Detecting Spambot as an Antispam Technique for Web Internet BBS

[1]Juned M Laliwala [2]Girish Khilari
[1]Gujarat Technological University PG School, Ahmedabad, Gujarat
[2]MS (Software Systems),BITS Pilani

Abstract— Spam which is one of the most popular and also the most relevant topic that needs to be understood in the current scenario. Everyone whether it may be a small child or an old person are using emails everyday all around the world. The scenario which we are seeing is that almost no one is aware or in simple sentence they do not know what actually the spam is and what they will do in their systems. Spam in general means unsolicited or unwanted mails. Botnets are considered one of the main source of the spam. Botnet means the group of software's called bots and the function of these bots is to run on several compromised computers autonomously and automatically. The main objective of this paper is to detect such a bot or spambots for the Bulletin Board System (BBS). BBS is a computer that is running software that allows users to leave a message and access information of general interest. Originally BBSes were accessed only over a phone line using a modem, but nowadays some BBSes allowed access via a Telnet, packet switched network, or packet radio connection. The main methodology that we are going to focus is on Behavioural-based Spam Detection (BSD) method. Behavioral-based Spam Detector (BSD) combines several behaviours of the spam bots at different stages including the behaviour of spam preparation before the spam session when the spammers search for an open relay SMTP service to send e-mails through, and the behaviour of spammers while connecting to the mail server. Detecting the abnormal behaviour produced by the spam activities gives a high rate of suspicion on the existence of bots.

Keywords— SPAM, Botnet detection, Network threat detection, Network worm detection.

I. INTRODUCTION

There are several challenges facing the e-mail systems; for example, the increase of harmful techniques has forced e-mail users to search for the higher degree of safety and privacy to ensure the security of the transmitted information [1]. This is due to the recent spread of viruses, hackers, malwares, worms, and Botnets.

Spam is one of these challenges that abuse the electronic messaging system by sending a huge amount of unrequested bulk messages randomly that makes up of 80% of the emails as a spam [2]. The reason behind this high percentage is due to the armies of the harmful bots that are controlled by a bot master. Botnets refer to a group of software called „bots" or „robots". The function of these bots is to run on several computers autonomously and automatically [3]. This kind of software usually `works at the end-user system that has been infected. Once these bots are installed, they send an identification message to the bot master.

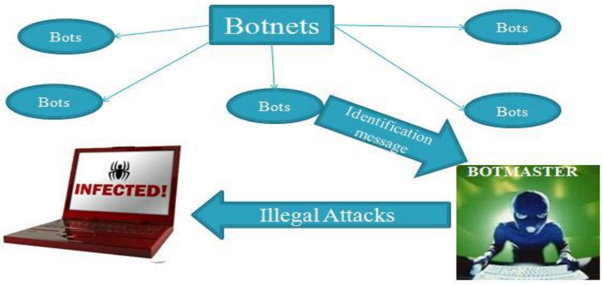

Fig1. Botnet Architecture

The bot master can start any command and control session by using these infected computers that are called „zombies". The bot master performs illegal attacks on all these zombies. Bots work under the shadow to avoid being detected by an antivirus or observed by the user. Bots software has the ability to disable the antivirus effect by producing an anti-antivirus [2]. The best time for the bots to start performing their activities is during the idle period of the host computer. This happens especially when the bots sense the low CPU utilization; hence they start to take advantage of the infected host resources to do their desired activities.

As the Internet becomes more and more popular, the Internet becomes not only sources of providing information, but also channels of effective communication. Since the Internet may cause severe effects to their society [1], most of political organizations and media companies keep monitoring information posted on popular websites. Web sites like CNN or Facebook, which have millions of daily visits, provide service spaces for communication. Most of people regard such a space as a powerful way to communicate with others, but spammers think this space as a powerful way to advertise their contents in an illegal manner. Advertisements by spammers exist especially on BBS (Bulletin Board System) in major media web sites (like CNN and The Times) or YouTube comment pages. So developing a BBS anti-spam module for public websites is no option for website builders. So many kinds of spam-bots on BBS were invented and used to generate spams by attackers. On the other hand, security experts have developed antispam engines against spam attacks. At first, it works well but causes another side effect. Some BBS spam-bots use bypass attack to neutralize spam filters (aka false negative), and some spam filters block even normal users" articles (aka false positive).

