Traffic Shaping of Spam Botnets

Husain Husna, Santi Phithakkitnukoon,and Ram Dantu
Department of Computer Science and Engineering
University of North Texas, Denton, TX 76203 USA.
{*hjh0036, santi, rdantu*}*@unt.edu*

***Abstract*—Compromised computers, known as bots, are the major source of spamming. Detecting them can help greatly improve control of unwanted traffic. In this paper, we develop a traffic control mechanism to detect and delay the traffic of suspicious senders and bots. By delaying spammer's traffic, it has been reported that 90% of spam emails can be eliminated.**

In our proposed mechanism, we group spammers based on their behavior and transmission patterns. These patterns of spammers show high correlation between group members irrespective of geographic location, network ID, content, and kind of receivers. After identification of these Botnet groups we applied traffic shaping techniques a pre-filtering analysis to avoid use of automated machines(Bots) to spam a particular domain. Thus the source for majority of spam is blocked before reaching email servers. We also identify how randomly the Botnets behave and how easy it is to capture a Botnet behavior, based on Information theory. To our knowledge, there is no work reported on detecting and mitigating Botnets based on their behavior and in particular transmission patterns.

I. Introduction

The most challenging aspect of dealing with spam has been the seriousness, threat and growth of 'Botnets'. According to a recent survey, spammers sent an estimated 80% of email spam by using zombie PCs. One of the most common usages of Botnets is to launch massive spams. Spam remains an annoying problem because a majority of spam filtering techniques focus on the content of an email, which is in complete control of the spammers. So, such techniques are not of use as their classification strategies depend upon the message's meaning. Our approach avoids this limitation as we base classification on the individual user's behavior. So one needs understanding of the network of Botnet, its growth dynamics, threat level and an approach to avoid automated 'Bots'. Not much research on Botnet is stated as yet on the behavior analysis, there are few papers related to this work but they have more to do with the IRC relays and their study.

An in-depth understanding of Botnet behavior is a precursor to building effective defenses against this serious and fast growing threat in emails and in future it would be the Voice over IP (VoIP) applications. Using our technique we are able to perform a range of experimental study on new methods and tools for characterizing, comparing, identifying, tracking, dismantling, and preventing Botnets.

Main Contribution

We develop a Traffic control mechanism which, categorize email senders into categories such as legitimate, suspicious and Bots. We analyze traffic from their behavior patterns and delay the traffic of their spamming which allows real time filtering techniques to be used without the risk of false positives. We set up a simple traffic shaping technique called Turing test in our analyzer for all the suspicious and spam mails, which will eliminate the use of automated machines to send spam mails. In addition, we are able to find the trace-ability of botnets based on Information theory.

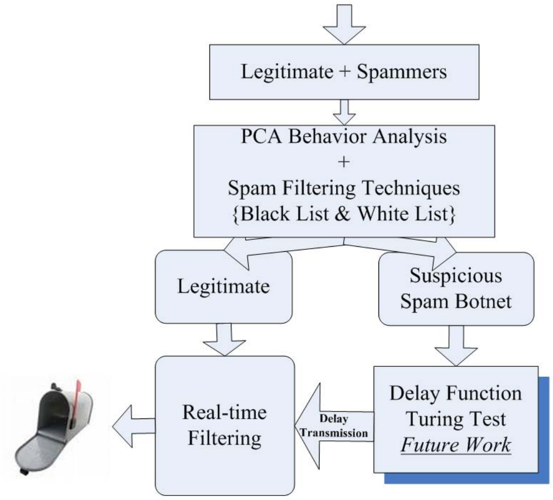

Fig. 1. Flow Diagram of Pre-filtering Analysis to avoid Spam Bots using traffic shaping techniques.

II. Methodology

For our analysis we have considered a corpus of size 8000 emails which includes both legitimate as well as spam emails. Based on the spammers' locations, we categorized the traffic profile of the botnet groups.

Step 1 : First step is to separate Legitimate senders from spammers using Principal component analysis (PCA) [1] [2]. From the spammers only traffic, we apply clustering techniques on their feature set and identify botnet groups [3]. Our study only focuses on the header analysis which is not under the control of the spammer whereas a spammer can spoof the content of an email. Spammers obfuscate their spam emails' content; however, because our analysis does not focus on an email's content, such content is irrelevant to our results. We categorized each email spammer based on features like IP address, Content Length, time of arrival, frequency of spamming and content type. Because we are analyzing

1-4244-1457-1/08/$25.00 © IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE CCNC 2008 proceedings.

spammers' behaviors, other parameters such as reciprocity, read emails, and storage time do not apply, as we assume that users do not read telemarketing email. Using our analysis we were able to separate legitimate and spammers traffic and also identify Bots with a precision of more than 90%.

