A Reliable & Scalable Frame Work for HTTP BotNet Detection

Dr.R.Kannan ,
Associate Professor, Department of Computerscience ,Sri Ramakrishna Mission Vidyalaya College of arts and science
Mrs.Poongodi
Department of Computerscience ,
Sri Ramakrishna Mission Vidhyalaya College of arts and science,
Coimbatore
Tamilnadu, India

Abstract- *With growing number of internet based applications, smart phones and mobile computing devices to connect such applications has increase the use of internet technology. The growing network capabilities enable distributing and resources sharing even among cross platform devices. Parallel processing enables to utilize resources exceeding more than one machine. Exploiting such development's and computational freedom, attacker's use botnets for various tasks including data stealing, denial of service, and other illegal activities. To detect botnets from regular activity in the network involves distinguishing regular traffic and botnet activity. The resilient nature of botnets can't be predicted with regular time intervals and the activity may resume at any given time. This paper aims to classify the patterns of the botnets and to mitigate the effects of Botnet through detection and prevention framework proposed. The framework classifies the malicious activity using information mining methods to distinguish internet traffic from malicious traffic, once the traffic patterns are identified; future patters can be identified to remove botnets from the network.*

Keywords:*Botnet,SVMhyperplane,Clustering,BDoS,IRC andC&C*

I. INTRODUCTION

Botnet comprises of interconnected computers, servers, mobile devices and other network components. Botnet is a sophisticated program that can perform autonomous work of sending and receiving information. Botnet are highly equipped to work without any traces of its presence. Over the years, botnets are evolved to infect, involve in phishing attacks, stealing on data, behavior, adware and denial of services. The present information sharing through social networks have enabled botnets to exploit numbers machines silently for monetary benefits. The major types of botnet activity include spam, Trojan, DDoS which are primarily targeted for information and personal identity, as millions of data are shared over internet in a second.

The activity of botnet is similar to that of regular network behavior such that to distinguish between regular traffic and botnet traffic has become difficult. But with its unique network behavior, communication botnets can be easily differentiated through their network characteristics. To enhance the network security, performance and communication it is important to detect, classify and remove the botnets from the network. Generally botnets are controlled by command & control centers which are termed as botmasters. There are host-based botnets and network based botnets. Network based botnets are primarily located in the network snooping the network traffic with incoming packets. Host based botnets are present in the machines in the registry, files and process.

The botnets have become more advanced and gain capabilities to show quality of polymorphism, rootkit, and many other qualities to avoid detection. In order to detect the traffic patterns of botnet and their characteristics classification of network data is required to find the variants of traffic characteristics exhibited by the botnets. But the traffic characteristics are independent to the network and it is complex to indentify the traffic flow of each network type. Generally the botnet detections are confined to a particular network type and the emerging patterns are distinguished as botnets to that particular network. However there are some common botnet variants that are capable of spreading across different networks. These botnets when left undetected pose a serious threat to network infrastructures, machines connected over the network and more over they take down the network communication and traffic data as they are capable of disguising their identity.

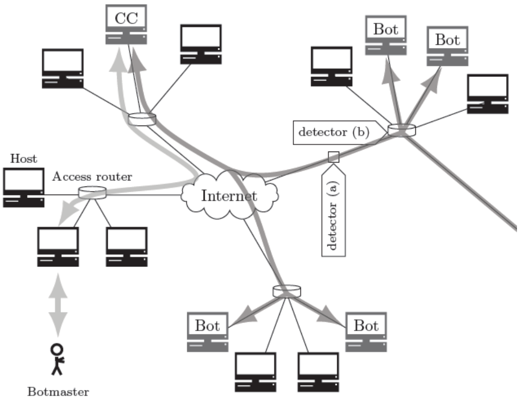


Fig 1 Botmaster, C&C and communication channels

The botnet command and control centers (C&C) follow a typical pattern where a botmaster send a command to the control center, control center takes up the command from the botmaster and acts accordingly by forwarding the command to the botnet and finally botnet sends back to the botmaster it details of activity. There are three types of communication takes place by the C&C namely relay based chat (IRC), HTTP-based and P2P, where the C&C pushes commands to the botnet in the IRC model, HTTP-based utilizes a regular server requests from bots for new commands and P2P uses a placeholder type to communicate the control centre. The main advantage of network based botnet detection is that the network based detection uses the inter relation among the patterns. They do not require much information about botnet.

II RELEATED WORKS

(Basil & Morab, 2013) analyzed command and control communications traffic between the members which exhibit a periodic behavior. They applied Walkers large sample test to find out the periodic components of the traffic. By using aggregated control panel traffic to trace the communication between different hosts.

