Abstract - One of the key security threats on the Internet is compromised machines. They create various security attacks like spamming and spreading malware, DDoS, and identity theft. Spamming assist with attackers to deploy large number of compromised machines, we focus on the identification of the compromised machines in a network that are involved in the spamming activities, commonly known as spam zombies. Using an effective spam zombie detection system named SPOT we can monitor outgoing messages of a network. SPOT is designed based on the statistical tool called Sequential Probability Ratio Test, which has bounded false positive and false negative error rates. SPOT is an effective and efficient system in automatically detecting compromised machines in a network. In addition, we also compare the performance of SPOT with other spam zombie detection algorithms based on the number and percentage of spam messages originated by internal machines, and show that SPOT outperforms these two detection algorithms.

Keywords - DDoS, Security, Zombies, Misbehavior, PRT, WMN

I. INTRODUCTION

A major security challenge on the Internet is the existence of the large number of compromised machines. Such machines have been increasingly used to launch various security attacks including spamming and spreading malware, DDoS, and identity theft. Two natures of the compromised machines on the Internet sheer volume and widespread render many existing security counter measures less effective and defending attacks involving compromised machines extremely hard. In this paper, we focus on the detection of the compromised machines in a network that are used for sending spam messages, which are commonly referred to as spam zombies. Given that spamming provides a critical economic incentive for the controllers of the compromised machines to recruit these machines, it has been widely observed that man compromised machines are involved in spamming.

Rather than the aggregate global characteristics of spamming botnets, we aim to develop a tool for system administrators to automatically detect the compromised machines in their networks in an online manner. The approaches developed in the previous work cannot be applied here. The locally generated outgoing messages in a network normally cannot provide the aggregate large-scale spam view required by these approaches. Moreover, these approaches cannot support the online detection requirement in the environment we consider. The nature of sequentially observing outgoing messages gives rise to the sequential detection problem. In this paper, we will
develop a spam zombie detection system, named SPOT, by monitoring outgoing messages. SPOT is designed based on a statistical method called Sequential Probability Ratio Test (SPRT). SPRT is a powerful statistical method that can be used to test between two hypotheses (in our case, a machine is compromised versus the machine is not compromised), as the events (in our case, outgoing messages) occur sequentially.

II. SYSTEM ANALYSIS

A. Existing System

A major security challenge on the Internet is the existence of the large number of compromised machines. Such machines have been increasingly used to launch various security attacks including spamming and spreading malware, DDoS, and identity theft. 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. On the other hand, identifying and cleaning compromised machines in a network remain a significant challenge on characterizing spamming botnets by leveraging both spam payload and spam server traffic properties. Our investigation is based on a novel framework called AutoRE that identifies botnet hosts by generating botnet spam signatures from emails.

AutoRE is motivated in part by the recent success of signature based worm and virus detection systems. These trends for evading existing detection systems suggest that we need to take a holistic view of various mechanisms and explore the invariable attack features in order to get an upper hand in the spam arms race.

B. Proposed System

The detection of the compromised machines in a network that are involved in the spamming activities, commonly known as spam zombies. We develop an effective spam zombie detection system named SPOT by monitoring outgoing messages of a network. SPOT is designed based on a powerful statistical tool called Sequential Probability Ratio Test, which has bounded false positive and false negative error rates. In addition, we also evaluate the performance of the developed SPOT system using a two-month e-mail trace collected in a large US campus network.

Our evaluation studies show that SPOT is an effective and efficient system in automatically detecting compromised machines in a network. For example, among the 440 internal IP addresses observed in the e-mail trace, SPOT identifies 132 of them as being associated with compromised machines. Out of the 132 IP addresses identified by SPOT, 126 can be either independently confirmed (110) or highly likely (16) to be compromised. Moreover, only seven internal IP addresses associated with compromised machines in the trace are missed by SPOT. In addition, we also compare the performance of SPOT with two other spam zombie detection algorithms based on the number and percentage of spam messages originated or forwarded by internal machines, respectively, and show that SPOT outperforms these two detection algorithms.

One based on the number of spam messages and another the percentage of spam messages sent from an internal machine, respectively. For simplicity, we refer to them as the count-threshold (CT) detection algorithm and the percentage-threshold (PT) detection algorithm, respectively. Advantages are

1. An effective and efficient system in automatically detecting compromised machines in a network.
2. Operation workload very low because using the CT and PT techniques.
3. It effectively identifies any machine sending a single spam message as a compromised machine if a machine sent one outgoing message carrying a virus/worm attachment.

III. ALGORITHMS AND MODULES DESCRIPTION

A. Network and Topology Invention

Network topology is the arrangement of the various elements such as links and nodes of a computer. Essentially, it is the topological structure of a network, and may be depicted physically or logically. Physical topology refers to the placement of the network's various components, including device location and cable installation, while logical topology shows how data flows within a network, regardless of its physical design. Distances between nodes, physical interconnections, transmission rates, and/or signal types may differ between two networks, yet their topologies may be identical. We assume that messages originated from machines inside the network will pass the deployed spam zombie detection system. Receiving the node information and IP address through the network administrator. After configuring the topology and finally it is innovated.