Step 2 Future Work : After separating out the legitimate traffic from spammers or suspicious spam Bots, we pass the legitimate for advanced filtering techniques. Whereas suspicious and spam Bot traffic is passed ahead for a Turing Test which will delay the transmission. According to [4] 90% of spam mails are avoided if we delay their transmission channel. In our proposed approach, Turing test will not only delay spam traffic but also mitigate automatic use of zombie machines (bots) which are used to send large chunk of spam mails. Methodology of the proposed filter is a pre-filtering technique, which is controlled by the network administrator.

III. Randomness of Bot Structures

How easy it is to detect/trace botnet spam machines?

To quantify the randomness or amount of predictable structure in an individual Botnet group, the information entropy can be used. The information entropy or Shannons entropy is a measure of uncertainty of a random variable. The information entropy as given in (1) was introduced by Shannon [5].

$$H(X) = -\sum_{x} p(x) log_2 p(x), \tag{1}$$

where X is a discrete random variable, and the probability mass function $p(x) = Pr(X = x)$. The spam botnet pattern can be observed from the active time, time of arrival, frequency and content length of email spam. Let A, T, N and C be random variables representing active time, time of arrival, frequency and content length respectively. The entropy of time of arrival can be calculated by (2).

$$H(T) = -\sum_{t} p(t) log_2 p(t), \tag{2}$$

where the probability $p(t)$ is a ratio of the number of mails during t^{th} hour slot to the total number of mails of all time slots. By the same token, the randomness in the spammers active time $H(A)$, content length $H(C)$and frequency $H(N)$, can also be quantified using information entropy which is defined in (1).

Results of Entropy values shows that behavior structure of Botnets are more predictable and less random as compared to normal spammers. Also by comparing legitimate traffic and spam bots, we conclude that they vary a lot in terms of its entropy value. Thus one can easily group Bots, spammers and legitimate from a given traffic analysis. Normal spamming patterns are quite random and thus they have a high entropy value, where as Bots usually tend to have a similar structure of spamming a domain, which results it being less entropic. As an example, a low entropic Botnet group is considered to have a similar structure of spamming in terms of its time of arrival, active time, frequency and content length. It is much easier to trace and identify low entropic Bot in terms of its behavior pattern. Where as a high entropic Bot having random patterns shows that it is changing its pattern of spamming regularly to avoid being traced. Such Bots are difficult to trace in terms of their spamming pattern.

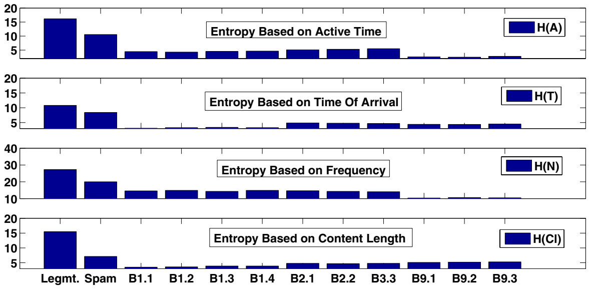

Fig. 2. Figure shows Entropy (Randomness) values of Legitimate, Spammers and Botnets.

IV. Conclusion

Spammers are aggressive and smart and continue to use new techniques to send large volumes of spam. The result is not only an enormous quantity of spam in users mailboxes, but also email servers that are brought down by excessive quantities of unwanted content using Bots. This leads to an increasing investments in servers, network security administrators and software to keep up with the growing deluge of spam. Therefore, we propose a model that blocks spam based on spammers readily identifiable behavior. By applying a pre-filtering analysis of senders reputation, header based filters can operate much more efficiently, since the vast majority of spam is blocked before reaching email servers. However, Traffic shaping does not block email, allowing real-time filtering techniques to be used without the risk of false positives.

In future, we will be working on traffic shaping techniques to reduce some legitimate delay, if any. Also, we will apply Principal-factor analysis on email traffic in order to study the relationship of the randomness (Entropy) levels of Botnets based on their underlying parameters. Furthermore, using factor analysis we will be able to capture the trends of Botnets behavior on various parameters and interesting relationship between randomness levels in Bots, spammers and legitimate traffic.

References

[1] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer Series in Statistics, 1986, New York, USA.

[2] L. Fulu, M. Hsieh, "An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies," *CEAS*, 2006.

[3] H. Husna, S. Phithakkitnukoon, R. Dantu, "Behavior Analysis of Spam Botnets," *IEEE COMSWARE'08*, In submission.

[4] Osterman Research, Inc., "The Advantages of Using Traffic- Shaping Techniques to Control Spam," January 2007.

[5] C.E. Shanon, "A mathematical theory of communication," *Bell System Technical Journal,* vol. 27, pp.379-423 and 623-656, July and October 1948.

[0] This work is supported by the National Science Foundation under grants CNS- 0627754, CNS-0516807, CNS-0619871, and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.