(Fedynyshyn, Chuah, Tan, 2011) presented a host based method to distinguish the different types of botnet types and the communication they involve with command and control units. They achieved a overall accuracy of 0.929 in classifying the C&C channels. By classifying they were able to distinguish between original network traffic and C&C network traffic.

(Kwon, Lee, Lee, 2011) investigated botnets by investigating the attack host by tracking the non-human attack types. Their proposed mechanism detects the bots and distinguishes from human attacks. They studied the botnets using their structural characteristics and once identified, they halt the botnet traffic at the host level and Achieved detecting bots that are involved with flooding attacks and spam attacks.

(Gu, Zhang & Lee 2008) introduced a network based anomaly detection method. Their proposed technique detects C&C and infected hosts in the IRC and HTTP network. Using the spatio temporal similarity they proposed techniques detect the communication between the botnets and C&C servers. Their system identified DDos, fake communications, botnet multiplication through correlation within their network properties.

(Tegeler, 2012) suggested a system that can find botnets present in the network. They system can detect communications in the network with respect to their trained communication pattern. They used flow based characteristics to exploit the botnet behavior by arriving an average flow rates which are transmitted between the source and the destination. Using Fourier transformation, they improved the accuracy of detecting botnet and C&C actions which detects both IRC and HTTP channels

(Zang, 2011) proposed a flow base classification method to identify the botnets using RTT of the data packets. He identified the flow rate using clustering algorithm and found that the difficulty in achieving a common system to capture the botnets that are available in the network.

(Wang, Huang, Lin, & Lin, 2011) introduced a method based on fuzzy pattern recognition to recognize the behaviors of botnets. By auditing the network flow traces, they intend to detect botnet related domain names and their IP address. Their system used the similarity functions in failed network connection, failed DNS queries, DNS interval and payload sizes to differentiate the infected network flows.

(Lu, 2011) studied on a system that studies the patterns of the network communication. Using traffic payload, they classify the network traffic using payload and flow properties. For classification, they filtered features using n-gram feature selection method and grouped each of the traffic flows corresponding to each application. The improved the detection rate using traffic payload.

(Garg, 2013) showed data mining algorithms such as Naïve bayes, J48 and K nearest to detect p2p botnets using network traffic features and found that J48 accuracy was good at classifying however, the classification accuracy cannot detect in real time detection due to lack of traffic properties.

(Junjie, 2011) proposed a system to detect botnet that were sneaky. Using bytes sent, bytes received, packets received, packets sent between the source and the destination, his study pointed in traffic supervising with C&C. The features selected were able to detect the botnets but this method suffers from huge network traffic data.

(Wen-Hwa and Chia-Ching, 2010) used packet size features to differentiate between botnet traffic and authorized traffic. They lured botnets by sending information's and botnets which are transmitting data to a minimum threshold were classified accurately using Naive Bayes and J48 algorithms. They achieved a classification accuracy of about 85% in their study. In their study they found that the size of the packets was very small when compared to the overall network packet size in the P2P network.

(Nogueira, 2010) developed a neural network based traffic recognition to distinguish the real traffic and botnet traffic. Their system was robust in detecting botnets even in the encrypted mode and has various advantages in splitting the traffic flows with rigorous training. They were able to achieve an accuracy of 87% on the network traffic.

(Saad, 2011) proved that machine learning algorithms are potential enough to detect the botnet traffic. Using machine learning methods, he achieved an accuracy of 89% and showed that ANN and SVM have huge potential but higher training period and traffic volume. But Machine learning environments required high computational and processing power than other models and more over have restrictions to certain kind of botnets features.

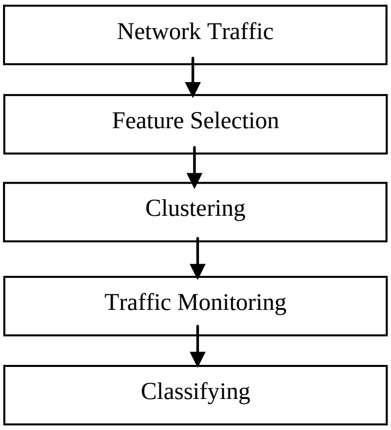
III RESEARCH PROPOSED

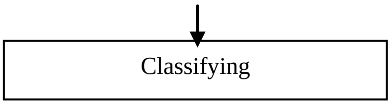
Increase of attacks and exploitation of data demands a strong intrusion detection system with detection and prevention capabilities. This study focuses on learning and preventing intrusions on internet based attacks. To detect and filter attacks, a framework is developed to learn and detect intrusions, analyze patterns of existence of botnets that are present in the internet communications. Using data mining techniques, the frame work captures network traffic data, selects the best features using feature selection method and classifies according to the classified network behaviors using SVM. The classification is done with respect to network flow characteristics present in UNSW_NB15 dataset and the patterns of such network flows are detected with respect to the type of malicious activity and are removed from the communication network.

