ISSN:0975-9646

Stealthy P2P Botnet Detection Based on Automated Threshold Computing Using Genetic Algorithm

Vishnu S Sekhar, Neena Joseph

Department Of Computer Science
,Mangalam College Of Engineering
Ettumanoor,Kottayam,Kerala,India

Abstract— **Botnet constitutes several hosts connected in a network which are utilized to perform malicious activities including DDOS attacks,spamming,Phishing etc..by a illegitimate user namely the botmaster using a command and control mechanism(c&c).When the hosts in a botnet communicate in a P2P(peer to peer) fashion,it forms a P2P botnet.We aim at detecting the P2P botnet with minimized value for false postiveness by introducing a method to calculate the threshold automatically so that the attacks with value of certain parameters above that threshold are only declared as an attack.**

Keywords— **Botnet,P2P,GA,Probe,R2L,U2R**

I. INTRODUCTION

The botnets in the earlier times made use of a centralized communication mechanism for example by relying on an internet relay chat(IRC) server. The attacks via this mechanisms were easily detected and not so prominent as the centralized control becomes a bottleneck. In the recent times the P2P applications like Bittorent,Skype and many others became so popular and were widely employed in the networks. With the emergence of such P2P applications, a decentralized approach towards botnet formation and malware spreading came into play. Some of the most popular threats or malwares spreading making use of P2P applications includes Trojans like Storm,nugache etc..The attacking patterns became more and more stealthier due to the mixed behavior of the hosts in the network which exhibits the behavior of a legitimate node as well as a P2P botnet.Apart from that in a huge network traffic, the detection of threats or attacks becomes even more complex. Thus there arises the need for a new mechanism for detecting the P2P botnets. Even if there are many botnet detection mechanisms already existing none of them could make a clear distinction between the legitimate users and the P2P botnets.The detection of malware can be based on certain factors like port scans,number of packets send to the same interface,password guessing,ip spoofing,session hacking ad so on. But the problem to be addressed is how to make up a clear distinction between an attack and a genuine communication.This is based on seeting up a threshold value for each of the above cases so that the traffic flows exceeding ths thresholdwill only be classified as a threat and others will be listed as normal communications.

II. RELATED WORKS

There are many researches going on in the field of network security and intrusion detection. The P2P traffic in the network could be identified by using certain application level signatures[1].In the P2P applications the peers communicate directly without the help of the DNS.That is by a a direct IP communication.For ecxample in the P2P file-sharing applications,a peer directly searches for the content on another peer which posess the same. For detecting P2P traffic the dowloading phase is only considered.It provides signatures usually corresponding to the protocols being used.Based on that the P2P traffic is analysed and threats are detected if any. The signatures of infections includes registry keys,keystroke log files etc..But such detection methods are not fool proof.It may also possess a high value for false positiveness sometimes.

The other methods of finding P2P botsincludes botnet detection by structured graph analysis[4].It deals with localizing the botnet members based on unique communication patterns.It uses Botgrep algorithm to make up a clear distinction between the traffic papperns of P2P botnets and legitimate P2P.It separates the P2P communications from rest of the packet flows using information from communication graphs.The Botgrep architecture constituted the following steps.

- Collecting the communication graph.it may be provided by the ISP.
- Misuse detection
- Isolate the botnet communication sub graph
- Once the c&c structure is found as malicious, their list can be used to install blacklists in routers.

The pitfall of the above technique was the privacy issues regarding the collection and analysis of communication graphs.

The detection of storm P2P/spambot communication patterns[2] can be done only by traffic analysis.Infection increases the volume and breadth of UDP and TCP communication from local host to external targets.But the major drawback is setting up a threshold factor for the volume of communication flows,such that the traffic flows above the threshold is declared to be an attack.That is the major task associated with it is also how to bring about what is called a clear distinction between an attack and a legitimate P2P traffic.

The above discussions reveal the need for a mechanism to set up the threshold separately for each kind of attacks via

the network in such a way that it makes a clear cut distinction between the the threats or P2P botnet and a legitimate P2P application.In the following sections we discuss the methods to find the most optimal value for the threshold associated with each of the attacks including probe attack,ddos attack,R2L attack,U2R attack,session hacking,spamming .

