International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Spam and Zombie Detection System with Machine Learned Spot Algorithm

Manishankar .S[1], Sobin E. S[2]

[1]Assistant Professor, Amrita Vishwa Vidhyapeetham, Mysore Campus, #114, 7th Cross Bogadi 2nd Stage

[2]Final Year, M.Sc. CS, Amrita Vishwa Vidhyapeetham, Mysore Campus, #114, 7th Cross Bogadi 2nd Stage

Abstract: ***Email world has grown so much that there is a wide increase in the number of email that fells in to the category of Spam and Zombie attack. There is always a need for trained system that predicts spam and Zombie attack, this paper deals with a novel approach of machine learning to build a tool which predicts an email spam or not with the help of SPOT detection with SPERT algorithm, paper also deals with Zombie attacks and DDOS attacks***

Keywords: SPAM ,ZOMBIE ,DDOS,SPOT ,SPERT,BOTNET.Learning Algorithm

1. Introduction

Major challenge that arises in the current internet Security world is the existence of the numerous systems which get affected by zombie attack or malicious spam .Such machines have been increasingly used to launch various security attacks including spamming and spreading malware, DDoS, and identity theft [1]. Mainly this type of attacks is targeting a particular object in web. They will create uncontrolled traffic and they will easily penetrate or destroy the load capacity of the target application. If we block a compromised pc in the entry level of a network or blocking a spread of zombie in mail server is ideal. Here this paper we are trying to implement a concept to protect the network from within the network itself. Exactly mean that healing from the beginning itself. by using some statistical techniques. We try detecting the spam speeding compromised system within a limited observation. In this paper we are trying devolving a system (SPOT) that will monitor all outgoing mail from a network (mainly the port no:25) .We use the Sequential probability ratio testing to train the machine to learn to detect a Spam Zombie.

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. On the other hand, identifying and cleaning compromised machines in a network remain a significant challenge for system administrators of networks of all sizes. The paper focuses 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. A number of recent research efforts have studied the aggregate global characteristics of spamming botnets [7] (networks of compromised machines involved in spamming) such as the size of botnets and the spamming patterns of botnets, based on the sampled spam messages received at a large e-mail service provider 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. We consider ourselves situated in a network and ask the following question: How can we automatically identify the compromised machines in the network as outgoing messages pass the monitoring point sequentially? 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 [4], by monitoring outgoing messages. SPOT is designed based on a statistical method called Sequential Probability Ratio Test (SPRT), developed by Wald in his seminal work. 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. This means that the SPOT detection system can identify a compromised machine quickly. Moreover, both the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds. Consequently, users of the SPOT system can select the desired thresholds to control the false positive and false negative rates of the system. In this paper, we develop the SPOT detection system to assist system administrators in automatically identifying the compromised m our evaluation studies spot is effective.

2. Existing Works on the Field

Existing approach mainly depend upon two area One is effectively detecting the spam mails from the outgoing mails network/system. Commonly the Botnet attacks are showing some common characteristics. These studies provided important insights into the 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 [5]. The common approaches to the botnet attacks are quite different they are not consider about healing from a single network

they only consider about whole network. They try to only in detect the spamming activity at only in the receiving side. However, their approaches are better suited for large email service providers to understand the aggregate global characteristics. In this paper we mainly concentrate to make an administrative tool for detecting the compromised machine in that network SPOT is a light-weight spam zombie detection system; it does not need the support from the network intrusion detection system as required by Bot Hunter [6] As a simple and powerful statistical method, Sequential Probability Ratio Test (SPRT) [2] has been successfully applied in many areas. In the area of networking security, SPRT has been used to detect port scan activities proxy-based spamming activities and MAC protocol misbehavior in wireless networks

3. Problem Formulation

Problem formulation deals with training the machine in the network based on unsupervised machine learning to plot the different cluster among the email is zombie spam or not, here the sequential probability testing plays a vital role in classifying the test data in to zombie spam or not and then training the system to clearly identify the zombie spam [9]. Here we have a trained set of data and the emails which were successfully classified by spot using the sequential probability testing In this section we formulate the spam zombie detection problem in a network. In particular, we discuss the network model and assumptions we make in the detection problem. giving the logical view of network this network that consists number of pc may be compromised or not .We assume that the messages are will be originated from this network. Each outgoing messages are passed through the spot detection system. In this outgoing window only we will capture all the email id that coming from that particular port no (For the testing we took port no 25) Outgoing email traffic (with destination port number of 25) from all other machines in the network is blocked by edge routers of the network. In this situation this application we can embed with the local smtp server .In the local smtp server that will capture the mail from that network. By using the port name detection approach we can avoid the some behavior zombie attackers. see in fig 1 such as they does not follow a specific path or application to attack then .By using this technique we can capture the email from any application. [10].

