e.g., nurturing the cyber army). The 3.20 Cyber Terror that occurred in 2013 destroyed the hard disk in many Nodes and paralyzed the computing network. At that time, the Node of the vaccine update server administrator was infected with malicious code. The Node of general users was then infected when the vaccine was updated. During the Nonghyup Hacking Incident that occurred in 2011, more than 270 servers were destroyed when the laptop that had visited the website infected with malicious code was connected to the internal network. The SK Communications Incident that occurred in 2012 disclosed more than 35 million users’ personal information because the internal user’s Node was infected with malicious code through the Alzip updater server. Similarly, the Nexon Incident that occurred in 2011 disclosed 13 million users’ personal information due to the infringement attack caused by an infected internal user's Node. In other words, most infringement incidents begin...
with infecting the user's Node on the internal network with malicious code. An attacker understands the configuration of the internal network using the infected Node and obtains the connection information of key systems by monitoring the keyboard input information. The attacker then installs more malicious code needed to make an actual infringement attack and carries out a cyber attack. To prevent these types of infringement incidents, it is important to detect and take actions against the Node infected with malicious code.

This paper proposes a method of automatically detecting a Node infected with malicious code and a botnet group, which is an operational zombie IP pool, when the attacker makes an infringement attack by analyzing incoming e-mails. Section 2 describes the characteristics of existing studies and improvements to make regarding the detection of the botnet group in the e-mail. Section 3 analyzes the attack type of three spam e-mails that can be found in the zombie Node, including the detection method and botnet clustering method. Section 4 verifies the performance of the proposed methods using spam e-mails, which are actually distributed in a commercial environment. Section 5 presents the conclusion and describes further studies.

2. Related Work

Most infringement attacks including the recent APT (Advanced Persistent Threat) exploit the zombie Node infected with malicious code. The attacker infects the user Node of the enterprise or other organizations with malicious code and understands the configuration of the internal network, including the identification and connection information of key systems, through long-term monitoring. Then, the attacker installs more malicious codes that fit the attack purpose and causes damages such as personal information leak. Botnet group detection becomes an important issue to prevent an infringement incident, and there are two main approaches [1–3]. One approach configures a honeynet and analyzes incoming malicious code in a virtual environment. If malicious behavior is found, an attack IP is detected [4–6]. Even though this honeynet approach provides a good method of detecting the botnet group, malicious code that becomes more intelligent cannot be analyzed easily if packing or encryption is used. In addition, an attack technique that deactivates malicious code is used when the virtual environment is recognized. The other approach detects the botnet group by monitoring and analyzing the network using a passive method [7–9]. The traffic characteristics of the secured botnet groups are analyzed to detect communication traffic between the zombie Node and C&C and carry out relation analysis. This approach has a significant level of false positive ratio considering the various normal traffic types, with the traffic type of PMS (Patch Management System) making analysis more difficult.

As described above, the existing botnet group detection studies have their own advantages and disadvantages. Note, however, that the detection ratio is low, and the false positive ratio is high, because all traffic is analyzed. In this study, the analysis target was limited to e-mail considering the characteristics of the spam e-mail; that is, 83.2% of spam e-mails are sent by the botnet group. In addition, actual cases of detailed attack characteristics, which can appear when the botnet group sends a spam e-mail, were analyzed and monitored to propose a more improved botnet group method than that of existing studies. The next section describes the characteristics and problems of existing studies related to e-mail-based botnet group detection and presents the result of comparison with the method proposed in this paper.

Jeong proposed a method of detecting a botnet group based on the IP and group's pollution level of the incoming spam e-mail by developing a spam e-mail trap system [10–13]. In calculating pollution, a weighted value is derived regarding the status of RBL registration, number of affected e-mails, and direct-to-MX use status. Analysis is performed based on the threshold. Note, however, that RBL is used to retrieve known information and cannot be used to determine a new zombie IP. The number of affected e-mails is not significantly different between normal e-mail and spam e-mail, except in the case of transmitting large-quantity e-mails. In addition, direct-to-MX use is an important factor, but a false alarm can occur if a false Received header is used at the same time. It is also difficult to set the threshold that distinguishes the zombie Node and the normal Node.

Zhuang proposed a method of detecting botnet groups from spam e-mails collected from the MS Hotmail system [18, 19]. Botnet groups that seem to be infected with the same malicious code, based on similar content (e.g., URL included in the spam e-mail body) and spam e-mail creation time and creation intervals, are detected. This approach can be effectively utilized to detect a botnet group if malicious code is included in the spam e-mail. Note, however, that only a limited number of botnet groups can be detected, since most spam e-mails play a simple role of delivering advertisements.