II. EXISTING WORK ON SPAM DETECTION

A. *List Filters*

At present, there are many proposed and developed software and filters aimed to mitigate spamming. These are different attempts; each one has to fight spam from different places and perspectives. For instance, list filters (i.e. the white, black, and gray filters) [6] concern more with the trustable sender's address that is usually placed at the top of the DNS server (i.e. as in the DNSBL applications). Whitelist maintains the list of the e-mails that the users wish to receive mail Blacklist maintains the list of the e-mails that the users does not wish to receive mails. Graylist maintains the intermediate list of the e-mails where the users who have frequently sends the mail within N hours/minute can be stored and accordingly the user can wish to enter that particular e-mail in blacklist or whitelist.

B. Domain Name System-BlackList (DNSBL)

One of the well-known techniques that are being used is the DNSBL and it is used by the Mail Abuse Prevention [5]. Basically, it is a real-time database that contains IPs of all the discovered spams such as bots and Trojans. The DNSBL is built at the high level of the DNS server, which is the largest distributed database that contains IPs, and records of names for each domain. The record that points to the mail service is the Mail Exchanger (MX). The mail server checks for MX record that belongs to the receiver‟s e-mails. However, if the sender's IP is already listed in the black list, the server will reject the connection and aborts from providing the MTA with the MX record.

C. Signature-based Spam Detection

Signature-based Spam Detection is a widely used in many mail server systems and it depends on some statistical methods to produce hash value, which is attached with each e-mail to become a marker or signature that classifies the e-mail. By making a comparison with the spam e-mails discovered earlier, the received e-mail is recognized and marked as a spam. Then, this e-mail hash value is stored and distributed to all the filters that use signature technique. It is difficult to calculate the hash value because it depends on specific structures and words that e-mails contain, such as (porn materials, Click Here, Join Us) which give a suspicious value to the e-mail weight. Hash technique or signature gives accepted prevention an improbable percentage to classify legitimate e-mails as spam because it depends on the calculated hash value of the e-mails that are reported as spam.

III. BEHAVIORAL-BASED SPAMMING DETECTOR

Information extracted from the network-level traffic has several advantages, Network-level information is clear and closer to the spam source than e-mails header and content [11]. One of the advantages of the behaviour-based technique is the ability to detect zero-day spam bots and its potential for early detection. It can be used to detect the spamming attempts before the end of the spam sending session. Hence, recent studies focus on the abnormal behavioural detection. Since spam is an abnormal behaviour, and it uses and consumes the network resources illegally. This abnormal behaviour leaves evidence that point to its occurrence on the network.

– *A BSD framework*

The proposed system as shown in Figure 2 above is divided into four main stages; preparing the system's settings, capturing, analysis, and results. At the first stage, the system is prepared by configuring and adding the required variables. At the second stage, the packets are captured and filtered. Then, the third stage is decoding and classifying the packets and inserting the required traffic into the database.

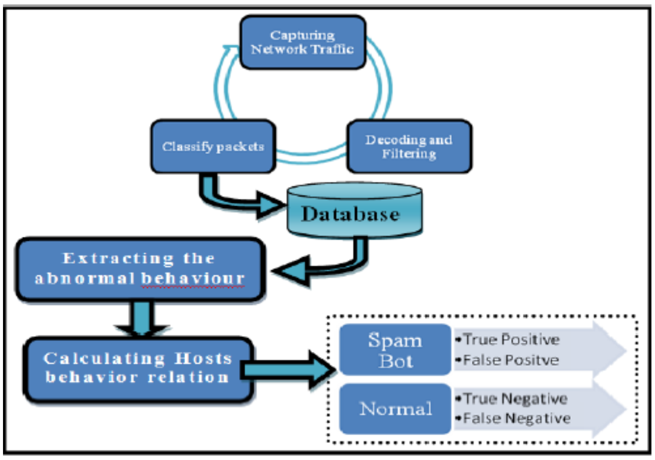

Fig 2: BSD Framework

After that the timer (T) equalizes the observation period (Op). The next process is to analyse and extract the useful information from the captured packets that are inserted into the database, and then extracting the abnormal behaviours in the captured traffic. Finally, the relation rate will be calculated to decide whether the detected host behaviours during the observation period can lead to detect a spam bot activities and spamming attempts on the network.

– *PREPARING THE SYSTEM:*

The following points summarize all the necessary values and information required to start the system:

1) Mail Server MAC and IP Address
2) Mail Servers Priority as stated in the MX Query reply
3) SMTP port used in the SMTP server (default value 25)
4) Other variables used in the programming application

– *CAPTURING AND DECODING STAGE:*

The packets that traverse the network are captured and filtered. The filters are used to reduce the number of the packets to be processed are listed below. • IP Source and IP Destination within local network IP range. AND • TCP/IP "0x06". AND • Source Port is SMTP port (default 25). Finally, the packets are inserted into the buffer. The buffer is based on the Queue technique; FIFO (first in first out) to keep the packet in the same order as sniffed. The buffer places all the packets directly in the system, and keeps capturing them without delay.