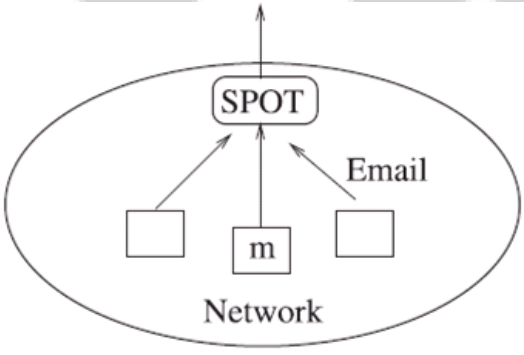

Figure 1: SPOT system Design

Term a *compromised machine* to denote a *spam zombie*, comprised machine learns to identify the spam zombie or not and use the two terms interchangeably. Let Xi 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 $Xi = 1$ if message i from the machine is a spam, and $Xi = 0$ otherwise. The main part of the algorithm (SPRT). We assume the probability to producing a spam message from compromised pc is more than a normal pc.

$$Pr\,(Xi = 1|H1) > Pr(Xi = 1|H0), \qquad (1)$$

H1 is denoted as compromised system and H0 is normal pc. By using this classification system can classified as the incoming Message characteristics. By using the above formula we need not observe the sequence of message stream .Because in a within a limited observation we can predict system is compromised system or not

4. SPOT Detection and SPERT Algorithm Overview

SPRT is a statistical method for testing a simple null hypothesis against a single alternative hypothesis. Intuitively, SPRT can be considered as a one dimensional random walk two user-specified boundaries corresponding to the two hypotheses. In this Algorithm we will specify a boundary or a threshold value .before reaching the threshold value it will filter each incoming messages. We can use any type spam filtering technique in this paper. we used Bayesian classification for finding the spam mails. If the incoming mail is spam we add this mail and count to the database it will count the number of email that coming from the particular system and the number of spam mails from that system. By this counts are using for input values for our algorithm,

Algorithm 1: SPOT spam zombie detection system

Input Values: M,T,S
Output: False or True
1: An outgoing message arrives at SPO
2: Get system name of sending machine *m*
3: Let *n* be the message index, and spam count S
4:Threshold value T
5 :Let $Xn = 1$ if message is spam, $Xn = 0$ otherwise
6: if ($Xn == 1$) then
7: n++,s++;
8: S=S/n;
End if
9: If(T>S) Then
Machine *m* is compromised. Test terminates for *m* , True
10: Else
Test continues with an additional observation
False

The algorithm here helps to train the machine more than a regression testing or supervised learning machine automatically learns the to do the sequential probability testing that is carried forward with in the zombie detection system.

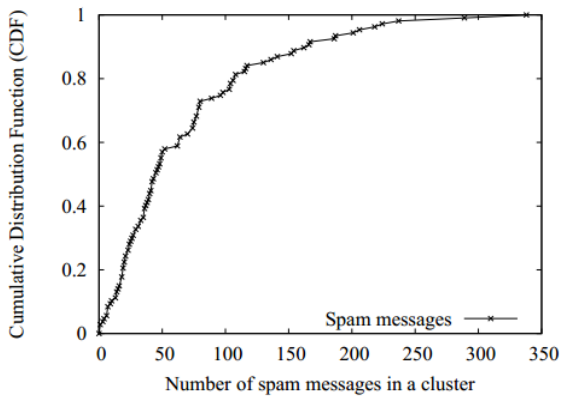

Figure 2: SPAM Detection Ratio Graph

In this algorithm have four input values T is the threshold value that can specify by the administrator if the calculated value is more than threshold then the algorithm return the true value s is the total number of count that that particular system sent before. N is the total number of messages. Fig 2 represents the That sent by that particular system(m). if the incoming message is spam we assign the one to the Xi else it zero. After reaching the target or assigned observation the algorithm that return a probability ratio value .By analyzing the value we can detect the systems are being infected

5. Working model of Proposed System

Architecture and working proposed system has following which help to train the machine by collecting the data. system has a spot detection circuit which helps to classify the mail in to a Zombie or not which is a machine learning environment which is unsupervised .each mail that is outgoing from the mail server on which the tool is deployed is being captured .Each mail is being cross checked by the spam detection system .spam findings are recorded in to a database for the use of training the statistical machine in the mail server .then mail count is being retrieved to find out the count of the number of spam in fact present in the total number of outgoing email ,then calculation of value is being performed which help to analyze the threshold value ,the trained system with the help of SPERT and SPOT detection compares the email with the threshold value and the trained value of the system to predict it as a spam zombie or not. comparison is performed and the machine is compromised with the learning process.