Yinglian Xie clustered the spam e-mails based on the URL included in the spam e-mail and detected the botnet group by checking whether the spam e-mail sender's IP was listed in several autonomous systems (AS) or a large quantity of spam e-mails were sent within a short period of time [20–22]. If several IPs sending the same spam e-mails in the same time slot are listed in several ASs, this method can be regarded as an effective way of detecting the botnet group considering the social engineering characteristics. Nonetheless, this method has limits in terms of accurate analysis, since the spam e-mail sending technique of the attacker (e.g., open relay vulnerability, faked Received) is not analyzed in detail.
Table 1: Comparison of related works and our proposed model.

| Author | Approach | Problem | Advantage of our proposed model |
|--------|----------|---------|---------------------------------|
| Jeong  | IP, group pollution level-based detection (RBL, MTA, threshold). | RBL can be applicable with only known IP; MTA with fake Received header cannot be detected. | 3 mechanisms can detect MTA with fake Received header, open relay vulnerability, and unknown IP. |
| John   | Botnet analysis - (network behavior-) based Detection. | Most spam mails do not have malicious network behaviors (short coverage). | Focus on spam attack types rather than malicious behavior. |
| Zhuang | Same EML contents clustering-based detection. | High false positive rate occurred. | Confirmed zombie Node-based clustering can reduce the false positive rate. |
| Xie    | URL clustering and AS time-based detection. | There are many spam mails that do not have URL. | Subject, attached file, and URL are used in the hosting server-based advanced approach. |

The method proposed by this paper has several characteristics. First, direct-to-MX, fake Received header, and open relay vulnerability, which are the spam e-mail types found in an actual commercial environment, are analyzed. Then, 8 cases are automatically detected and classified depending on whether those techniques have been used or not. In particular, we can accurately check whether spam e-mails were created in the zombie Node or not, since three spam e-mail types do not appear in normal e-mail. In the paper written by Jeong, a suspicious sign of an attack is determined by the threshold of the IP and group pollution level. As a result, a false positive or a false negative can occur. Furthermore, this approach has limits in accurate detection, since the most important factor, "direct-to-MX" attack technique, has been recently used together with the fake Received header. Second, a zombie Node can be detected in real time by analyzing only one e-mail. Existing studies like John and Zhuang cannot process many e-mails in real time through relation analysis, causing false positive cases because those studies are not based on accurate attack pattern analysis. Third, delicate clustering analysis is possible in botnet group detection based on the subject and attached file, besides the URL. The hosting server-based clustering technique is additionally applied as well. Existing studies like Zhuang and Yinglian Xie cannot cluster spam e-mails with no URL. Our approach involves the subject and the attached file, which are commonly included in all e-mails. In addition, the clustering level is increased by adding analysis with focus on the hosting server operated by the attacker, even though the spam content does not seem to be identical. Table 1 describes the difference between the existing studies and the proposed model.

3. Designs of Zombie Node and Botnet Group Detection System

3.1. Spam Mail Pattern Analysis. Currently, most of the spam e-mails are created by the Node infected with malicious code. According to the Kaspersky report, the zombie Node was used to create 80~85% of spam e-mails. The following three methods are mainly used to send the spam e-mail from the zombie Node [23]. First, the direct-to-MX technique is used; it sends e-mails to the receiving e-mail server by receiving a command from the attacker without passing through the e-mail sending server, using the built-in SMTP function of the zombie Node. The e-mail sending server can manage the e-mail accounts belonging to the server in question, and it can be easily filtered because monitoring is possible. In contrast, the e-mail receiving server can be an effective means of sending spam e-mails because it must transfer the e-mail received from various e-mail servers. When the direct-to-MX technique is used, there is no separate e-mail sending server since the zombie Node sends e-mails using the built-in SMTP function. Note, however, that the e-mail is altered as if it was sent by Gmail or Yahoo e-mail account to gain the recipient’s trust. The e-mail presented in Figure 1 shows that the sender’s e-mail address is speedy.net.pe, but the receiving...
e-mail server (211.252.150.22) has actually received the e-mail from 190.42.216.96.

