DETECTION OF BOTNET USING FUZZY C-MEANS CLUSTERING BY ANALYSING THE NETWORK TRAFFIC

Sindhu Arumugam, Dr. V. Vanitha, Ms. V.P. Sumathi

Abstract— Botnets are networks formed by malware-compromised machines which become a serious threat to the Internet. Detection of Botnet can be done in two ways either by using signature based or anomaly based techniques. signature based techniques detect the presence of the known bots accurately but it is not possible for this approach to detect evolving new types of bots whereas anomaly based detection techniques are popular as its capability to detect the bot variants, even unknown bots. Botnet detection is a challenging task, since the creators of botnets continue to adopt innovative means in creating botnets. This paper presents the detection of bot variants by implementing the Fuzzy C-Means clustering algorithm in storm tool.

Index Terms— Botnet, Centralized, Decentralized, C&C Server, DDOS, IRCbot, P2Pbot , HTTPbot, DNSbot

1 Introduction

A Cyber Attack is an attack which employed by individuals or whole organizations from a computer that targets another computer or a website, with an aim of compromising the confidentiality, integrity, or availability of target and the information stored in it. There are many methods of Cyber Attacks that provide a distributed platform for many cyber-crimes such as Distributed Denial of Service (DDoS) attacks against critical targets, malware dissemination, click fraud, phishing, etc. Bots are one of the most sophisticated and popular types of cyber attack nowadays. Botmaster controls the many computers at a time, and turn them into 'zombie' computers, which operate as part of a powerful 'botnet' for performing criminal activities.

Bot's life cycle [1] consists of four stages which shown in the Fig 1. Injection stage is the first stage of any computer to become real bot by downloading malicious software from websites or infected files attached to emails unknowingly. In connection stage, the bot communicate with the server by using the IP addresses which is encoded directly as a list of IP addresses or domains, which can be static or dynamic. While this makes it more difficult to take down or block a specific C&C server. In the execution stage, the bot perform the malicious activities such as information theft, DDoS attacks, spreading malware, stealing computer resources, monitoring network traffic, spamming, phishing, etc as directed by the botmaster. Maintenance and upgrading stage is the last stage of the life cycle. Maintenance is necessary if the botmaster wants to keep his army of zombies or to update codes for many reasons like adding new features, moving to another C&C server or escaping from detection techniques.

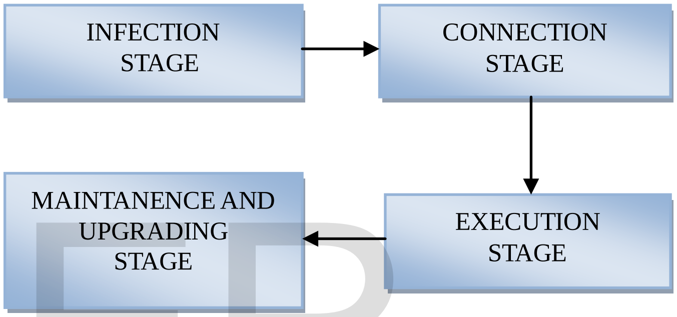


Fig. 1. Life Cycle of Botnet

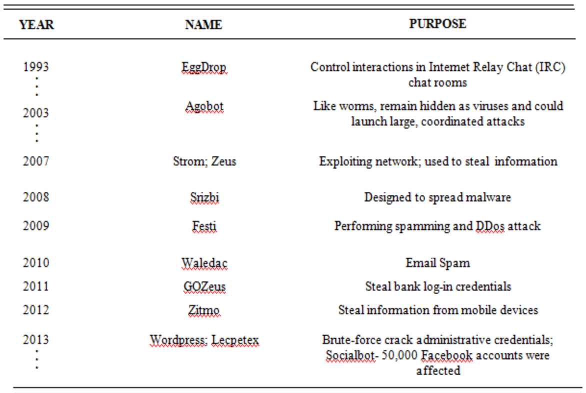
According to [1], the architecture of botnet is based on how they communicate with the bots. It is classified into three types, they are centralized, decentralized and hybrid. In a centralized architecture, all the bots report to and receive commands from a single C&C server whereas in decentralized architecture, the communication between bots and server or among the bots will be of peer to peer contact. Hybrid architecture is the combination of the centralized and decentralized structure. Bots are classified into four types based on their architectural design namely, IRC bot, P2P bot, HTTP bot, and DNS bot. The IRC bots, follow the PUSH approach as they connect to selected channels and remain in the connect state till it receives command from botmaster. P2P bot architecture, instead of having a central C&C server, the botmaster attack the target host by sending a command to one or more bots, and they deliver it to their neighbors. HTTP bot uses the HTTP protocol to send the commands via web servers which enables these bot to periodically visit certain web servers to get updates or new commands. DNS bots are the bots which make use of DNS to receive commands to perform malicious activities.