III. SYSTEM DESIGN

The system design basically constitutes of the following basic blocks as shown in the architectural diagram below.

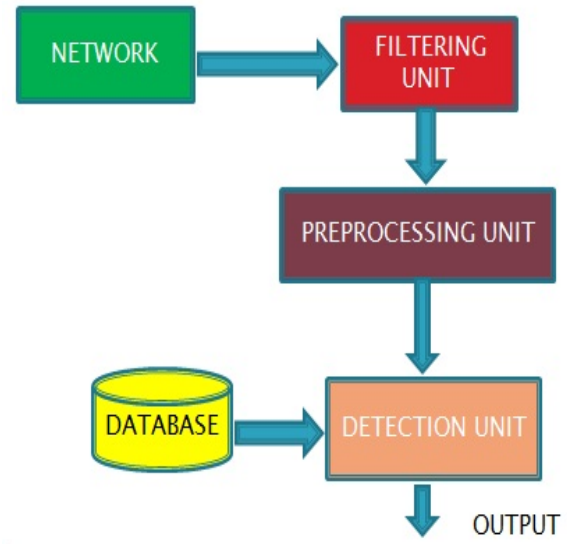

Fig. System Architecture

A. *Filtering Unit*

The filtering unit captures the packet moving in and out of the network.The tapping of the traffic can be done by using packet capturing softwares like Jpcap,Wincap etc..Along the details of the packets like packet size,protocol,source IP,destination IP can be extracted.Another function of filtering unit is that it filters out all traffic other than the P2P.

B. *Preprocessing Unit*

This unit groups the further the packets send between the real P2P applications using certain signatures which is represented by [Pkts, Pktr, Bytes, Byter].It further filters out all other communication traffic which are non P2P by finding out the packets that require DNS resolving.This can be used to distinguish P2P and non P2P applications as P2P uses communication without the help of DNS.

C. *Detection Unit*

This unit is concerned with the detection of the threats or the malware. The detection is based on factors like number of portscans, unusually frequent logins and signups, high rate of packets send to a single port. There is a threshold associated with each of them described. This threshold value is automatically generated by applying a data mining algorithm over the past history of traffic flows present in the database. The packet with fingerprints or signatures with values of the above given parameters exceeding their respective threshold will be classified as an attack say DDOS,U2R,R2L and so on. The algorithm we use here is the genetic algorithm. The genetic algorithm can be applied to the past history of network traffic constituting of the spread of various threats and its consequences. According to the unusual behavioural patterns like continuous port scans, unlimited connection request to the same IP or multiple packets destined to the same interface, extremely large number of packets with urgent pointer set, session hijacking etc..Based on that the threshold is set for each of the attacks.

IV. IMPLEMENTATION

At the implementation point of view,the various modules involved in the botnet detection are as follows.

- Packet Capture
- Filter
- GA unit
- Detection System
- DDOS layer
- R2L layer
- U2R layer
- Probe layer
- Storm layer

A. *Packet Capture*

The WinPcap software provides facilities to capture raw packets, both the ones destined to the machine where it's running and the ones exchanged by other hosts(on shared media), filter the packets according to user specified rules before dispatching them to the application, transmit raw packets to the network and to gather statistical values on the network traffic. It captures all the packets in the Network Interface by using Jpcap captor.

B. *Filter*

Filters the traffic based on signatures like size of the packets,protocols used,destination and source addresses,mac address etc...

C. *GA Unit*

This unit work over the previous history of traffic flows by applying the so called genetic algorithm to it to evaluate the the threshold associated with various attacks including DDOS,R2L,U2R etc..The algorithm proceeds in the following steps.

1) set the sequence length N and number of rounds R
2) From the history select randomly N tuples constituting of number of packets with each attack in a given network traffic
3) Crossover: Copy the N tuples as itself and add N more tuples formed by crossing over initial N tuples
4) Mutation: Copy the Inital N tuples as such and add to it N more Tuples by making small changes in the different fields of tuples in previous step
5) Select N best solutions from the previous stage by considering the weight=(Number

of right assumptions-Number of Wrong assumptions)/N

6) Repeat steps 2 to 5 using the N tuples from step 5 for R rounds
7) Select solution with highest weight as the best solution

D. *Detection Unit*

It detects the various P2P botnet attacks based on the threshold provided by the GA unit.