Second, a false Received field is added. When the zombie Node sends e-mails, all information can be falsified/altered, such as the sender's e-mail address and header information. If the e-mail passes through the e-mail during transmission, however, the Received field is added by each relay server. Two Received fields are added to the normal e-mail if it is received by the sending e-mail server and receiving e-mail server. The e-mail system with a large number of users (e.g., Gmail, Naver) passes through many e-mail servers before reaching the receiving e-mail server. In this case, the number of added Received fields can increase significantly. The first attack technique, “direct-to-MX,” is an effective method of sending spam e-mails but has only one Received field, since those spam e-mails do not pass through the sending e-mail server. As a result, if the number of Received fields is not more than 2, it can be used to distinguish a spam e-mail. In response, many attackers add the false e-mail body as well as the false Received field that should be added by the e-mail relay server and send spam e-mails using the direct-to-MX technique. We can see that the Received header on the third line was added by the attacker in the example shown in Figure 2, because the relay server information that matches 119.165.16.3 on the first line does not appear after “by” on the third line, even though two Received headers are included.

Moreover, we can see that the direct-to-MX technique was used because the 119.165.16.3 on the first line does not match the e-mail sending domain (hosanna.net).

Third, the open relay technique, which sends spam e-mails via the e-mail server having relay vulnerability, is used. When spam e-mails are sent using the open relay technique, spam e-mail filtering can be prevented if the relay server is reliable. In the example shown in Figure 3, the sending e-mail server seems to be “yahoo.com,” but the Received header shows that the e-mail was received via the e-mail server “incheon.ac.kr” by exploiting its open relay vulnerability. At this time, 176.61.136.155 should be the IP address of the yahoo.com e-mail server, but we can see that it was manipulated using the direct-to-MX technique.

3.2. System Overview. The system proposed by this paper receives the e-mail as an input item to see if the e-mail is a spam e-mail sent by the zombie Node. This paper also proposes a technology that detects the zombie Node and botnet group (zombie Nodes infected with malicious code). As the analysis target, e-mail can be collected using the honeypot-type spam e-mail trap system; note, however, that commercial e-mails were analyzed to acquire high-quality spam e-mail samples. The proposed system is broadly composed of three modules. The first module collects the e-mail files
and performs preprocessing for easy analysis, with the second module analyzing the representative characteristics that appear when the zombie Node sends spam e-mails (direct-to-MX, fake Received header, open relay vulnerability); the third module determines the zombie Node based on the analysis results and detects the botnet group, which is the zombie Nodes infected with malicious code (Figure 4).

3.3. Zombie Node/Botnet Group Detection Mechanism. This section describes the interface and major functions of the system proposed in Section 3.2. Specifically, this section describes how to detect the zombie Node and botnet group using 7 detailed steps.

Step 1 (collecting spam mails). First, spam e-mails distributed in a commercial environment are collected. Among the incoming e-mails, spam e-mails filtered by the spam e-mail blocking solution “Spam Sniper” were extracted for analysis. Spam Sniper is a pattern-based spam e-mail blocking solution that secured more than 3,500 references sites. A total of 57,048 spam e-mails were collected for 8 days, or approximately 11.9 spam e-mails per person each day on average. The system developed in this paper is designed to analyze the e-mail file in real time when the e-mail is received. Note, however, that the botnet group can also be detected and analyzed from the large quantity of e-mails stored by the portal site, and so forth, using batch processing.

Step 2 (spam mail preprocessing). The basic trap log (e-mail received time, sender’s e-mail address, recipient’s e-mail address, URL included in the e-mail body, etc.) and the content included in the Received header, which was added while passing through the e-mail relay server, are extracted and saved in the RDBMS after normalization. The Received header is added automatically each time the e-mail passes through the relay server. Since the number of relay e-mail servers is different, depending on the environment where this system is applied at the e-mail receiving end, prior refinement work is required at the e-mail receiving end, which identifies and filters out all Received headers created afterward based on the information on the first relay server. Note, however, that Google has several relay servers at the e-mail receiving end, and the added Received headers should be identified and filtered in advance.