Our proposed framework is reliable and scalable to detect HTTP based Botnets. The frame work consists of components such as Filtering, Traffic clustering, and Traffic Classifier.

Data set description:

The UNSW_NB15 dataset consist of 43 features of Source IP address, Source port number, Destination IP address, Destination port number, Transaction protocol, Record total duration, Source to destination transaction bytes, Destination to source transaction bytes, Source to destination time to live value, Destination to source time to live value, Source packets retransmitted or dropped, Destination packets retransmitted or dropped, http, ftp, smtp, ssh, dns, ftp-data ,irc Source bits per second, Destination bits per second, Source to destination packet count, Destination to source packet count, Source TCP window advertisement value, Destination TCP window advertisement value, Source TCP base sequence number, Destination TCP base sequence number, Mean of the flow packet size transmitted by the src, Mean of the flow packet size transmitted by the dst, Represents the pipelined depth into the connection, data transferred from the server's http service, Source jitter (mSec), Destination jitter (mSec), record start time, record last time, Source interpacket arrival time (mSec), Destination interpacket arrival time (mSec), TCP connection setup round-trip time, TCP connection setup time, TCP connection setup time, If source (1) and destination (3)IP address, No. for each state (6) according to specific range of value, No. of flows that has methods such as Get and Post in http service, If the ftp session is accessed by user and password then 1 else 0, No of flows that has a command in ftp session, No. of connections that contain the same service (14), No. of connections that contain the same service (14), No. of connections of the same destination address (3), No. of connections of the same source address (1), No of connections of the same source address (1), No of connections of the same destination address (3) and the source, No of connections of the same source (1) and the destination (3) and the name of each attack category, 0 for normal and 1 for attack records.





I. Feature Selection

The feature selection is significant in high dimensional data. The feature selection and the quality of each features determines the accuracy of classification. The features are the predictors of the classes, there are 3 main reasons for a subset creation. Reducing the computational complexity, minimizing the training time and to derive the predicting power of the features. There are different feature selection methods available and of all the methods, each method of feature selection is distinct in its own merits and suits for different purposes and merits. The main groups of feature selection are wrapper, embedded and filter methods.

Wrapper method: this method subsets the features based on the explaining power and eliminates the features either backward or forward steps. Each feature strength depends on the each eliminations as the explain power changes with each feature removal

Embedded method: This method is a kind of search algorithm which treats the dependent feature as stimulus and search for best feature results.

Filter method: This method uses the relationship between the features as its strength for a given threshold. Correlation based filtering looks for high correlation with the dependent variable and removes features with low correlation. Regardless of the model, the features are eliminated.

We introduce a novel feature selection method based on permutation and the degree of variation using trees. The features are ranked based on the decreasing level of their variations in a dataset. The ranks are calculated through creating noises and each feature is permutated until their variations completely drained. The low variation features are removed and the higher variations are retained. The resulting features are created with a subset and finally the subset is applied for classification.

II. Clustering of Features (Monitoring)

In order to distinguish the real time traffic data from the botnets, clustering of each network flow is done using Kmeans clustering technique. This is to identify the variations within each botnet type. These variations are mapped to network characteristics and any newer pattern emerges, it can be used for clustering to the particular type of botnet. The classification of network flow is discussed in the next section. The frame work thus builds a network map based on network flow characteristics. Each time a new pattern emerges, the traffic points are added to the classification algorithm for training purposes. The training of newer patterns can help classifier to stay updated with the emerging newer patterns.

III. Classification

The feature selected subset is then passed on the SVM classifier. SVM is a kind of model used to classify data. Support Vector Machine is a promising system for learning machine mainly employed for two class problems. SVM utilizes both linear and nonlinear kernel functions and sorts out the entropy by finding the hyperplane. The hyper plane is the line that separates the data points between two classes. If a large margin is found then the model would be better.

The Support Vector Machine uses the linear classifier of the following form,

$$F(x) = WI + bias$$

Where W = weight factor, I = input vector and bias. The hyperplane which divides is defined by $f(x) = 0$. Therefore first class that falls above the hyperplane has $f(x) > 0$ and another class below the plane is $f(x) < 0$.