E. *Probe layer*

The probe attacks are aimed at acquiring information about the target network from a source that is often external to the network. Hence, basic connection level features such as the "duration of connection" and "source bytes" are significant while features like "number of files creations" and "number of files accessed" are not expected to provide information for detecting probes.Port Scans or sweeping of a given IP-address range typically fall in this category.

F. DDOS Unit

The DDoS attacks are meant to force the target to stop the service that is provided by flooding it with illegitimate requests. Hence, for the DDoS layer, traffic features such as the "percentage of connections having same destination host and same service" and packet level features such as the "source bytes" and "percentage of packets with errors" is significant. To detect DoS attacks, it may not be important to know whether a user is "logged in or not."

G. R2L Layer

The R2L attacks are one of the most difficult to detect as they involve the network level and the host level features.We therefore selected both the network level features suchas the "duration of connection" and "service requested"and the host level features such as the "number of failed login attempts" among others for detecting R2L attacks. Attackers do not have an account on the victim machine,hence tries to gain access, these are guess password, ftp write, multihop etc.

H. U2R Layer

The U2R attacks involve the semantic details that are very difficult to capture at an early stage. Such attacks are often content based and target an application. Hence, for U2R attacks, we selected features such as "number of file creations" and "number of shell prompts invoked," while ignores features such as "protocol" and "source bytes." An attacker has local access to the victim machine and tries to gain super user privileges; these are buffer overflow, rootkit.

I. Storm layer

This layer detects the misbehaviour in the network like IP spoofing.The attacker sends threats via different IPs generated making the threat detection much more difficult.

V. Performance Evaluation

The performance of the technique of threshold calculation based on the genetic algorithm can be measured based on three factors.

- False positiveness of detection
- False negativeness of detection
- Accuracy of detection

A.False Positiveness

It is the event of detecting a legitimate user or application as a threat.Increase in false positiveness causes degradation in the overall performance factor of a threat detection system.By applying genetic algorithm,the threshold value calculated such that all packet flows exceeding that threshold will be declared as an attack.This will reduce the false positiveness(FP) of the detection system when compared to
the old methods.The reduction in the false positiveness in the new method is show in the graph below.

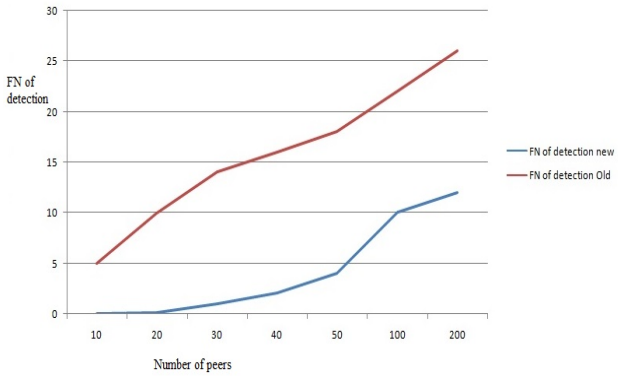

Fig.2 Reduction in false positiveness

B. False negativeness

It is the event of leaving a threat or attack undetected. Increase in false negativeness also causes degradation in the overall performance factor of a threat detection system.By applying genetic algorithm,the threshold value is calculated such that all packet flows exceeding that threshold will be declared as an attack.This will reduce the false negativeness(FN) of the detection system when compared to the old methods.The reduction in the false positiveness in the new method is show in the graph below.