Step 3 (direct-to-MX detection). The attacker uses the direct-to-MX technique, which sends the spam e-mail to the e-mail receiving server directly from the zombie Node equipped with the built-in SMTP function, since e-mail sending/receiving details are monitored and abnormal signs are detected by the security manager if spam e-mails are sent by the normal e-mail sending server. Since the direct-to-MX technique does not have the e-mail sending information, the sender’s e-mail address included in the e-mail body is forged and altered to well-known addresses (e.g., Google and Yahoo) to mislead the recipient regarding the trustworthiness of the e-mail. The IP recorded in the Received header of the normal e-mail refers to the e-mail sending server; thus, it matches the domain information of the e-mail sending address in the e-mail body. If the direct-to-MX technique is used, however, the IP recorded in the Received header refers to the IP address of the zombie Node, and the domain of the e-mail sending address in the e-mail body does not match (e.g., Google, Yahoo). The direct-to-MX technique cannot be used by normal e-mail. Therefore, if this characteristic is found, we can see that the e-mail was sent by the zombie Node. If the direct-to-MX technique is used, the extracted IP points to the zombie Node IP. Otherwise, it points to the IP of the e-mail sending server. Algorithm 1 shows the spam e-mail detection logic of the direct-to-MX technique.

Step 4 (From-by tracking-based fake received header detection). As described earlier, attackers use the direct-to-MX technique most frequently. The number of Received headers is 1 when the technique is used (Algorithm 2). Since there
must be more than one Received header in a normal e-mail, the e-mail with one Received header can be filtered out. Therefore, the attacker adds a fake Received header before sending the spam e-mail to bypass this attack detection mechanism when using the direct-to-MX technique. The Received header, which is automatically added by the relay e-mail server, records the information on "when the e-mail file was delivered to itself (by) by which server (from)." Therefore, if several Received headers are added, the incoming path of the e-mail file can be figured out if the relation of those Received headers is analyzed and tracked. If the fake Received header is added, the tracking chain cannot be connected because the location of each zombie Node is different. We can detect whether the fake Received header has been added arbitrarily using "From-by" tracking. We can see that the last “prior IP” after removing the fake Received header is the IP of the zombie Node. Like the direct-to-MX technique, this method cannot be used by normal e-mail. Therefore, if this characteristic is found, we can see that the e-mail was sent by the zombie Node.

**Step 5** (open relay vulnerability detection). Open relay vulnerability refers to a technique that sends spam e-mails via the e-mail server having relay vulnerability. If spam e-mails are sent using the open relay technique, spam e-mail filtering can be prevented if the relay server is reliable. Since spam e-mails pass through the reliable e-mail server on the way, the e-mail receiving server cannot detect these spam e-mails unless all incoming e-mails are blocked. The unique domain is limited in normal e-mail, including the e-mail sending Node, even though the e-mail has been relayed several times.

**Step 6** (zombie node extraction). Steps 3–5 presented a method of detecting spam e-mails created by the zombie Node, which cannot be found in a normal e-mail. Depending on the occurrence of the three attack types described above, separate work is required to determine whether the zombie Node has been used or not and detect the IP of the actual zombie Node. If the direct-to-MX technique is used, the IP that sends SMTP traffic to the e-mail receiving server should be directly extracted. If the fake Received header is used, the zombie Node IP should be extracted by identifying and removing the fake Received header. If open relay vulnerability is used, filtering is required so that the open relay server IP is not extracted. Figure 5 shows the process of extracting the zombie Node IP and detecting the attack type considering those requirements.

**Step 7** (botnet group clustering). Zombie Nodes and attack types are classified in Step 6. Attackers use hundreds or thousands of zombie Nodes under their control to make an attack, instead of a few zombie Nodes, when sending spam e-mails. Step 7 proposes a method of clustering the botnet group infected with the same malicious code among individual zombie Nodes. The botnet group is detected by two stages. In the first stage, zombie Nodes that have sent the spam e-mail with the same e-mail subject, attached file, and URL included in the e-mail body in the same time slot are grouped and detected among the detected zombie Nodes. In the second stage, spam e-mails having the same advertisement URL hosting server are grouped and detected even though the spam content is different.

### 4. Experimental Result

A total of 57,048 spam e-mails were collected for 8 days (November 29, 2013–December 6). Based on this data, Section 4.1 detects the zombie Node by attack type; Section 4.2 presents the result of botnet group clustering.