B. Spam Zombies Detection Machine Creation

The spam zombie detection problem in a network, the messages originated from machines inside the network will pass the deployed spam zombie detection system. This assumption can be achieved in a few different scenarios. For example, the outgoing e-mail traffic (with destination port number of 25) can be replicated and redirected to the spam zombie detection system. A machine in the network is assumed to be either compromised or normal (that is, not compromised).

We only focus on the compromised machines that are involved in spamming. Therefore, we use the term a compromised machine to denote a spam zombie, and use the two terms interchangeably. Let $X_i$ for $i = 1, 2, \ldots$. Denote the successive observations of a random variable $X$ corresponding to the sequence of messages originated from machine $m$ inside the network. We let $X_i = 1$ if message $i$ from the machine is a spam, and $X_i = 0$ 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 they send. Specifically, a compromised machine will with a higher probability generate a spam message than a normal machine.

C. Spot Detection and Analysis

SPRT can be considered as a one dimensional random walk with two user-specified boundaries corresponding to the two hypotheses. As the samples of the concerned random variable arrive sequentially, the walk moves either upward or downward one step, depending on the value of the observed sample. When the walk hits or crosses either of the boundaries for the first
time, the walk terminates and the corresponding hypothesis is selected. As a simple and powerful statistical tool, SPRT has a number of compelling and desirable features that lead to the widespread applications of the technique in many areas. First, both the actual false positive and false negative probabilities of SPRT can be bounded by the user-specified error rates. A smaller error rate tends to require a larger number of observations before SPRT terminates. Thus, users can balance the performance (in terms of false positive and false negative rates) and cost (in terms of number of required observations) of an SPRT test. Second, it has been proved that SPRT minimizes the average number of the required observations for reaching a decision for a given error rate, among all sequential and non-sequential statistical tests. In the following, we present the formal definition and a number of important properties of SPRT.

D. Spam Count Based Detection Analysis

The time is partitioned into windows of fixed length $T$. A user-defined threshold parameter $C_s$ specifies the maximum number of spam messages that may be originated from a normal machine in any time window. The system monitors the number of spam messages $n$ originated from a machine in each window. If $n > C_s$, then the algorithm declares that the machine has been compromised. They require a thorough understanding of the different behaviors of the compromised and normal machines in the concerned network and a training based on the behavioral history of the two different types of machines in order for them to work reasonably well in the network.

For CT, we set the maximum number of spam messages that a normal machine can send within a time window to be 30 ($C_s = 3$), that is, when a machine sends more than 30 spam messages within any time windows, CT concludes that the machine is compromised.

E. Spam Percentage Based Detection Analysis

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: one is the percentage of spam messages sent from a machine, another the total number of messages. Let $N$ and $n$ denote the total messages and spam messages originated from a machine $m$ within a time window, respectively, then PT declares machine $m$ as being compromised if $N < C_a$ and $N > P$, where $C_a$ is the minimum number of messages that a machine must send, and $P$ is the user-defined maximum spam percentage of a normal machine. The first condition is in place for preventing high false positive rates when a machine only generates a small number of messages. For example, in an extreme case, a machine may only send a single message and it is a spam, which renders the machine to have a 100 percent spam ratio. However, it does not make sense to classify this machine as being compromised based on this small number of messages generated.

F. Evaluation of Dynamic IP Address

The spam zombie detection algorithms, we have for simplicity ignored the potential impact of dynamic IP addresses and assumed that an observed IP corresponds to a unique machine. In the Following, we informally discuss how well the three algorithms fair with dynamic IP addresses. We formally evaluate the impacts of dynamic IP addresses on detecting spam zombies in the next section using a two-month e-mail trace collected on a large US campus network. Dynamic IP addresses can have a greater impact on the other two detection algorithms CT and PT. First, both require the continuous monitoring of the sending behavior of a machine for at least a specified time window, which in practice can be on the order of hours or days. Second, CT also requires a relatively larger number of spam messages to be observed.
from a machine before reaching detection. By properly selecting the values for the parameters of CT and PT (for example, a shorter time window for machines with dynamic IP addresses), they can also work reasonably well in the environment of dynamic IP addresses.

G. Spammer activities blocking

SPOT identifies the spam message, analysing the IP address of the sender machine. After analyzing the IP address, messages are sent to the Network Administrator in order to check the activities such as whether it exceeds the user defined threshold value. If the resultant activity seems to be an abnormal condition, it is then blocked. Otherwise, the process continues.

IV. CONCLUSION

We developed an effective spam zombie detection system named SPOT by monitoring outgoing messages in a network. SPOT was designed based on a simple and powerful statistical tool named Sequential Probability Ratio Test to detect the compromised machines that are involved in the spamming activities. SPOT has bounded false positive and false negative error rates. It also minimizes the number of required observations to detect a spam zombie. Our evaluation studies based on a two-month e-mail trace collected on the FSU campus network showed that SPOT is an effective and efficient system in automatically detecting compromised machines in a network. In addition, we also showed that SPOT outperforms two other detection algorithms based on the number and percentage of spam messages sent by an internal machine, respectively.

REFERENCES