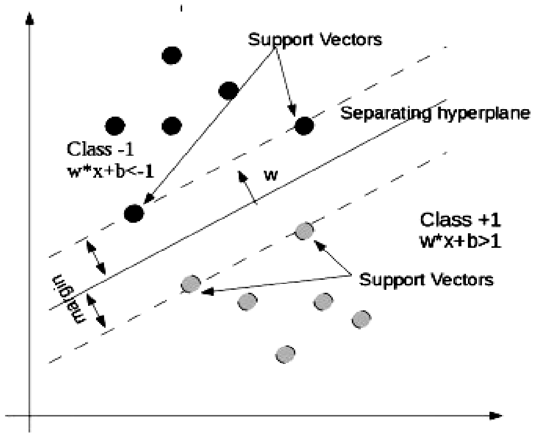


Fig 2 Svm hyperplane

IV CONCLUSION

This paper discusses a new proposed framework for botnet classification and detection. The frame work utilizes feature subset for dimensionality reduction; the features selected are then clustered to identify the newer patterns. The newer patterns in the network flow are grouped according to the type of activity it performs. The clustering of network flow is used for monitoring of traffic flows, and once a new pattern is registered, it is grouped according to the flow characteristics. The frame work processes the network flow in the order that it first extracts the features that are most relevant and then the classifier using SVM classifies the traffic flow either as attack category or normal. Once it is found to be an attack category, the network traffic data is clustered to appropriate attack types for future monitoring. Thus our frame work can be adapted to real time monitoring of traffic data for botnet activity based on the classified attack types.

REFERENCES

[1] (Basil,Mourab, 2013) An efficient method to detect periodic behavior in botnet traffic by analyzing control plane traffic, Journal of Advanced Research, Volume 5, Issue 4, July 2014, Pages 435-448,

[2] Fedynyshyn G., Chuah M.C., Tan G. (2011) Detection and Classification of Different Botnet C&C Channels. In: Calero J.M.A., Yang L.T., Mármol F.G., García Villalba L.J., Li A.X., Wang Y. (eds) Autonomic and Trusted Computing. ATC 2011. Lecture Notes in Computer Science, vol 6906. Springer, Berlin, Heidelberg

[3] J. Kwon, J. Lee, H. Lee, Hidden bot detection by tracing nonhuman generated traffic at the Zombie host, in: Information Security Practice and Experience, Springer, 2011, pp. 343–361.

[[4] Gu, G, Zhang, J & Lee, W 2008, 'BotSniffer:Detecting Botnet Command and Control Channels in Network Traffic', paper presented to 15th Annual Network and Distributed System Security Symposium (NDSS'08), San Diego, CA.

[5] Tegeler, F, Fu, X, Vigna, G & Kruegel, C 2012, BotFinder: finding bots in network traffic without deep packet inspection, ACM, Nice, France.

[6] Zang, X, Tang pong, A, Kesidis, G & Miller, DJ 2011, Botnet Detection Through Fine Flow Classification, The Pennsylvania State University, University Park, PA, 168

[7] Wang, K., Huang, C.-Y., Lin, S.-J., & Lin, Y.-D. (2011). A fuzzy pattern-based filtering algorithm for botnet detection. Computer Networks, (55), 3275-3286.

[8] Lu, W., Rammidi, G., & Ghorbani, A. A. (2011). Clustering botnet communication traffic based on n-gram feature selection. Computer Communications, 34, 502-514.

[9] Garg, S., Singh, A. K., Sarje, A. K., & Peddoju, S. K. (2013). Behaviour analysis of machine learning algorithms for detecting P2P botnets. Paper presented at the 15th International Conference on Advanced Computing Technologies (ICACT).

[10] Junjie, Z., Perdisci, R., Wenke, L., Sarfraz, U., & Xiapu, L. (2011). Detecting stealthy P2P botnets using statistical traffic fingerprints. Paper presented at the IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN), Hong Kong.

[11] Wen-Hwa, L., & Chia-Ching, C. (2010). Peer to Peer Botnet Detection Using Data Mining Scheme. Paper presented at the the international Conference on Internet Technology and applications, Wuhan, China.

[12] Nogueira, A., Salvador, P., & Blessa, F. (2010). A Botnet Detection System Based on Neural Networks. Paper presented at the Fifth International Conference on digital Telecommunications (ICDT), Athens, TBD, Greece.

[13] Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., Felix, J., Hakimian, P. (2011). Detecting P2P botnets through network behavior analysis and machine learning. Paper presented at the Ninth Annual International Conference on Privacy, Security and Trust (PST), Montreal, QC.