---

**Algorithm 1: Direct-to-MX detection.**

\[
n = \text{number of Received Headers} \\
\text{for } i = 1 \text{ to } n \\
\quad \text{if (from domain in Received Header } (i) \neq \text{ by domain in Received Header } (i + 1)) \\
\qquad \text{then} \\
\qquad \quad \text{Received Header } (i + 1) \sim \text{Received Header } (n) \text{ are faked} \\
\qquad \quad \text{by attacker} \\
\qquad \quad \text{from domain in Received Header } (i) \text{ is Zombie IP} \\
\quad \text{end} \\
\text{all Received Headers are not faked}
\]

**Algorithm 2: Fake Received header detection.**
Table 2: Spam category analysis.

| Section | Attack category | Number of Samples | Ratio | Remarks |
|---------|----------------|------------------|-------|---------|
| Case 1  | X              | X                | X     | 33      | 6.87%   |
| Case 2  | X              | X                | O     | 35      | 7.29%   |
| Case 3  | X              | O                | X     | 1       | 0.21%   |
| Case 4  | X              | O                | O     | 19      | 3.96%   |
| Case 5  | O              | X                | X     | 0       | 0       |
| Case 6  | O              | X                | O     | 391     | 81.45%  |
| Case 7  | O              | O                | X     | 0       | 0       |
| Case 8  | O              | O                | O     | 1       | 0.21%   |
| Total   |                |                  |       | 480     | 100%    |

Proposed approach coverage: 93.13%

4.1. Zombie Node Detection. At least 480 spam e-mails were sampled randomly among the spam e-mails as analysis target, and the ratio of attacks was manually analyzed according to the three attack types proposed in this paper (fake Received header, open relay vulnerability, and direct-to-MX). The random sample result shows that 447 spam e-mails (93.19%), among the 480 spam e-mails collected from the commercial environment, could be detected using the system proposed by this paper (Table 2).

Analyzing the detailed characteristics of spam e-mails, 92.9% (446 spam e-mails) were based on the direct-to-MX technique (Cases 2, 4, 6, and 8). Note, however, that cases of using the direct-to-MX technique numbered only 35 (7.29%). Apparently, fake Received header and open relay techniques are also used, because only the number of Received headers can indicate the abnormal sign. When these techniques are used together, it cannot be detected using the existing method. Moreover, whether the spam e-mail has been sent by the zombie Node can be determined, but the IP of the zombie Node cannot be identified easily. The fake Received header technique is assumed to be used rarely in sending spam e-mails, since the attack effects cannot be achieved easily when used alone. Regarding the 480 spam e-mails identified above, the detection rate and false negative...
Table 3: Detection, false negative rate according to the spam cases.

| Case   | Number of samples | Detection rate | False negative rate |
|--------|-------------------|----------------|---------------------|
|        | Number of samples | Ratio          | Number of samples   | Ratio   |
| Case 1 | 33                | 31             | 93.94%              | 2       | 6.06%   |
| Case 2 | 35                | 34             | 97.14%              | 1       | 2.86%   |
| Case 3 | 1                 | 0              | 0%                  | 1       | 100%    |
| Case 4 | 19                | 18             | 94.74%              | 1       | 5.26%   |
| Case 5 | 0                 | —              | —                   | —       | —       |
| Case 6 | 391               | 384            | 98.21%              | 7       | 1.79%   |
| Case 7 | 1                 | 0              | 0%                  | 1       | 100%    |
| Total  | 480               | 467            | 97.29%              | 13      | 2.71%   |

Table 4: Zombie IP detection result.

| Date   | Zombie IP | Total IP | Ratio  |
|--------|-----------|----------|--------|
| 11-29  | 1,255     | 1,425    | 88.07% |
| 11-30  | 1,269     | 1,449    | 87.58% |
| 12-01  | 1,095     | 1,228    | 89.17% |
| 12-02  | 1,105     | 1,307    | 84.54% |
| 12-03  | 1,301     | 1,510    | 86.16% |
| 12-04  | 1,270     | 1,475    | 86.10% |
| 12-05  | 1,176     | 1,346    | 87.37% |
| 12-06  | 1,325     | 1,639    | 80.84% |
| Total  | 9,796     | 11,379   | 86.09% |

Table 5: Zombie IP mail detection result.

| Date   | Zombie IP mail | Total IP mail | Ratio  |
|--------|----------------|---------------|--------|
| 11-29  | 6,299          | 6,622         | 95.12% |
| 11-30  | 7,100          | 7,446         | 95.35% |
| 12-01  | 6,632          | 6,822         | 97.21% |
| 12-02  | 6,591          | 6,857         | 96.12% |
| 12-03  | 7,145          | 7,642         | 93.50% |
| 12-04  | 6,905          | 7,168         | 96.33% |
| 12-05  | 6,855          | 7,111         | 96.40% |
| 12-06  | 6,996          | 7,380         | 94.80% |
| Total  | 54,523         | 57,048        | 95.57% |

rate of each case in the developed system are as follows: the ratio of accurate detection (8 cases) was 97.29%, and the false negative rate was 2.71% (see Table 3).