– *DECODE AND CLASSIFY PACKETS:*

This stage begins when the buffer is not empty; it starts to get packets from the buffer, then decodes and classifies the packets into either SMTP connection packets or service scanning attempts packets. Then, the packets are stored into the system's database to be analysed and mined later in the next stage. After filtering the packets, they are presented as either established SMTP connection packets, or initiating

SMTP connection packets through the TCP three handshakes with the destination SMTP port.

– *IDENTIFYING THE HOST'S BEHAVIOUR:*
This stage starts once the observation period ends and is determined by the system administrator (tn: is the observer period). Analyzing the packets collected from the previous stage is done at this stage. It involves several tests and analysis procedures to extract and count the total number of occurrence in the network during the experiments of each behaviour.

IV. ANTISPAM TECHNIQUE FOR BBS

A. *Introduction*

According to the survey of BBS spams[4], other secure input modules for web applications to block BBS spam-bots [3] were also developed. Figure 1 shows the characteristics of each technology.

1) *WORD FILTER:* It checks suspicious words that frequently used in spam articles, e.g. bacara, sex, adult. If a filter detects these words in an article, then it blocks the article. But so many variants and bypass attacks can be possible.
2) IP BLACK / WHITE LIST: A source IP address is examined and if it is on the black list, then articles posted from the IP address are blocked. But it cannot block spam articles posted via a proxy server since the IP address of the proxy server may not on the black list. In addition, it is hard to block attacks that are using CSRF(Cross Site Request Forgery).

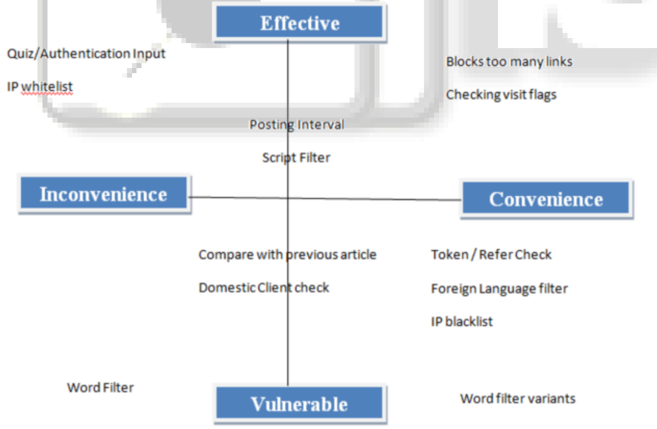

Fig 3: Characteristics of Antispam technologies

3) BLOCKING ARTICLES WITH TOO MANY LINKS:
It is based on characteristics of spam articles. There usually are many links in an article if the article is for advertisement, which was reported in 2008. It exploits the Zeroboard (origin of Xpress Engine) administration module vulnerability.
4) FILTERING ARTICLES WITH A FOREIGN LANGUAGE ONLY (domestic service site only) This filtering mechanism is effective only in non-English-speaking countries. Most of articles written in a foreign language are spams.
5) SCRIPT FILTER: Some spam-bots launch XSS attacks. These bots usually change a victim's URL address to their own site address. So blocking script tag is very effective to block spams.

6) *CHECKING WHETHER A LIST/INDEX PAGE WAS ACCESSED* Most of spam bots directly request write pages without visiting an index / list page.
7) *MANDATORY INCLUSION OF CERTAIN KEYWORDS:* A system can block article posting if the article doesn't include specific keywords (usually configured by an administrator) [5]. It is effective for block spam articles but it can cause users‟ inconvenience.
8) *TIME INTERVAL CHECK:* Spam bots usually post many articles in a short time interval. This filtering mechanism checks a time interval between two postings from the same user. If it is too short, the system will block postings from this user.
9) *DOMESTIC CLIENT CHECK (domestic service site only)* Similar to the IP blacklist mechanism and the foreign language filtering mechanism, this mechanism filters articles posted foreign servers.
10) QUIZ / AUTHENTICATION INPUT: This mechanism uses a quiz or a distorted image which only humans can solve [3]. It is one of the most effective ways to block spam articles, but it can also cause users‟ inconvenience.
11) *TOKEN / REFERRER CHECK* This mechanism is usually used for open source boards. Each site builder adds extra hidden form fields to verify automated postings which exploit open-source web boards.

