Detecting Targeted Malicious Email

Lakshmi.H.Thampi [1], Vaishnavi.A [2], G.Shaalini [3], T.Rahul [4]
[1, 2, 3, 4] UGC Scholar, Department of Computer Science, SRM Institute of Science & Technology, Chennai, Tamil Nadu, India.

Abstract – Recent studies have shown that TME(Targeted Malicious Emails) have been exploiting computer networks and is causing a lot of damage . Beyond spam or phishing designed to trick users into revealing personal information, TME can exploit computer networks and gather sensitive information. They can consist of coordinated and persistent campaigns that can span years. This paper, how to detect a targeted malicious packet (email) for normal network into modern network. A compromised router detection protocol that dynamically infers the precise number of congestive packet losses that will occur. Developing an alternative filtering procedure by using TME specific feature extraction. These protocols automatically predict congestion in a systematic manner.

Index Terms – Non Targeted Malicious Mails, Targeted Malicious Mail, Naïve Bayesian Classifier, Random Forest Method

1. INTRODUCTION

Nowadays email has made a deep impact in the society as most of the research efforts have been made for making email technology more convenient, intuitive to use and costing virtually nothing. Thus, an email system has become an important and essential communication approach for millions of people since one can conveniently transfer messages electronically to anyone within seconds at visibly zero cost. In order to use email, one has to use a mail client to access the mail server. The mail client and mail server use a variety of protocols for exchanging information with each other .The users can access email in several ways, but most popular ones are Post Office Protocol (POP), Interactive Mail Access Protocol (IMAP) and Webmail. POP is designed to support offline mail processing. With POP protocol, messages are delivered to the mailboxes and users can access their mailboxes and download messages from the mail server to their computers by using mail client programs. Once the messages are delivered to the computer the messages are deleted from the mail server. IMAP is more complex and recent development which is designed for the users to stay connected to one or more email servers while reading, creating and organizing messages. With IMAP, the mails can be accessed by connecting to the servers only. The mails cannot be viewed when one is offline. Webmail offers complete access to one's email without any email being downloaded to one's computer. The users of email face various difficulties due to the attacks which may destroy the whole system. According to the statistics around 90% of email messages are spam. Spam is not only irritating and nuisance; it is also a persistent problem which can cause significant harm negatively affecting the internet users and administrators. It has also increasingly become extremely dangerous as 83% of spam contains a URL so phishing sites and Trojan infections are just one click away. Email spam is not only wastage of time but it also consumes storage on the server and blocks communication channels until the recipient takes some action on it. Also there is a chance of deletion of an important email while deleting spam emails. Spam email is also a great malware carrier in order to infect computers with viruses. TME on the other hand is more dangerous than spam and phishing. Spam and phishing is easy to detect as it is mass generated sent to millions of people. It is possible to gather mails with similar characteristics and message content probably for identifying spam. But TME is designed to target a single individual and is difficult to detect. So, we develop an alternative filtering procedure by using TME specific feature extraction.

2. RELATED WORK

In the existing system, the detection methods used were Spam Assassin and ClamAV. They used distributed protocols to detect such traffic manipulations, typically by validating that traffic transmitted by one router is received unmodified by another. Transmission Control Protocol (TCP) is designed to cause such losses as part of its normal congestion control behaviour. The attacker may subvert the network control plane (e.g., by manipulating the routing protocol into false route. The problems faced by the existing system is that traffic manipulation occurred in the router due to the usage of distributed protocols. The failure in the existing method also occurred by using TCP .The static threshold mechanism in the existing system is inadequate as it allows multiple vigorous attacks to take place without being detected.

3. PORPOSED MODELLING

The main problem in the current scenario is the attack on the mails. Sometimes this may lead to the destruction of the entire system. The main aim is to acknowledge TME and to inform it to the user. We develop a compromised router detection protocol that identifies congestive packet losses be. To identify TME a special feature extraction algorithm is proposed in this paper. A simplified view of our classification consists of pre-processing the mail for leveraging company

information. Persistent threat and recipient oriented features are extracted and the associated mails are classified using Random Forest classifier. In this paper, we also propose Naive Bayesian classification for classifying the mails.