Detection of botnet at early stage is necessary since botnet creators are designing powerful bots in order to perform spamming, DDos, steal bank log-in credentials, and even started to attack the emerging technologies such as cloud computing, social media, smart phone technology, etc which cause huge financial losses. Effects of some bots are shown in

- *Sindhu Arumugam is currently pursuing masters degree program in computer science and engineering in Kumaraguru College of Technology, India. E-mail: asindhukct@gmail.com*
- *Dr. V. Vanitha is currently working as Professor, department of computer science and engineering in Kumaraguru College of Technology, India. E-mail: vanitha.v.cse@kct.ac.in*
- *Ms. V.P. Sumathi is currently working as Assitant Professor, department of computer science and engineering in Kumaraguru College of Technology, India. E-mail: sumathi.vp.cse@kct.ac.in*

the Table 1

TABLE 1
SOME OF THE BOT'S EFFECTS



YEAR	NAME	PURPOSE
1993 ⋮	EggDrop	Control interactions in Internet Relay Chat (IRC) chat rooms
2003 ⋮	Agobot	Like worms, remain hidden as viruses and could launch large, coordinated attacks
2007	Strom; Zeus	Exploiting network; used to steal information
2008	Srizbi	Designed to spread malware
2009	Festi	Performing spamming and DDos attack
2010	Waledac	Email Spam
2011	GOZeus	Steal bank log-in credentials
2012	Zitmo	Steal information from mobile devices
2013 ⋮	Wordpress; Lecpetex	Brute-force crack administrative credentials; Socialbot- 50,000 Facebook accounts were affected

Botnet detection can be done in two ways either by using signature based or anomaly based techniques. signature based techniques detect the presence of the known bots accurately but it is not possible for this approach to detect evolving new types of bots whereas anomaly based detection techniques are popular as its capability to detect the bot variants, even unknown bots.

It can be concluded from the previous studies that Botnet connection has specific network flow characteristics in which frequent communication happens between C&C and infected machine and hence to detect anomaly attack, network flow analysis is the best approach.

In this paper, botnet detection method is designed to detect bots by analyzing the captured network traffic. The contribution of the paper is divided into two parts. In First part, preprocessing the network traffic and extraction of features form the traffic that are necessary for detecting the bots are done. In second part, Fuzzy C-Means clustering algorithm is implemented in storm tool to cluster the dataset into malicious (bots) and non-malicious traffic, and then these trained dataset is used to cluster the test dataset which is feed as input to the storm. Experimental results shows that the proposed bot detection system has good detection rates and it is capable of detect the various types of bots including IRC, HTTP, P2P and DNS bots. The rest of this paper is formulated as follows. Section II deals withthe related work. Section III consists of the explanation about the proposed work. Section VI shows the implementation of the system. Section V presents the experiment environment and results. Finnally, conclusion is given in the Section VI.

2 RELATED WORK

IRC Bot were the first bot designed for legal activity but later on the hackers begin to create bots for performing illegal task to gain their own benefits. Hence it is imperative to detect and prevent the bots in order to avoid the malicious behaviors both in the host and network.

An adaptive blacklist-based packet filter using a statistic-based approach was developed by Meng et al. (2014) [2] to improve the performance of a signature-based NIDS. Two main parts of the filter were blacklist packet filter and a monitor engine. The blacklist packet filter removes the network packets by comparing payloads of the packets with signatures that were stored in the database. A monitor engine's tasks were monitoring the NIDS, collecting the statistical data, calculating the confidences of IP addresses and updating the blacklist in a fixed time. Advantages of this approach are reduction of the processing time, defend against IP spoofing and adaptive to the real context.

A behavior-based botnet detection system based on fuzzy pattern recognition techniques proposed by Wang et al. (2011) [3] to identify bot-relevant domain names and IP addresses by analysing network traces. The Domain names and IP addresses used by bots were identified then that information can be further used to prevent protected hosts from becoming one member of a botnet. DNS phase and the network flow phase were the two phases of this system. Detection of bots based on DNS features and bots based on network flow features were carried out in DNS and network flow phases respectively. This algorithm has the ability to find inactive bots.