B. *Mechanism for adaptable Antispam techniques*

Adaptable anti-spam engine has 3-step filters followed by:

- Low level filter : simple / small overhead
- Mid-level filter : more strict policy / allows large overhead
- High level filter : only humans can pass / can cause inconvenience

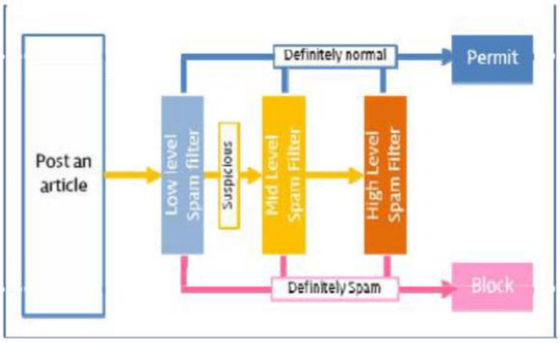

Fig 4: Mechanism for Antispam techniqueh

– LOW LEVEL ANTISPAM POLICY should cause the least overhead to server and allowed to cause some false positives (like simple word filter method). If spam-suspicious article is blocked by low level spam filter, then it will be checked with mid level anti-spam policy.

– MID LEVEL FILTER has more strict policy and may cause some performance down (like comparing with large spammer IP list table). Site builders can be free from server performance problem caused by spam filter, because most of normal article will be posted through just only low level spam filter. It means that anti-spam engine can embed effective mid/high level spam filters which have side effect like performance down. By the flexibility of this engine, low level filter can use anti-spam policy which can cause frequent false positive.

High level which is the final step filter has the strictest anti-spam policy. It needs extra input that only humans can solve, like typing authentication word from distorted image [3,7]. Or getting answer from simple quiz can be another solution (of course, it should not be solved automatically). These methods are effective to block spam articles, but may cause inconvenience to visitors. So site builders should embed and design properly first two filters that high level filters should be rarely applied to normal articles.

V. CONCLUSION

The behavioral-based Spamming Detector (BSD) system is the one and only one method that we can find more relevant because it more focuses on multiple behaviors on the network which are considered as spam related activity. The main function of the BSD system is to constantly monitor the network level traffic where its information is constant and closer to the sources. Network traffic information is more trustable than the information extracted from the e-mail header at the application layer. BSD system is implemented by using java in the java RCP-eclipse environment.

As there is some misunderstanding between the BBS spammers and the site builders, many customers or normal visitors feels incompatibility for such spam articles and their respective strict spam policies. But the solution that proposed in this paper can solve these two problems flexibly. Of course, some variant forms of spam-bots will appear someday and more advanced anti-spam technologies will be preceded, as it has been done. We expect this adaptable anti-spam technology can block the most of spam articles for a while.

REFERENCES

[1] Mohammed Fadhil Zamil, Ahmed M. Manasrah, Omar Amir, Sureswaran Ramadass “A BEHAVIOR BASED ALGORITHM TO DETECT SPAM BOTS” National Advanced IPv6 Center, Universiti Sains Malaysia.

[2] Joonmo Hong, Boo Joong Kang, Eul Gyu Im “ADAPTABLE ANTI-SPAM TECHNIQUE FOR THE INTERNET WEB BBS” Division of Computer Science & Engineering, Hanyang University, Seoul, 133-791

[3] R. Qiong, M. Yi, & W. Susilo, “SEFAP: An Email System for Anti-Phishing,” ICIS 6th IEEE/ACIS International Conference on Computer and Information Science, pp. 782-787, 2007.

[4] J. S. Sauver, “Spam Zombies and Inbound Flows to Compromised Customer Systems,” AAWG

[5] Z. Zhaosheng, L. Guohan, C. Yan, Z.J. Fu, P. Roberts, and H. Keesook, “Botnet Research Survey,” Computer Software and Applications. COMPSAC '08. 32nd Annual IEEE International, pp. 967-972, 2008.

[6] Y. Miao, J. Qiu-Xiang, and M. Fan-Jin, “The Spam Filtering Technology Based on SVM and D-S Theory,” Knowledge Discovery and Data Mining, 2008. WKDD. First International Workshop on, pp. 562-565, 2008.

[7] Introduction to the Realtime Blackhole List (RBL). MAPS-Global Secure Systems [website], Nov 28, 2008, Available: http://www.mailabuse.com/wp_introrbl.html.