SPAM ZOMBIE DETECTION USING SPOT

Mrs. Chaitrali Chaudhari, Ms. Sonali G. Doiphode

[1, 2] *Computer Engineering, Mumbai University, (India)*

ABSTRACT

Email spam are measure problem on Internet.These email spam message may contain code which is used to execute different malicious activities ranges from online searching of data, phishing, accessing lists, moving files sharing channel information to DDoS attacks against click fraud. Finding machines which are used to send spam messages is very important to prevent these type of activities.These compromised machines in a network that are involved in the spamming activities, are known as spam zombies. In this paper we present effective solution for detecting spam zombies named "SPOT". SPOT is designed based on a powerful statistical tool called Sequential Probability Ratio Test(SPRT), which has bounded false positive and false negative error rates. For comparison two more spam zombie detection algorithms are studied based on the number and the percentage of spam messages originated by the machine.

Keywords: Compromised Machines, Spot Detection System, Spam Filter, Spam Zombies.

I. INTRODUCTION

E-mail spam, also known as unsolicited commercial e-mail or unsolicited bulk e-mail. These are unwanted e-mail messages frequently sent with commercial content in large quantities to an indiscriminate set of recipients. Spam is technically delivered the same way as legitimate e-mail, using the Simple Mail Transfer Protocol. Network of compromised machines is called as botnet. Botnet is the serious threat which occurs commonly in today's cybercrimes and cyber-attacks. A large fraction of spam comes from botnets,. E-mail spam detection is an effective strategy for subsequent botnet detection. Botnet performs predefined functions in an automated fashion, and executes different malicious activities ranges from online searching of data, accessing lists, critical targets, phishing, moving files sharing channel information to DDoS attacks etc. Command and control(C&C) infrastructure makes the functioning of Botnet unique; in turn throws challenges in the mitigation of Botnet attacks. This paper, focuses on the detection of the compromised machines in a network that are used for sending spam messages, also called as spam zombies. Two natures of the compromised machines on the Internet—sheer volume and widespread—render many existing security countermeasures less effective and defending attacks involving compromised machines extremely hard. A number of recent research have studied the aggregate global characteristics of spamming botnets (networks of compromised machines involved in spamming) such as spamming patterns of botnets and the size of botnets. Instead of studying aggregate global characteristics of spamming botnets, we develop a tool for system administrators to automatically detect the compromised machines in online manner. This paper, present a spam zombie detection system, called as SPOT, by analyzing outgoing messages. SPOT system is designed based on a statistical tool called Sequential Probability Ratio Test (SPRT), developed by Wald [1]. As a simple and powerful statistical method, SPRT has many desirable characteristics. It minimizes the required number of observations for decision among all the sequential and non-sequential statistical tests. This means that the SPOT detection system can identify a compromised machine quickly. Both the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds.

II. LITERATURE SURVEY

In this section, we discuss related work in detecting compromised machines.Recent studies [2], [3] investigated the aggregate global characteristics of spamming botnets including the size of botnets and the spamming patterns of botnets. These studies provide aggregate global characteristics of spamming botnets by clustering spam messages received at the provider into spam campaigns using embedded URLs and near-duplicate content clustering, respectively. These schemes are better suited for large e-mail service providers to understand the aggregate global characteristics of spamming botnets than the individual networks to detect internal compromised machines. Also online detection is not supported by them. DBSpam tool detect proxy-based spamming activities in a network relying on the packet symmetry property of such activities [4], which is developed by Xie et al.,Not only the spam proxies but the aim is to detect all types of compromised machines which are involved in spamming. Here we have few botnet detection schemes. Gu et al., developed BotHunter [5] detects compromised machines by correlating the IDS dialog trace in a network. BotHunter relies on the specifics of the malware infection process,while SPOT focuses on the economic incentive behind many compromised machines and their involvement in spamming. An anomaly-based detection system named BotSniffer [6] identifies botnets by exploring the spatial-temporal behavioral similarity commonly observed in botnets. It focuses on HTTP-based and IRC-based botnets. BotMiner [7] is both structure and protocol independent. In this technique, flows are classified into groups based on similar malicious activity patterns and similar communication patterns. The intersection of the two groups is considered to be compromised machines. Compared to general botnet detection systems such as BotHunter, BotSniffer, and BotMiner, SPOT is a lightweight compromised machine detection system.

III. PROBLEM FORMULATION AND ASSUMPTIONS

Figure 1 illustrates the logical view of the network model. Assume that messages originated from machines inside the network will pass the deployed spam zombie detection system. SPOT only requires a sufficient view of the outgoing messages originated from the network in which it is deployed. A machine in the network is assumed to be either compromised or not compromised. In this paper, focus is on the compromised machines that are involved in spamming. The term a compromised machine is used to denote a spam zombie. Let Xi for i = 1, 2, . . . denote the successive observations of a random variable X corresponding to the sequence of messages originated from machine m inside the network. Let Xi = 1 if message i from the machine m is a spam, and Xi = 0 otherwise. The detection system assumes that the behavior of a compromised machine is different from that of a normal machine in terms of the messages sending. Specifically, a compromised machine will generate spam messages with a higher probability than a normal machine.