The botmaster can control and send commands to the bots to perform attacks or launch more infections through the communication channel. Chen et al. (2014) [4], deployed anomaly score based botnet detection technique to find the bots activities by analysing the similarity measurement and the periodic characteristics of botnets. The proposed system employs two-level correlation to improve the detection rate. This system can separate the malicious network traffic generated by infected hosts from the normal IRC clients by analysing the features such as Source IP Address, Destination IP Address, Source Port, Destination Port, Timestamp, Payload of IRC Traffic and also it has ability to find the IRC bots at the communication stage.

AsSadhan et al. (2014) [5], proposed a detection method which analyze C2 traffic and find that it exhibits a periodic behaviour. To detect botnet C2 communication channels traffic, applied discrete time series analysis to examine the aggregate traffic behaviour. The detection involves evaluating the periodogram of the monitored traffic. Then applying Walker's large sample test to the periodogram's maximum ordinate in order to determine if it is due to a periodic component or not. If the periodogram of the monitored traffic contains a periodic component, then it is highly likely that it is due to a bots' C2 traffic. The test looks only at aggregate control plane traffic behaviour, which makes it more scalable than techniques that involve deep packet inspection (DPI) or tracking the communication flows of different hosts.

Network traffic monitoring and analysis-related research has struggled to scale for huge amounts of data in real time. A scalable real time intrusion detection system using the open source tools like Hadoop, Hive and Mahout designed by Kamaldeep et al. (2014) [6] . This implementation is used to detect Peer-to-Peer Botnet attacks using Random Forest machine learning algorithm. Libpcap, Hadoop, MapReduce and Mahout technologies were used in this framework. The Random Forest Algorithm was chosen to achieve high accuracy of prediction and it has the ability to handle diverse bots.
Modern botnet tend to be stealthier in the way they perform

malicious activities, making current detection approaches ineffective. Junjie Zhang [6], proposed a novel scalable botnet detection system capable of detecting stealthy P2P botnet. The detection system divided into two phases in which detection of all hosts within the monitored network that engage in P2P communications is done in first phase then the traffic generated by the P2P clients were analysed and classifies them into either legitimate P2P clients or P2P bots by using clustering algorithm in second phase.

3 Dection of Botnet

The main goal of the proposed detection system is to identify all types of bots by analyzing the network traffic behavior. The proposed system consists of two parts. In First part, preprocessing the network traffic and extraction of features form the traffic that are necessary for detecting the bots are done. In second part, Fuzzy C-Means clustering algorithm is implemented in storm tool to cluster the dataset into malicious (bots) and non-malicious traffic, and then these trained dataset is used to cluster the test dataset which is feed as input to the storm. The proposed system architecture is shown in the Fig. 2

Most of the botnet detection is based on payload analysis methods which inspect the contents of TCP and UDP packets for malicious signatures. Payload inspection involves very high identification accuracy when comparing other approaches but it violates the privacy. Furthermore, new bots utilize encryption and other methods to hide their communication and defeat packet inspection techniques. Hence Traffic analysis exploits the idea that bots within a botnet typically have uniformity of traffic behavior then these behaviors may be analyzed and cluster using a set of features which differentiate malicious and non-malicious traffic. Traffic analysis does not depend on the content of the packets and is therefore unaffected by encryption.

In preprocessing, the captured network traffic of malicious and non-malicious are merged and saved as pcap(packet capture) file. Then data cleaning and normalization are done. In a large subset of features, the feature subset that is necessary for detection of bots must be identified because some irrelevant features may exist. If all these unnecessary features are considered then the detection rate will get decreases. Feature selection is used for identifying the necessary features of a large amount of multidimensional data.

There are eight features namely Source IP address, Source port address, Destination IP address, Destination port address, Protocol, Time, Duration, The number of packets transferred in both forward and backward directions that are essential for clustering the given data set into malicious and non-malicious data.

Fuzzy C-Means clustering algorithm is implemented in Storm to cluster the given dataset into malicious and non-malicious data. These trained information is feed to Storm so that, it can detect the bot if any present in the test dataset.

4 Implementation

Storm is a reliable, fault tolerant and distributed system for processing streams of data. It consists of different types of components and each component is responsible for a simple specific processing task. The input stream for Storm cluster is handled by spout component. The spout passes the data to another component called bolt, which transforms the given data in to some specified form. A bolt may store the data once the transformation done or passes it to some other bolt for further changes. In short, a Storm cluster as a chain of bolt components that each make some kind of transformation on the data exposed by the spout.