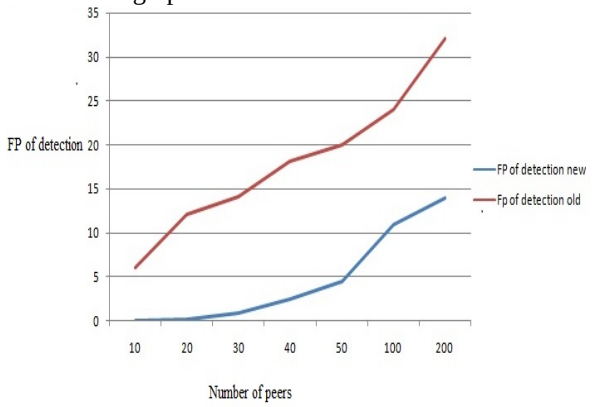

Fig.3 Reduction in false negativeness

C. Accuracy

The accuracy of threshold calculated is said to be optimal when it gives the maximum weight or rank.This is almost directly proportional to the number of rounds N which is fixed in the initial phase of the genetic algorithm .More the number of iterations or rounds carried out,the more accurate will be the threshold obtained.The graph below shows the increase in accuracy of the threshold calculation with the increase in number of rounds.

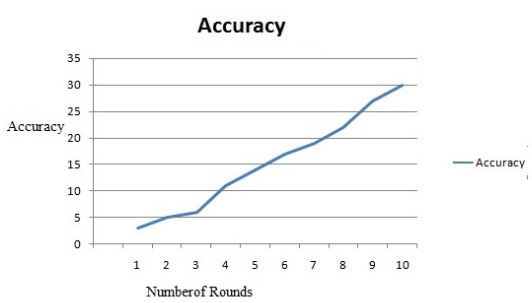

Fig. 4 Accuracy

VI. Conclusion And Future Scope

Many P2P BOTNET detection methods have been developed recently but couldnot work well mainly due to the mixed behaviour of the P2P traffic ie,exhibiting\ the behaviour of both legitimate as well as P2P BOTNETs and was also not scalable as compared to the huge network traffic.Here,scalable stealthy BOTNET detection mechanism has been implemented. To accomplish this task, we derive statistical fingerprints of the P2P communications to first detect P2P clients and further distinguish between those that are part of legitimate P2P networks (e.g.,file-sharing networks) and P2P bots.We also identify the performance bottleneck of our system and optimize its scalability.The difficulty associated with the method is how to set the threshold values which differentites legitimate P2P traffic and P2P BOTNETs.Here we have implemented the automated threshold calculation based on the genetic algorithm and analyzed its performance also.Even if there is reduction in the false negativeness and false positivess of detection,there is factor of time complexity associated with theshold calculation which is a ground for Improvement.

Acknowledgment

With prayers to God for his grace and blessings, for without his unforeseen guidance, this work would have remained only in my dreams. I would like to acknowledge with thanks the contributions given by the management of our college. I express my sincere gratitude to our Principal, Head of the Department of Computer Science for permitting me to do the same and for encouraging me. I would like to extend my regards to my Guide for her valuable ideas, inspiration, cooperation and technical guidance at every stage of the project which helped to keep the work on right track. Last but not the least, I would like to thank my parents, friends, teaching, and non-teaching staff of the college.

References

1. S. Stover, D. Dittrich, J. Hernandez, and S. Dietrich, “Analysis of the storm and nugache trojans: P2P is here,” in Proc. USENIX, vol. 32. 2007, pp.18–27..
2. P. Porras, H. Saidi, and V. Yegneswaran, “A multi-perspective analysis of the storm (peacomm) worm,” Comput. Sci. Lab., SRI Int., Menlo Park,CA, USA, Tech. Rep., 2007.
3. W. Liao and C. Chang, “Peer to peer botnet detection using data mining scheme,” in Proc. IEEE Int. Conf. ITA, Aug. 2010, pp. 1–4.
4. S. Nagaraja, P. Mittal, C.-Y. Hong, M. Caesar, and N. Borisov, “BotGrep:Finding P2P bots with structured graph analysis,” in Proc. USENIX Secu-rity, 2010, pp. 1–16.
5. Zhang, R. Perdisci, W. Lee, U. Sarfraz, and X. Luo, “Detecting stealthy P2P botnets using statistical traffic fingerprints,” in Proc. IEEE/IFIP 41st Int. Conf. DSN, Jun. 2011, pp. 121–132.
6. Subhabrata Sen,Oliver Spatscheck and Dongmei Wang,“Accurate Scalable In-Network Identification of P2P Traffic Using Application Signature“,2004,pp.512-521.