Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

A HOST–BASED P2P HOST IDENTIFICATION APPROACH WITH FLOW–BASED IN DETECTION OF P2P BOTS

GADDE RAMESH[1] and SURESH PABBOJU[2]

[1]Research Scholar, University College of Engineering (Autonomous), Osmania University, Hyderabad-500 007, Telangana State, India.
[2]Professor of Information Technology, Chaitanya Bharathi Institute of Technology-CBIT, Hyderabad-500075, Telangana State, India.

Abstract- In today's era of Internet 59.5% people globally are connected to Internet. Few utilize for entertainment, banking, communicating, information retrieving. Some others utilize for business purpose by interacting with the suppliers, distributors, partners, customers etc., with such an abundant internet usage the criminals lead to the era of Cyber Crimes. As per PurpleSec, cyber-crimes have risen to 600% due to COVID – 19 pandemic as remote working has increased, and the security levels needed to an individual's system may not be available at home. Though various cyber-attacks are happening, one of the major attacks via Bot is buzzing in the stream of cybercrime. The significance of the botnets made the researchers work on them and approaches to assuage them. Botnet's new architecture called Peer – to – Peer (P2P) has made strong against detection over conventional client – server. Due to the superior resiliency next to detection, P2P botnets reputation commenced growing namely ZBot/Zeus, the largest botnet globally estimated to mark 3.6 million PCs [1]. This newly proposed approach is a 2 – step process, initially the engaged hosts in P2P activity are identified and later detects P2P botnets using PeerClear methodology. This methodology works with an accuracy of 99.6% and low false – positive rate < 0.28%.

Keywords: Botnet; Bot; P2P bots, P2P Host; Network flow; Peer Clear; P2P Network.

1. INTRODUCTION

Bots are afflicted systems operated by an intruder, and the botnet is a chain of these malicious machines. Botnet assaults account for a significant proportion of cyber warfare. Bot spyware shifts the machine into yet another robot that does tasks relying on orders delivered to it via the Web. Botnets really had no limitations besides the harmony; they are meant to create to machines even if expected to remain anonymized [7]. Bandwidth, Diversity in IP, Computational power, and few more are the various resources Botnet provides for the attackers. 25 percent of computer systems globally are piece among one or more botnets [4] as stated by Vinton Gray Cerf, the "Father of the Internet".

Botnet run automatically and autonomously in the switch of a distinct or organization identified as BotMaster or Botherder [11]. Distinct channels are used by Botmaster called Command & Control (C & C) to connect with the bots.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

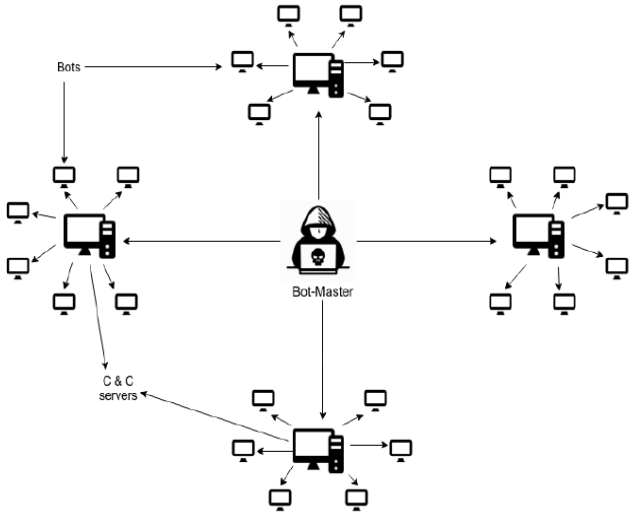

Figure 1: Graphical outline of a botnet [10]

Botnets with small domain attack, botmaster can connect to bots directly but, bots with larger domain attack botherder connect to the bots via Command & Control Servers (C&C Servers are the intermediate hosts) to hide the bots identification. Thus, detecting forms of botnet by chasing C & C servers is always a hot cake for scholars. Bot throughout its life undergoes various segments namely S1 to S6, where S1 is initial in infection, examines for the compatibility and tries to take advantage of vulnerabilities or susceptibilities. After few effective exploits, bot not only informs the botherder about the compromised bot via C & C network but also the target's information like its abilities, backdoor data and many more. This activity is categorized as second segment. In the third, to make bot full functional it downloads new binary files which holds full commands but are executed in fourth segment by receiving the instructions from the servers to execute the commands. The results are reverted to the server by the bot in the last segment.

Botnet undergoes Propagation and Response phases. Initially, botnet malware is generated, and then the botnet starts growing in the Propagation phase, establishing C & C Servers, and attempting to recruit bots by extending botnet malware to invade vulnerable systems. Following the establishment of the botnet, the bots await instructions from the C & C Servers. During the Response phase, the botnet initiates to strike targets, with C & C servers transferring target information and list of command to the bots till everything is sensed by security software or ISPs.

When users initiate sharing their system resources to avail the services proposed, that distributed network architecture can be named as P2P network. The resources and their services can be directly utilized by other peers and the users here are therefore not only resource requesters but also resource providers. The existence of servants (a host

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

which is both client and server at the same time) differentiate among client – server networking and P2P networking. The latter implements virtual overlay network over the physical network topology and nodes form subset of nodes. Data is directly traded over TCP / IP network, but peers connect directly at Application layer by means of logical overlay links. Pure and Hybrid are 2 P2P networks which differentiates P2P networking to those with a central entity and to those without.