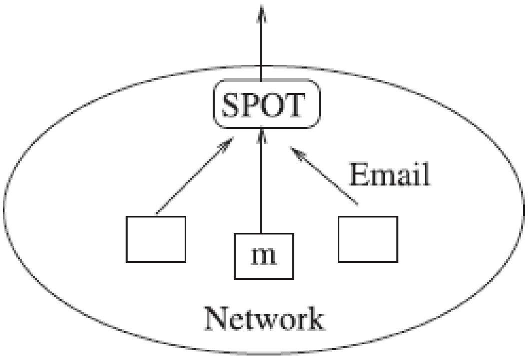

Figure 1: Network Model

Formally,

$$P\,r(X_i = 1|H_1) > P\,r(X_i = 1|H_0), \qquad (1)$$

where H_1 denotes machine m is compromised and H_0 denotes the machine is normal.

For the convenience assume that a sending machine m as observed by the spam zombie detection system is an end-user client rather than a mail relay server. The proposed SPOT system can manage the case where an outgoing message is forwarded by a few internal mail relay servers before leaving the network. Detection system uses spam filter (Content-based), which classifies an outgoing message as either a spam or non-spam. The spam filter does not need to be perfect in terms of the false negative rate and false positive rate. The spam zombie detection problem can be formally stated as : Xi arrives sequentially at the SPOT system, the system determines with a high probability if m has been compromised. After decision is taken, the detection system declare the result, and further required actions can be taken, e.g., to clean the machine.

IV. ALGORITHMS

4.1 Spot Detection Algorithm

SPOT is designed based on the statistical tool SPRT. In SPOT, H_1 is considered as a machine is compromised and H_0 as machine is normal. In addition, let X_i =1 if the i^{th} message from the concerned machine in the network is a spam, and $X_i = 0$ otherwise. When an outgoing message arrives at the SPOT system, it records the IP address of message sending machine. Then using content-based spam filter message is classified as either ham or spam. Spot maintains the logarithm value of the corresponding probability ratio Λn for every IP address of sender machine. When a machine is identified as being compromised it is added into the list of potentially compromised machines. Once the machine is declared as compromised, it not to be further monitored by SPOT. On the other hand, a machine which is currently normal may get compromised at a later time. Therefore, normal machines are continuously monitored by SPOT. Once such a machine is identified by SPOT, the records of the machine in SPOT are reset so that a new monitoring phase starts for the machine.

Algorithm 1: spam zombie detection system SPOT

```
An outgoing message arrives at SPOT
Get IP address of sending machine *m*
// all following parameters are specific to machine *m*
Let n be the message index
Let $X_n = 1$ if message is spam, otherwise $X_n = 0$
if ($X_n$ = = 1) then
    //spam , 3
    $\Lambda_n$ += $ln\ (\theta_1/\ \theta_0)$
else
    // nonspam
    $\Lambda_n$ += $ln((1-\ \theta_1)/(1-\ \theta_0))$
end if
if ($\Lambda_n$ >= B) then
    Machine m is compromised. Test terminates for *m.*
else if ($\Lambda_n$ < = A) then
    Machine *m* is normal. Test is reset for *m*
```

```
$\Lambda_n = 0$
Test continues with new observations  values
else
Test continues with an additional observations
end if
```

4.2 Spam Count and Percentage Based Detection Algorithms

Two more algorithms in detecting spam zombies are discussed here. First one is based on the number of spam messages and second is on the percentage of spam messages sent from an internal machine. We call them as the count-threshold (CT) detection algorithm and the percentage-threshold (PT) detection algorithm. In CT, the time is partitioned into windows of fixed length T. A user-defined threshold parameter C*s* specifies the maximum number of spam message that may be originated from a normal machine. The system observes the number of spam messages n originated from a machine in each window. If n > C*s*, then the algorithm identify that the machine as compromised. Similarly, in the PT detection algorithm, the time is partitioned into windows of fixed length T. PT monitors two e-mail sending properties of each internal machine in each time window: the percentage of spam messages sent by a machine and the total number of messages. Let n and N denote the spam messages and total messages originated from a machine m within any time window, respectively. If N >=Ca and n/N > P ,then PT declares machine m as being compromised, where Ca is the minimum number of messages machine must send and P is the user-defined maximum spam percentage of a normal machine.