As for the result of detecting 57,048 spam e-mails for 8 days (November 29, 2013–December 6), spam e-mails were sent from 11,379 IPs to 600 actual users for 8 days. Among them, 9,796 IPs were determined as zombie Node (average detection rate of 86.09%). Judging from the number of spam e-mails, zombie Nodes have sent 4,523 spam e-mails (95.57%) out of 57,048 spam e-mails for 8 days, telling us that each user received 11.89 spam e-mails from 2.04 zombie Nodes each day on average (Tables 4 and 5).

Figure 6 shows the detailed detection result of each case. When the domestic and overseas distribution of detected zombie IPs was analyzed, most IPs were found to be located abroad (97.13%), with spam e-mails received from the zombie Nodes abroad. When domestic zombie IPs were analyzed by ISP, major ISPs accounted for the largest proportion (Tables 6 and 7).

4.2. Botnet Group Clustering. Section 4.1 has explained the zombie Node detection result by attack type in a commercial environment. This chapter presents the detection result of the botnet group that sends spam e-mails by receiving a command from the same attacker, since zombie Nodes are infected with the same malicious code. The botnet group was clustered by content first and then clustered by hosting.

4.2.1. Contents-Based Botnet Group Detection. To determine a botnet group, the IP of the identified zombie Node was
Table 8: Botnet group detection result.

| Date          | Identifier (subject)                  | Number of mails | Number of countries | Number of IPs |
|---------------|---------------------------------------|-----------------|---------------------|---------------|
| 2013-12-01 13:00 | Where can I buy the cheapest sexual?   | 147             | 1                   | 9             |
| 2013-12-02 12:00 | Strongly arouses man’s!               | 143             | 1                   | 9             |
| 2013-12-01 13:00 | Strongly arouses man’s!               | 125             | 1                   | 9             |
| 2013-12-06 21:00 | Strongly arouses man’s!               | 125             | 1                   | 7             |
| 2013-12-03 03:00 | Where can I buy the cheapest sexual?   | 123             | 1                   | 9             |
| 2013-12-05 23:00 | Where can I buy the cheapest sexual?   | 123             | 1                   | 7             |
| 2013-12-04 16:00 | Strongly arouses man’s!               | 119             | 1                   | 9             |

Table 9: Zombie IP lists in the same botnet group.

| Date          | Nation | IP              | Case# | Number of mails |
|---------------|--------|-----------------|-------|-----------------|
| 2013-12-01 13:00 | C**    | 112.225.***.*** | 6     | 16              |
| 2013-12-01 13:00 | C**    | 119.165.***.*** | 6     | 20              |
| 2013-12-01 13:00 | C**    | 119.165.***.*** | 6     | 10              |
| 2013-12-01 13:00 | C**    | 218.57.***.***  | 6     | 10              |
| 2013-12-01 13:00 | C**    | 119.165.***.*** | 6     | 21              |
| 2013-12-01 13:00 | C**    | 112.255.***.*** | 6     | 10              |
| 2013-12-01 13:00 | C**    | 112.255.***.*** | 6     | 15              |
| 2013-12-01 13:00 | C**    | 124.135.***.*** | 6     | 25              |
| 2013-12-01 13:00 | C**    | 119.165.***.*** | 6     | 20              |

4.2.2. Hosting-Based Botnet Group Detection. Previously, we have detected a botnet group if the zombie Nodes have sent the same spam e-mail in the same time slot. Spam e-mails usually contain the advertisement link in the e-mail body. Even though the URL is different, those spam e-mails can be analyzed to see if the spam e-mail with the same content (e-mail subject, URL included in the e-mail body, and attached file) was sent in the same time slot. If it was, those zombie Nodes were regarded as a botnet group infected with the same malicious code. Table 8 shows an example of botnet group detection during the test period. The first line shows that the same 147 spam e-mails related to the sexual enhancer were sent by 9 IPs between 12:00 and 13:00 on December 1, 2013. All those 9 IPs were determined to belong to the zombie Node. The size of the botnet group was found to be small, because spam e-mails were grouped from the small group. If the system is applied to the service securing more than 10 million users like Naver and Daum, however, a significant number of botnet groups can apparently be detected.