A. SYSTEM ARCHITECTURE

The detection of TME is done by using Naive Bayesian Classification features. From the below figure 1 the user must login using mail id and password. During the training, a model is built based on the characteristics of each category in a pre-classified set of e-mail messages. The training dataset should be selected in such a way that it is varying in content and subject. Each sample message is labelled with a specific category. We first perform pre-processing to extract tokens and determine the number of occurrences of each token in each category. Spam filtering is based on calculating the fuzzy similarity measure between the received message and each category i.e. spam and legitimate.. The token with the maximum number of occurrences is assigned with a value of 1, and all other tokens are assigned with proportional values. The mails are then classified using Naive Bayesian classification which detects the mails with highest probability of spam

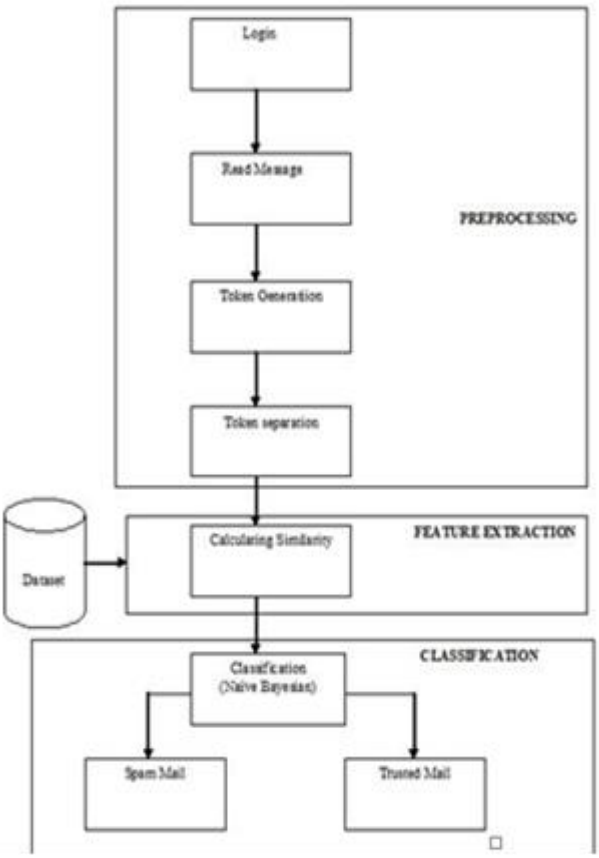

Figure 1.Basic system architecture

B. PREPROCESSING AND FEATURE EXTRACTION

Stemming Algorithm: Stemming is reducing the word to the root form, where lemmatization is concerned with linguistics. Lemmatization is ”go”, “gone”, “goes”, “going”, “been” and ”went”, where stemming a word would be reducing a word from "gone" to "go", so it can be matched to other stemmed words such as "going", as "going" stemmed would also be "go".

A better example is: "engineering", "engineers", "engineered", "engineer". These four words would not match up if they were tested for equality, however by stemming these words we can reduce them to a more basic form,

engineering --> engineer

engineers --> engineer

engineered --> engineer

engineer --> engineer.

Now the stemmed words will match for equality. So, now if we try searching using the word engineer, documents on engineering, engineers and engineered would be returned from a stemmed index/database .Stemming usually means to cut off characters from the end of the word, e.g. walked -> walk, walking -> walk. However, this does not necessarily produce a real word e.g. a stemmer could also change house and houses to "hous". Also, cutting of characters isn't enough for irregular words, e.g. you cannot get from "went" to "go" by just cutting of characters. A lemmatizer solves these problems, i.e. it always produces real words, even for irregular forms. It usually needs a table of irregular forms for this.

Random Forest Algorithm: Random forests is a learning method in which we make use of large number of trees for classification, regression etc. .It helps in finding the mean prediction of the individual trees. A choice tree is a k-exhibit tree in which every inside hub indicates a test on a few qualities from data list of capabilities speaking to information. Every branch from a hub relates to conceivable feature values determined at that hub. Furthermore, every test results in branches, speaking to fluctuated test results. The choice tree prompting fundamental calculation is a ravenous calculation developing choice trees in a top down recursive partition and-vanquish way. The calculation starts with tuples in the preparation set, selecting best quality yielding greatest data for classification. It produces a test hub for this and after that a top down choice trees affectation partitions current tuples set by test trait values. Classifier era stops when all subset tuples fit in with the same class or on the off chance that it is not qualified to continue with extra partition to further subsets, i.e. on the off chance that more quality tests yield data for classification alone underneath a pre specified. In the proposed feature selection a Decision tree impelling chooses significant features. Choice tree actuation is the learning of choice tree classifiers building tree structure where every inside hub (no leaf hub) signifies quality test. Every branch speaks to test result and every outside hub (leaf hub) signifies class forecast. At each hub, the calculation chooses best segment information credit to

individual classes. The best credit to apportioning is chosen by characteristic selection with Information pick up. Trait with most astounding data increase parts the characteristic. Data addition of the trait is found by

1. $info(D)=-\sum p_i \log_2 (p)$

Where p_i is the probability that arbitrary vector in D belongs to class c_i. A log function to base 2 is used, as information is encoded in bits. Info (D) is just average information amount required to identify vector D class label. The information gain is used to rank the features and the ranked features are treated as features in hierarchical clusters. The proposed Manhattan distance for n number of clusters is given as follows:

2. $MDist= -\sum (a_i-b_i)$

A cubic polynomial comparison is determined utilizing the Manhattan values and the limit rule is resolved from the incline of the polynomial mathematical statement. The features are thought to be superfluous for arranging if the slant is zero or negative and pertinent when the incline is sure. Pseudo code for the random forest algorithm:

To generate c classifiers:

For i = 1 to c do

Randomly sample the training data D with replacement

to produce D_i

Create a root node, N_i

containing i D

Call BuildTree(N_i)

end for

BuildTree(N):

if N contains instances of only one class then

return

else

Randomly select x% of the possible splitting features

in N

Select the feature F with the highest information gain

to split on

Create f child nodes of N ,N_1,..., N_f, where F has f

possible values (F_1, … ,F_f)

for i = 1 to f do

Set the contents of N_i to D_i

, where D_i

is all instances in

N that match

F_i

Call BuildTree(N_i)

end for

end if

C. CLASSIFICATION

Naïve Bayesian Classification: Naïve Bayesian Classification is a method in which classification takes place using Bayesian theorem. This classification method is also known as independent featured model.Naive Bayesian belongs to a bunch of applied mathematics techniques that square measure referred to as 'supervised classification' as hostile 'unsupervised classification.' In 'supervised classification' the algorithms square measure told concerning two or additional categories to that texts have antecedently been assigned by some human(s) on no matter basis. In Naive Bayesian classification method a particular feature from a category is independent of other features. In several sensible applications, parameter estimation for Naive Bayesian models uses the strategy of most likelihood; in alternative words, one will work with the Naive Bayesian model while not basic cognitive process in Bayesian chance or victimization any Bayesian strategies. This method needs only little information to find mean and variance. As a result of freelance variables square measure assumed, solely the variances of the variables for every category have to be compelled to be determined and not the whole variance matrix.

4. CONCLUSION

This paper successfully presents a new email filtering technique focused on persistent threat and recipient- oriented features that outperforms other available techniques. Targeted malicious emails (TME) for computer network exploitation have become more insidious and more widely documented in recent years. In this paper we develop an alternative filtering procedure by using TME specific feature extraction. The protocols automatically predict congestion in a systematic manner and that it is necessary. Through this method we can easily identify and detect TME.

REFERENCES

[1] D. Erickson, M. Casado, and N. McKeown, ―The Effectiveness of Whitelisting: A User-Study,‖ Proc. Conf. Email and Anti-Spam, 2008.
[2] M. Tran and G. Armitage, ―Evaluating the Use of Spam-Triggered TCP Rate Control to Protect SMTP Servers,‖ Proc. Australian Telecom. Networks and Applications Conf. (ATNAC 04), ATNAC, 2004, pp. 329–335.
[3] N. Ianelli and A. Hackworth, ―Botnets as a vehicle for online crime,‖ Coordination Center, CERT cMellon University, Canegie CERT, 2005.
[4] S. Lab., March 2011 Intelligence Report. Symantec Report 2011, Symantec Corp., Mountain View, CA, USA. 2011.
[5] C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, and S. Savage, ―The heisenbot uncertainty problem: Challenges in separating bots from chaff,‖ in Proc. 1st USENIX Workshop LEET, Apr. 2008, p. 10.
[6] Jeevanandam J, Kumaraswamy YS. Feature reduction using principal component analysis for opinion mining. International Journal of Computer Science and Telecommunications. 2012; 3(5): 118–21.

[7] Chen J, Liu Y, Zhang G, Cai Y, Wang T, Min H. Sentiment analysis for cantonese opinion mining, emerging intelligent data and web technologies (EIDWT), 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies; Xi'an. 2013. p. 496–500.

[8] Z. Zhu, G. Lu, Y. Chen, Z. Fu, and R. P. K. Han, ―Botnet research survey,‖ in Proc. 32nd IEEE Int. Annu. COMPSAC, Turku, Finland, Jul. 28–Aug. 1, 2008, pp. 967–972.

[9] M. Ye, T. Tao, F.-J.Mai, and X.-H. Cheng, ―A spam discrimination based on mail header feature and SVM,‖ in Proc. 4th Int. Conf. WiCOM, Netw. Mobile, Dalian, China, Oct. 12–14, 2008, pp. 1–4.

[10] C.-H. Wu, ―Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks,‖ Expert Syst. Appl., vol. 36, no. 3, pp. 4321–4330, Apr. 2009.

[11] Devi KS, Ravi R. A new feature selection algorithm for efficient spam filtering using adaboost and hashing techniques. Indian Journal of Science and Technology. 2015; 8(13):1–8.