Storm process unbounded streams of data easily, and it also used for real-time processing. Storm can be used with any programming language. Implementation of machine learning algorithm with Storm is possible. There are three approaches for such implementation are Storm-Pattern, Storm-R, and Trident-ML. Here Storm-pattern approach is used. Storm-Pattern: Pattern is a port of the Cascading/Pattern project. It aims to support the operational deployment of all of the most common models, imported via PMML(Predictive Model Markup Language).

Installing Storm package in Ubuntu operating system is the first step of implementation and then Fuzzy c-means algorithm is employed in that storm tool.

Algorithm: Fuzzy c-means clustering
Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of cluster centers.
1. Randomly choose '*c*' cluster centers.
2. Compute the fuzzy membership 'μ_{ij}'
3. Calculate the fuzzy centers 'v_j'
Vj-jth center.
c-number of cluster centers.
µij-membership of ith data to jth cluster center.
4. Repeat step 2 and 3 until the minimum 'J' value is obtained.

Fuzzy c-means algorithm gives best result for overlapped data set and comparatively better than k-means algorithm.

5 Experimental Result

This experimental result shows that the proposed detection approach has a good performance. Clustering of malicious and Non-malicious represented in the Fig.3

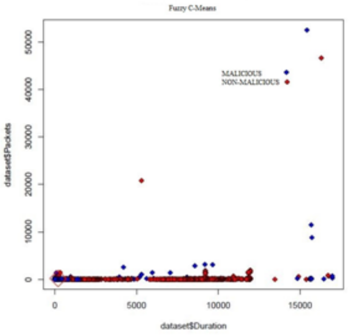


Fig.3 Fuzzy C-Means Clustering Graph

The proposed system is compared with the K-Means algorithm and it shows that the accuracy of fuzzy C-Means is higher than the K-Means algorithm.

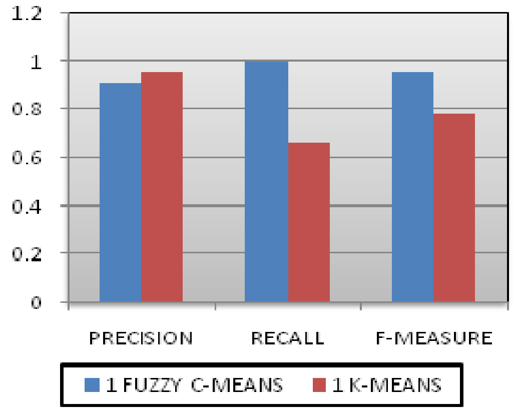


Fig. 4 Comparison of Fuzzy C-Means and K-Means for Dataset 1

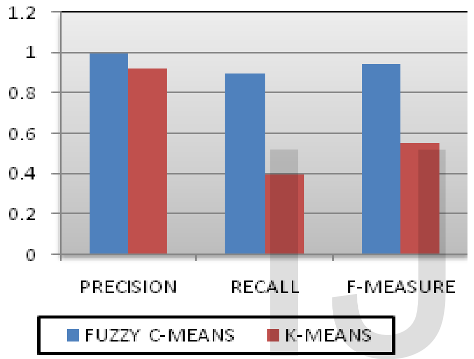


Fig. 5 Comparison of Fuzzy C-Means and K-Means for Dataset 2

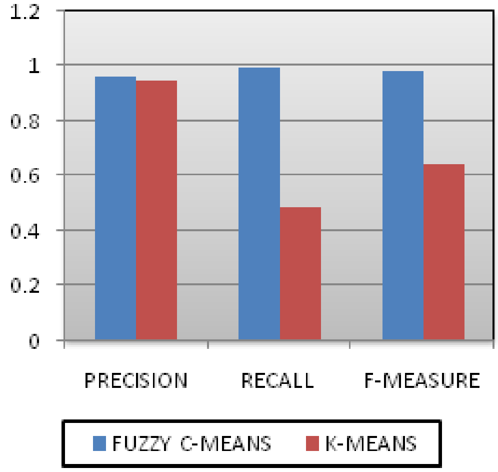


Fig. 6 Comparison of Fuzzy C-Means and K-Means for Dataset 3

In this work, three malicious datasets are downloaded and network traces captured using wireshark tool are taken as non-malicious traffic. Each malicious dataset is merged with the non-malicious traffic then the required features are extracted from these three datasets. The comparison between Fuzzy C-Means and K-Means for dataset 1 is shown in the Fig.4

Recall is the measure that referred to the true positive rate or sensitivity whereas precision is referred as positive predictive value. F-measure or F-score is the measure that combines precision and recall is the harmonic mean of precision and recall.

