Analyzing Spam Messages And Detecting Zombies

V.ANNIE
II YEAR M.E (CSE)
Department of computer science & engineering
Mount Zion college of Engineering & technology
Tamilnadu, India

Abstract—In this paper we focus on the detection of the compromised machines in a network that are involved in the spamming activities, commonly known as spam zombies. Here applied our technique against a Detecting of spam email from Hotmail Web mail services. In this Detecting method, successfully identified hundreds of Zombies. These present new findings about Zombie attacks and behavior while also confirming other researcher's observations derived by different methods. In this paper, Network administrator analyzes Spam messages by using the heuristic search, and reported to the end user. After checking the activities along with IP address, If the resultant activity is seems to be abnormal, then it is said to be compromised machine then it should undergo clean process. While checking activities, it is not the non spam message, and then it is normal machine.

Index term-Zombies, spam, botnet, heuristic

1. 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. As a side-product of free email services, spam has become a serious problem that afflicts every Internet user in recent years.

In this paper, 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 many compromised machines are involved in spamming.

Although a number of anti-spam mechanisms have been proposed and deployed to foil spammers, spam messages continue swarming into Internet users' mailboxes. A more effective spam detection and suppression mechanism close to spam sources is critical to dampen the dramatically-grown spam volume. The IP address of a spammer is obfuscated by a spam proxy during the protocol transformation, which hinders the tracking of real spam origins.

2. RELATED WORK

In this section, we present *AutoRE* – a framework for *automatically* generating URL

signatures to identify Zombies-based spam campaigns. As input, AutoRE takes only a set of unlabeled email messages (messages are not tagged as spam/non-spam), and pro duces two outputs: a set of *spam URL signatures*, and a related list of *Zombies host IP addresses*.

Zombies, or more properly Unsolicited Commercial E-mail (UCE), are an increasing threat to the viability of Internet E-mail and a danger to Internet commerce. UCE senders take away resources from users and service suppliers without compensation and without authorization. A variety of counter-measures to UCE have been proposed, from technical to regulatory.

Among the technical ones, the use of filtering methods is popular and effective. UCE filtering is a text categorization task. Text categorization (TC) is the classification of documents with respect to a set of one or more pre-existing categories. In the case of UCE, the task is to classify e-mail messages or newsgroups articles as UCE or not (that is, legitimate). The general model of TC makes use of a set of pre-classified documents to classify new ones, according to the text content (i.e. words) of the documents

In this paper, the adoption of polymorphic URLs increased significantly, and the number of static IP address based bots doubled from Nov 2006 to July 2012. These trends for evading existing detection systems suggests that 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 to provide improve zombie detecting technique need to developed

3. OVERVIEW

Here first study the quality of the extracted URL signatures. Here used the human classified labels to compute the spam detection false positive rate. To better understand the effectiveness of using signatures for future spam detection, we performed cross-month evaluation by applying signatures generated in a previous month to emails received in a later month. These experiments also demonstrated the importance of having regular expression signatures.

Second, Here examined whether the identified Zombie hosts were indeed spamming servers – to this end, we used the Hotmail server log that records the sending history of *all* email servers that communicate with Hotmail over time. This log includes the email volume and the spam ratio 4 of each server on a daily basis. In this paper, these statistics to evaluate the identified Zombie hosts.

The spam ratio was computed using the existing spam filtering system configured. The current filter leverages both email content and email server sending history for spam detection. Finally, They are interested in finding whether each set of emails identified from the same spam campaign were correctly grouped together. To answer this question, for every set, we examine the similarity between the corresponding destination Web pages. In this paper destination web pages were shown to be strongly correlated to the corresponding spam campaign.

4. TECHNIQUE USED BY ZOMBIE

Zombie/Spammers use various techniques to send large volumes of mail while attempting to remain untraceable. Here describe several of these techniques, beginning with .conventional. methods and progressing to more intricate techniques.

4.1 Direct spamming.

Spammers may purchase upstream connectivity from .spam-friendly ISPs., which turn a blind eye to the activity. Occasionally, spammers buy connectivity and send spam from ISPs that do not condone this activity and are forced to change ISPs. Ordinarily, changing from one ISP to another would require a spammer to renumber the IP addresses of their mail relays IP address spoofed to appear as if it came from the dialup connection, and proxy the reverse traffic through the dialup connection back to the spamming hosts .