V. PERFORMANCE EVALUATION

5.1 Performance of SPOT

In this section, performance of SPOT is evaluated based on the collected emails. In all the studies, set $\alpha = 0.01$, $\beta = 0.01$, $\theta 1 = 0.9$, and $\theta 0 = 0.2$. Assume that the deployed spam filter has a 90% detection rate and 20% false positive rate. SPOT depends on the spam messages to detect whether the machine has been compromised or not. Table 1 shows the performance of SPOT detection system.

Table 1: Spam Sending Machine Detail

Ip address	Total messages	No. of Ham	No. of spam	Machine status
127.0.0.1	46	34	12	Compromised
192.168.43.5	19	8	11	Compromised
192.168.209.181	3	0	3	Compromised
128.30.52.37	1	1	0	Not Compromised
128.30.52.45	1	1	0	Not Compromised
192.168.35.105	6	1	5	Compromised
192.168.43.142	4	1	3	Compromised

5.2 Performance of Count Threshold

Table 2 shows the performance of count threshold which include the machine IP addresses, spam count and machine status. The count threshold value is defined to 10. The machine status field is used to define, whether the machine is compromised or uncompromised.

Table 2: Normal spam's count for threshold

Ip address	Spam count	Machine status
127.0.0.1	12	Compromised
192.168.43.5	11	Compromised
192.168.209.181	2	Not Compromised
128.30.52.37	0	Not Compromised
128.30.52.45	0	Not Compromised
192.168.35.105	5	Not Compromised
192.168.43.142	2	Not Compromised

5.3 Performance of Percentage Threshold

Table 3 shows the performance of Percentage Threshold which includes the machine IP address, percentage of spam and the machine status fields. The machine IP address field denote the sender machine IP address. The percentage field shows percentage of spam messages sent by any machine.The machine status field is used to define, whether the machine is compromised or uncompromised, based on the performance.

Table 3: Normal spam percentage 40%

Ip address	percentage of spam	Machine status
127.0.0.1	26	Not Compromised
192.168.43.5	58	Compromised
192.168.209.181	100	Compromised
128.30.52.37	0	Not Compromised
128.30.52.45	0	Not Compromised
192.168.35.105	100	Compromised
192.168.43.142	66.67	Compromised

Both CT and PT require a fixed time window T and an appropriate user defined threshold values. In this evaluation, set T to 1 hour. For CT, set the threshold to 10 messages, which means a machine will be detected as a zombie if it sends more than 10 spam messages in any one hour window; for PT, set the threshold to 40%, which means a machine will be thought to be a zombie if it sends more than 40% spam messages in any one hour time window. The result shows CT detects 2 zombies out of 7 total IP addresses, which is 60% of what SPOT has detected and PT detects 4 zombies out of total 7 IP addresses, which is 80% of what SPOT has detected.

VI. RESULT

An effective and efficient system in automatically detecting compromised machines in the network is achieved successfully. The machine which is entering into the network will be observed by the SPOT. It will monitor the spam messages sent by the system. If the message exceeded the level, SPOT will do some process and decide that system as Spam Zombie. This detection is based on the outgoing messages. SPOT detection system can identify a compromised machine using minimum number of observations. SPOT is a lightweight process to identify the compromised machine.

VII. CONCLUSION

In this paper, we have discussed an effective spam zombies detection system called SPOT for detecting an compromised machine in a network. SPOT works well as compared to count threshold & percentage threshold algorithms. SPOT is designed based on the statistical tool Sequential Probability Ratio Test(SPRT).It also minimizes the number of required observations to detect a spam zombie.

REFERENCES

[1] Wald, Sequential Analysis. John Wiley & Sons, 1947.

[2] Y. Xie, F. Xu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, “Spamming Botnets: Signatures and Characteristics,” Proc. ACM. SIGCOMM,Aug. 2008.

[3] L. Zhuang, J. Dunagan, D.R. Simon, H.J. Wang, I. Osipkov, G. Hulten, and J.D. Tygar, “Characterizing Botnets from Email Spam Records,” Proc. First Usenix Workshop Large-Scale Exploits and Emergent Threats, Apr. 2008.

[4] M. Xie, H. Yin, and H. Wang, “An Effective Defense against Email Spam Laundering,” Proc. ACM Conf. Computer and Comm. Security, Oct./Nov. 2006.

[5] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee, “BotHunter: Detecting Malware Infection through Ids-Driven Dialog Correlation,” Proc. 16th USENIX Security Symp., Aug. 2007.

[6] G. Gu, J. Zhang, and W. Lee, “BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic,” Proc. 15th Ann. Network and Distributed System Security Symp. (NDSS ’08), Feb. 2008.

[7] G. Gu, R. Perdisci, J. Zhang, and W. Lee, “BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection,” Proc. 17th USENIX Security Symp., July 2008.