Table 9 shows detailed information on the zombie Node belonging to the botnet group detected between 12:00 and 13:00 on December 1, 2013. A total of 9 zombie Nodes belonging to the same botnet group were located in China, having sent spam e-mails using the same attachment type (fake Received header and direct-to-MX technique).
Table 10: Same hosting based botnet group detection.

| Botnet group | Hosting IP    | Number of IPs | Number of nations | Number of URLs |
|--------------|---------------|---------------|-------------------|---------------|
| Group 1      | 89.33.*.*.*   | 340           | 50                | 290           |
| Group 2      | 199.59.*.*.*  | 174           | 38                | 10            |
| Group 3      | 199.59.*.*.*  | 127           | 33                | 12            |
| Group 4      | 110.45.*.*.*  | 1271          | 20                | 20            |

Table 11: Same hosting based botnet group lists.

| URL          | Number of IPs | Number of nations |
|--------------|---------------|-------------------|
| **** pills.com | 49            | 16                |
| **** generics.com | 12           | 9                 |
| **** iowe.com   | 2             | 1                 |
| **** shop.com   | 2             | 1                 |
| Total          | 340           | 50                |

5. Conclusions

Existing studies on the detection of the botnet group, which is a key factor of a cyber infringement incident, analyzed all network traffic, resulting in a significant false negative rate as well as limits in detecting intelligent attacks. By paying attention to the report that more than 80% of spam e-mails are sent by the botnet group, this paper has limited the analysis target to e-mails. We have analyzed the three attack characteristics (direct-to-MX, fake Received header, open relay) that appear when the botnet group sends spam e-mails, which are not found when sending a normal e-mail, and proposed how to detect those characteristics automatically. We have found that three attack types account for 85.36% when 57,048 spam e-mails collected in a commercial environment were analyzed. In addition, botnet groups were detected from the zombie Nodes using the URL included in the e-mail body, e-mail subject, attached file, and hosting server-based clustering. Since the botnet group is exploited to send spam e-mails and make preparations for various malicious behaviors (e.g., personal information disclosure, DDoS attack), it is meaningful to take proper actions against the detected botnet group. In addition, when the botnet group is analyzed, the attacker can be traced because the C&C server operated by the attacker can be tracked, and the characteristics and protocol of the malicious code can be analyzed. Moreover, the known spam e-mail is blocked using the pattern, with the pattern of suspicious e-mails added manually. When the technique proposed by this paper is applied to suspicious e-mails, blocking patterns can be created more conveniently.

Figure 7 shows that the proposed technique can be used together with existing spam response technologies.

According to the Kaspersky report in 2013, 72.1% of all e-mails are spam e-mails. Taking into account the portal...
site that receives more than 10 million e-mails each day on average, 7.2 million spam e-mails are received every day on average, and each zombie Node sends about a handful of spam e-mails. As a result, at least 150,000 zombie Nodes are expected to be detected each day on average if the proposed technique is applied. We are considering the following work to apply the system proposed by this paper to portal sites in the future. First, the system will be modified to have a scalable structure, so that new attack types that might appear in the future can be analyzed. Second, direct-to-MX and fake Received header analysis triggers numerous queries about the domain and IP, so the analysis performance will be improved using cache and other methods. Third, the system will be supplemented to support batch processing for application to portal sites, since the proposed system is currently designed to process e-mail received in real time. Fourth, the clustering result will be improved by analyzing the time series change pattern of the botnet group controlled by each attacker, similarity of the attack type cases used to send spam e-mails, and similarity of the fake domains and accounts used. Lastly, the analyzed botnet groups will be compared with the time series, attack type, and attack route of botnet groups obtained by infringement response agencies at home and abroad to increase the level of utilization.

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 the ICT R&D program of MSIP/IITP (14-824-06-001, The Development of Cyber Blackbox and Integrated Security Analysis Technology for Proactive and Reactive Cyber Incident Response) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (no. NRF-2014R1A2A1A1050818).