4.2 Open relays and proxies.

Open relays are mail servers that allow unauthenticated Internet hosts to connect and relay email through them. Originally intended for user convenience (*e.g.*, to let users send mail from a particular relay while they are traveling or otherwise in a different network), It appears that the widespread deployment and use of blacklisting techniques have all but extinguished the use of open relays and proxies to send spam .

4.3 Zombie/bot.

Conventional wisdom suggests that the majority of spam on the Internet today is sent by Zombie. Collections of machines acting under one centralized controller . The W32/Bobax (.Bobax.) worm (of which there are many variants) exploits the DCOM and LSASS vulnerabilities on Windows systems , allows infected hosts to be used as a mail relay, and attempts to spread itself to other machines affected by the above vulnerabilities, as well as over email. This studies the network level properties of spam sent by Bobax drones. Agobot and SDBot are two other bots purported to send spam .

4.5 BGP spectrum agility.

This study has discovered a new type of cloaking mechanism.BGP .spectrum agility..where by spammers briefly announce (often hijacked) IP address space from which they send spam and the routes to that IP address space once the spam has been sent. our study thoroughly documents this activity, and further finds that spammers may be using spectrum agility to complement spamming by other methods.

5. SPAM ZOMBIE DETECTION ALGORITHMS

In this section, develop spam zombie detection algorithms. The first one is SPOT, which utilizes the Sequential Probability Ratio Test presented in the last section. Here discuss the impacts of SPRT parameters on SPOT in the context of spam zombie detection. The other two spam zombie detection algorithms are developed based on the number of spam messages and the percentage of spam messages sent from an internal machine, respectively.

5.1 SPOT Detection Algorithm

SPOT is designed based on the statistical tool SPRT we discussed in the last section. In the context of detecting spam zombies in SPOT, we consider H1 as a detection and H0 as a normality that in the context of spam zombie detection, from the viewpoint of network monitoring, it is more important to identify the machines that have been compromised than the machines that are normal.

After a machine is identified as being compromised it is added into the list of potentially compromised machines that system administrators can go after to clean. The message-sending behavior of the machine is also recorded should further analysis be required.

Before the machine is cleaned and removed from the list, the SPOT detection system does not need to further monitor the message sending behavior of the machine. On the other hand, a machine that is currently normal may get compromised at a later time. Therefore need to continuously monitor machines that are determined to be normal by SPOT. Once such a machine is identified by SPOT, the records of the machine in SPOT are reset, in particular, monitoring phase starts for the machine.

6. PROPOSED ALGORITHM:

6.1. Heuristic Algorithm:

The features that are rare in normal messages but appear frequently in spam, such as non- existing domain names and spam-related keywords, can be used to distinguish spam from normal email. Spam Assassin is such an example. Each received message is verified against the heuristic filtering rules. Compared with a pre- defined threshold, the verification result decides whether the message is spam or not.

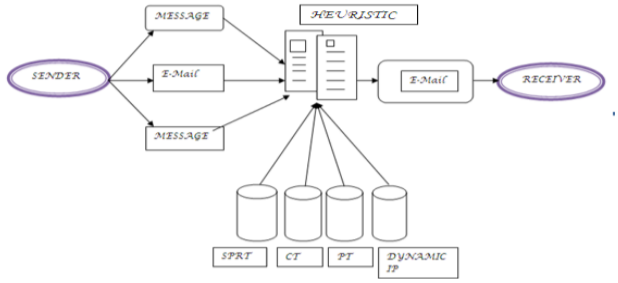

FIG 1.1 HEURISTIC ALGORITHM

6.2. Heuristics for ZOMBIE:

In the algorithm, Here proposed a set of heuristic features to complement the word Bayesian model in their work, including: a set of around 35 hand-crafted key phrases (like "free money"); some non text features (like the domain of the sender, or whether the message comes from a distribution list or not); and features concerning the non alphanumeric characters in the messages.

For this work, we have focused in this last set of features. The test collection used in our experiments, Spam base, already contained a set of nine heuristic features. Spam base is an e-mail messages collection containing 4601 messages, being 1813 (39%) marked as ZOMBIE. The collection comes in preprocessed (not raw) form, and its instances have been represented as 58-dimensional vectors. The first 48 features are words extracted from the original messages, without stop list nor stemming, and selected as

the most unbalanced words for the ZOMBIE class. The next 6 features are the percentage of occurrences of the special characters ";', "(", "[", "!", "$" and "#". The following 3 features represent different measures of occurrences of capital letters in the text of the messages.