The comparison between Fuzzy C-Means and K-Means for dataset 2 is shown in the Fig.5

The comparison between Fuzzy C-Means and K-Means for dataset 3 is shown in the Fig.6

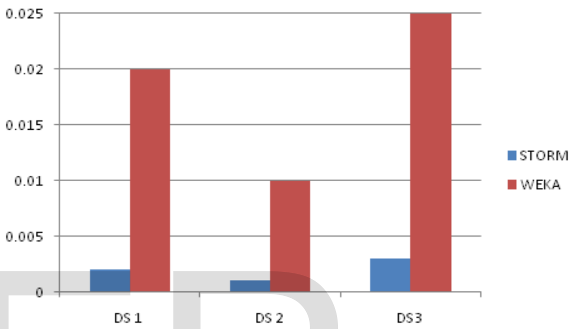


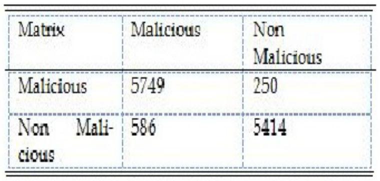
Fig.7 Comparison of execution time between Storm and Weka tool

From the above three graphs, it is well understood that Fuzzy C-Means has higher Recall and F-measure rate compared to K-Means algorithm.

Difference in execution time in seconds for three datasets among storm and weka is represented in the Fig.7 and the execution time in storm is less compared to weka tool.

Table 2 represents the Confusion matrix of Fuzzy C-Means for Dataset 1.

TABLE 2
CONFUSION MATRIX



Matrix	Malicious	Non Malicious
Malicious	5749	250
Non Malicious	586	5414

6 CONCLUSION

Botnet detection is a challenging task, since the creators of botnets continue to adopt innovative means in creating botnets. In this paper, Fuzzy C-Means clustering algorithm is implemented in Storm tool to detect bot. it is concluded that Fuzzy c-means algorithm shows high recall rate for all the three datasets compared to k-means algorithm which means returned most of the relevant results.

REFERENCES

[1] Sérgio S.C. Silva, Rodrigo M.P. Silva, Raquel C.G. Pinto, Ronaldo M. Salles,"Botnets: A survey", Computer Networks 57, 2013, pp 378–403.

[2] Yuxin Meng, Lam-ForKwok, "Adaptive blacklist-based packet filter with a statistic-based approach in network intrusion detection", Journal of Network and Computer Applications 39, 2014, pp. 83–92.

[3] Kuochen Wang, Chun-Ying Huang, Shang-Jyh Lin, Ying-Dar Lin, "A fuzzy pattern-based filtering algorithm for botnet detection", Computer Networks 55, 2011, pp. 3275–3286.

[4] Chia-Mei Chen, Hsiao-Chung Lin, "Detecting botnet by anomalous traffic", journal of information security and applications, 2014 pp. 1–10.

[5] Basil AsSadhan, Jose´ M.F. Moura, "An efficient method to detect periodic behavior in botnet traffic by analyzing control plane traffic", Journal of Advanced Research 5, 2014, pp. 435-448.

[6] Kamaldeep Singh, Sharath Chandra Guntuku, Abhishek Thakur, Chittaranjan Hota, "Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests", Information Sciences 278, 2014, pp. 488–497.

[7] Junjie Zhang, Roberto Perdisci, Wenke Lee, Xiapu Luo, and Unum Sarfraz, "Building a Scalable System for Stealthy P2P-Botnet Detection", IEEE transactions on information forensics and security, vol. 9, no. 1, 2014.

[8] Chun-Ying Huang, "Effective bot host detection based on network failure models", Computer Networks 57, 2013, pp. 514–525.

[9] Seungwon Shin, Zhaoyan Xu, Guofei Gu, "EFFORT: A new host–network cooperated framework for efficient and effective bot malware detection", Computer Networks 57, 2013, pp. 2628–2642.

[10] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, Matei Ripeanu, "Design and analysis of a social botnet", Computer Networks 57, 2013, pp. 556–578.

[11] Weizhang Ruana, Ying Liub, Renliang Zhaob, "Pattern Discovery in DNS Query Traffic", Procedia Computer Science 17, 2013, pp. 80 – 87.