References
[1] M. Feily, A. Shahrestani, and S. Ramadass, “A survey of botnet and botnet detection,” in Proceedings of the 3rd International Conference on Emerging Security Information, Systems, and Technologies (SECURWARE ’09), pp. 268–273, IEEE, Glyfada, Athens, June 2009.
[2] K. S. Han, Y. H. Shin, and E. G. Im, “A study of spam-spread malware analysis and countermeasure framework,” Journal of Security Engineering, vol. 7, no. 4, pp. 363–383, 2010.
[3] R. M. Algluiev, R. M. Algluiley, and S. A. Nazirova, “Classification of textual E-Mail spam using data mining techniques,” Applied Computational Intelligence and Soft Computing, vol. 2011, Article ID 416308, 8 pages, 2011.
[4] N. C. Paxton, G. J. Ahn, R. Kelly, K. Pearson, and B. T. Chu, “Collecting and analyzing bots in a systematic honeynet-based testbed environment,” in Proceedings of the 11th Colloquium for Information Systems Security Education, Boston, Mass, USA, June 2007.
[5] J. Zhuge, T. Holz, X. Han, J. Guo, and W. Zou, “Characterizing the IRC-based botnet phenomenon,” Tech. Rep., Peking University & University of Mannheim, 2007.
[6] P. Sroufe, S. Phithakkitnukoon, R. Dantu, and J. Cangussu, “Email shape analysis for spam botnet detection,” in Proceedings of the 6th IEEE Consumer Communications and Networking Conference (CCNC’09), pp. 1–2, January 2009.
[7] G. Gu, R. Perdisci, Z. Zhang, and W. Lee, “BotMiner: clustering analysis of network traffic for protocol-and structure-independent botnet detection,” in Proceedings of the 17th Conference on Security Symposium, pp. 139–154, 2008.
[8] H. Choi, H. Lee, H. Lee, and H. Kim, “Botnet detection by monitoring group activities in DNS traffic,” in Proceedings of the IEEE International Conference Computer and Information Technology (CIT’07), 2007.
[9] J. He, J. Xu, X. Zhu, Y. Zhang, T. Zhang, and W. Fu, “Reputation-based secure sensor localization in wireless sensor networks,” The Scientific World Journal, vol. 2014, Article ID 308341, 10 pages, 2014.
[10] H.-C. Jeong, H.-K. Kim, S. Lee, and E. Kim, “Detection of zombie PCs based on e-mail spam analysis,” KSII Transactions on Internet & Information Systems, vol. 6, no. 5, p. 1445, 2012.
[11] M. Qi, Y. Guo, and H. Yan, “A reputation system with anti-pollution mechanism in P2P file sharing systems,” International Journal of Distributed Sensor Networks, vol. 5, pp. 44–48, 2009.
[12] H. C. Jeong, H. K. Kim, S. Lee, and J. H. Oh, “Study for tracing zombie pcs and botnet using an email spam trap,” Journal of the Korea Institute of Information Security and Cryptology, vol. 21, no. 3, pp. 3–18, 2011.
[13] A. A. Akinyelu and A. O. Adewumi, “Classification of phishing email using random forest machine learning technique,” Journal of Applied Mathematics, vol. 2014, Article ID 425731, 6 pages, 2014.
[14] J. P. John, A. Moshchuk, S. D. Gribble, and A. Krishnamurthy, “Studying spamming botnets using BotLab,” in Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation (NSDI’09), pp. 291–306, 2009.
[15] M. Bailey, E. Cooke, F. Jahanian, Y. Xu, and M. Karir, “A survey of botnet technology and defenses,” in Proceedings of the Cybersecurity Applications & Technology Conference for Homeland Security (CATCH ’09), pp. 299–304, IEEE, Washington, DC, USA, March 2009.
[16] K. Singh, A. Srivastava, J. Griffin, and W. Lee, “Evaluating email’s feasibility for botnet command and control,” in Proceedings of the IEEE International Conference on Dependable Systems and Networks with FTCS and DCC, June 2008.
[17] K. Saraubon and B. Limthannaphon, “Fast effective botnet spam detection,” in Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology (ICCIT ’09), pp. 1066–1070, IEEE, Seoul, Republic of Korea, November 2009.
[18] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar, “Characterizing botnets from e-mail spam records,” in Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, 2008.
[19] L. Feng, H. Wang, Q. Han, Q. Zhao, and L. Song, “Modeling peer-to-peer botnet on scale-free network,” Abstract and Applied Analysis, vol. 2014, Article ID 212478, 8 pages, 2014.
[20] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, “Spamming botnets: signatures and characteristics,” in Proceedings of the ACM SIGCOMM Conference on Data Communication (SIGCOMM ’08), pp. 171–182, August 2008.
[21] K.-C. Lin, S.-Y. Chen, and J. C. Hung, “Botnet detection using support vector machines with artificial fish swarm algorithm,” *Journal of Applied Mathematics*, vol. 2014, Article ID 986428, 9 pages, 2014.

[22] P. Graham, “Different methods of stopping spam,” 2003, http://www.windowsecurity.com/.

[23] MessageLabs, http://www.messagelabs.com.
Submit your manuscripts at http://www.hindawi.com