Finally, the last feature is the class label. So, features 49 to 57 represent heuristic attributes of the messages. In our experiments, we have tested several learning algorithms on three feature sets: only 1 This collection can be words, only heuristic attributes, and both. We have divided the Spam base collection in two parts: a 90% of the instances axe used for training, and a 10% of the messages are retained for testing. This split has been performed preserving the percentages of legitimate and ZOMBIE messages in the whole collection.

6.3. Impact of Dynamic IP Addresses

In the above discussion of 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 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.

Heuristic can work extremely well in the environment of dynamic IP addresses. To understand the reason we note that Heuristic can reach a decision with a small number of observations which shows the average number of observations required for SPRT to terminate with a conclusion. In practice, we have noted that spam messages will be sent before the (unwitting) user shutdowns the machine and the corresponding IP address gets reassigned to a different machine. Therefore, dynamic IP addresses will not have any significant impact on Heuristic. Dynamic IP addresses can have a greater impact continuous monitoring of the sending behavior of a machine for at least a specified time window, (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.

7. PERFORMANCE EVALUATION:

In order to understand the performance of HEURISTIC in terms of the false positive and false negative rates, we rely on a number of ways to verify if a machine is indeed compromised. First, we check if any say we have a confirmation. Out of the 132 IP addresses identified by HEURISTIC, we can confirm 110 of them to be compromised in this way. For the remaining 22 IP addresses, we manually examine the spam sending patterns from the IP addresses and the domain names of the corresponding machines. If the fraction of the spam messages from an IP address is high (greater than 98 percent), we also claim that the corresponding machine has been confirmed to be compromised. We can confirm 16 of them to be compromised in this way. We note that the majority (62.5 percent) of the IP addresses

confirmed by the spam percentage are dynamic IP addresses, which further indicates the likelihood of the machines to be compromised.

All the compromised machines are detected with no more than 11 observations. This indicates that, HEURISTIC can quickly detect the compromised machines. We note that HEURISTIC does not need compromised machines to send spam messages at a high rate in order to detect them. Here, "quick" detection does not mean a short duration, but rather a small number of observations. A compromised machine can send spam messages at a low rate (which, though, works against the interest of spammers), but it can still be detected once enough observations are obtained by HEURISTIC.

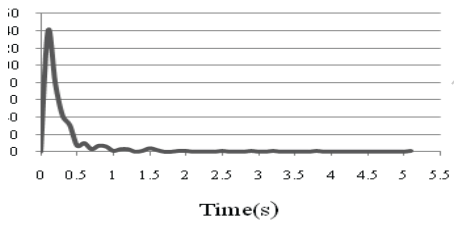

fig 7.1. heuristic detect spam zombies in quick time

8. CONCLUSION AND FUTURE WORK

In this paper, developed an effective spam zombie detection system named HEURISTIC by monitoring outgoing messages in a network. In future experiments, make plan to apply the uniform method Medicos to the algorithms tested in this work, for getting more comparable results. With respect to the use of heuristics, it can see that this information alone is not competitive, but it can improve classification based on words. The improvement shown in our experiments is modest, due to the heuristics used. Here are not able to add other heuristics in this case because the Spam base collection comes in a preprocessed fashion. For future experiments, they will use the collection from which is in raw form. This fact will enable us to search for more powerful heuristics.

In addition, these also showed that HEURISTIC outperforms two other detection algorithms based on the number and percentage of spam messages sent by an internal machine, respectively.

REFERENCES

[1] S. Arora, P. Raghavan, and S. Rao, "Approximation Schemes for Euclidean k-Medians and Related Problems," Proc. 30th ACM Symp. Theory of Computing (STOC), 1998.

[2] M. Baker, R. Buyya, and D. Laforenza, "Grids and Grid Technology for Wide-Area Distributed Computing," Software- Practice and Experience, 2002.

[3] A. Chervenak, E. Deelman, I. Foster, L. Guy, W. Hoschek, C. Kesselman, P. Kunszt, M. Ripeanu, B. Schwartzkopf, H. Stockinger, and B. Tierney, "Giggle: A Framework for Constructing Scalable Replica Location