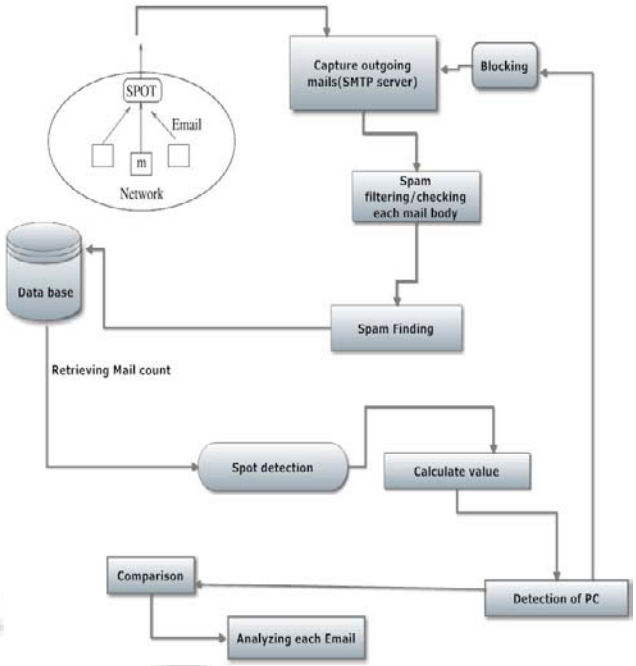

Figure 3: Working model of SPAM and zombie detection system

6. Conclusion and Future Recommendations

In this paper we developed machine learning approach which helps in an effective spam zombie detection system 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 subset of compromised machines that are involved in the spamming activities. Our process was to design an effective learning environment which tracks Spam zombie efficiently with high accuracy when worked on set of spam data successfully trained to become a efficient spam detection mechanism

7. Acknowledgment

We are grateful to almighty and all those who have been a part of making this approach possible in to a real time spam detection system. we hereby an opportunity to thank all the authors of the white papers which are a guide to the research we have carried forward.

References

[1] J. Markoff. Russian gang hijacking pcs in vast scheme.The New York Times, http://www.nytimes.com/2008/08/06/technology/06hack .html, August 2008. 1

[2] Y. Xie, F. Xu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets:signatures and characteristics. In Proc. ACM SIGCOMM, Seattle, WA, August 2008.1, 2

[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. In Proc. of 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, San Francisco, CA, April 2008. 1, 2

[4] A. Wald. Sequential Analysis. John Wiley & Sons, Inc, 1947. 1, 4, 4
[5] M. Xie, H. Yin, and H. Wang. An effective defense against email spam laundering. In ACM Conference on Computer and Communications Security, Alexandria, VA, October 2006. 2
[6] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee. Bothunter: Detecting malware infection through ids-driven dialog correlation. In Proc. 16th USENIX Security Symposium, Boston, MA, Auguest 2007. 2
[7] G. B. Wetherill and K. D. Glazebrook. Sequential Methods in Statistics. Chapman and Hall, 1986. 2
[8] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, May 2004. 2
[9] S. Radosavac, J. S. Baras, and I. Koutsopoulos. A framework for mac protocol misbehavior detection in wireless networks. In Proceedings of the 4th ACM workshop on Wireless security, Cologne, Germany, September 2005. 2
[10] S. Linford. Increasing spam threat from proxy hijacking.http://www.spamhaus.org/news.lasso?article=156.

Author Profile

Manishankar. S is a Research scholar and an Assistant professor in Amrita University. Has published many international and national conference and journal papers, completed M. Tech in Computer Science, guided many live projects and research projects ,and has also an vivid industrial experience ,specialized in cloud computing ,Big data ,machine Learning

Sobin E S is a student of Amrita Vishwa Vidhyapeetham Mysore campus, completed his MSc in Computer Science, he is a well versed with research around network security and machine learning concepts.