Pure P2P network = P2P network + any node randomly removed from the network should not exhibit any network misery or any failure of network service.

Hybrid P2P network = P2P network + a key entity to support elements of the recommended network services.

Command and control channel plays a key role in preventing the botnet from dismantling. Push – based and Pull – based are the approaches to establish communication among C & C Servers and bots or among the 2 bots. In the first approach, the botmaster pushes the commands into bots that are to be executed whereas in the second approach the bots obtain instructions occasionally from the server. In push - based approach the bothered can perform tasks instantly but pull - based have a random delay while providing the commands. High control over bots by botmasters is in the first case and control over generated traffic is in the second. As the amount of traffic generated traffic high in push – based, this leads in easier detection of bots as any 2 bots contaminated by the identical botnet have related traffic. As botnet has control over traffic in pull – based, this approach thwarts simpler recognition of bots and servers.

a) Internet Relay Chat protocol used by IRC Botnet handles botnets and facilitates real – time conversations. As IRC is client – server model, IRC servers became SPOF (Single Point of Failure) and can be found effortlessly.
b) As HTTP is mostly used in internet, HTTP botnets or Web based Botnets are trickier to detect and server – side languages are used to implement C & C Servers as web apps. Bots send HTTP requests to servers and in turn receive commands as a reply.
c) P2P Botnets eliminate or reduce the SPOF of the (a) and (b). In this case, all bots connect and share files among themselves constructing a network. At any point P2P botnets allow acceptance of authenticated instructions by using public key cryptography to prevent unauthorized instructions to enter the network.

2. LITERATURE SURVEY

The construction of the networks on the above of another network is referred as Overlay networks, for instance Application layer on the top of TCP / IP network which suggest self –organizing that offer effective routing structures. Reliability, redundancy, fault–tolerances, scalability, load – balancing are the assets of the overlay networking without altering the core IP network. Unstructured and Structured P2P overlay networks are the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

classifications. Unstructured arrange peers as a graph or in a hierarchical structure and structured P2P overlay networks forms Distributed Hash Tables (DHT). Gnutella and Bit – torrent are the applications of the former and Pastry, Chord, Content Addressable Network (CAN) are for the latter. The unstructured is flexible, robust, easy to use and the structured usually uses O (log (n)) [8] as a shortcut to reach the destination.

2.1 Several detection methodologies by Botnet

The impact of botnets made researchers come with innovative blends. Among them few analyze the botnets by static analytical techniques such as reverse engineering the bot binaries or by employing vibrant assessment using tools such as Cuckoo [3] or Norman Sandbox [6], few more use network sniffer to understand the network's traffic, and few others use Honeypots to witness malicious URLs.

1) In network – signature – based botnet detection, the network signature acts as pattern as an attempt of connection. This methodology extracts network – signatures of established botnets and are assessed against known botnet signatures. Until the signatures are updated frequently new botnets are not detected and this approach does not work on encrypted traffic.

(a) In network – behavior – based botnet detection, the discrimination among normal traffic and botnet traffic network behavior are studied. This approach can detect unknown botnets by comparing the behavior of the knowns. Higher false – positive rate is the major disadvantage here.

(b) Honeypot – based botnet detection is a fresher methodology to infiltrate and to detect botnets by using honeypots, where honeypots gather bot binaries. The gathered binaries are inspected with anti – virus or examined by static/dynamic evaluating techniques. The detected bot binaries execute in a supervised setting without destroying any target. This makes honeypot a bot by accepting instructions from C & C server and logs can be studied. The delay in analyzing and infiltrating the botnet is the main dropout thus attacks on logs too.

(c) Host - based botnet detection likely works as a traditional anti – virus software. Its observation over hosts, its related activities, logs help in predicting the occurrence of botnet. Before encryption, this approach can read the encrypted network traffic and places over the network [9] [12].

(d) Domain name system tracking is developed by Choi et.al [5] to track botnet DNS queries. DNS queries are executed in order to establish connection with C & C Server. These DNS queries are captured, and a blacklist of domain names of C & C server is prepared. To overcome, periodic change in domain name of the used server is done by botmaster and usage of own DNS servers in encouraged.

Flow – based, Node – based or host – based, conversation – based, resource sharing behavior monitoring are few P2P botnet detection techniques.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

3. PROPOSED SYSTEM

Here, we are proposing a fresh method for P2P botnet detection by introducing a 2 – step process. Initially we need to identify all the involved hosts in P2P activity by using host – based approach. In the second step, we need to detect P2P bots among P2P hosts by using flow – based approach which extracts inter – arrival time of packets, the frequency of packets and the count of bytes sent and received many more. These features are used for detecting P2P bots by building a classifier. At the end, both approaches are combined and constructed to have a GUI by entering IP address to detect P2P botnets or to retrieve he status of every host by monitoring the entire network.

Figure2 clearly mentions the 2 phases as P2P Host Detection and P2P Botnet Detection phases which are explained earlier. Every 10 minutes a host – based 14D feature vector is extracted periodically and is used to detect P2P host by training the classifier. A flow – based 18D feature vector is extracted every hour, used to train the classifier in detecting P2P botnet. The user enters the IP addresses to be tested and also saved in list of P2P detection. At initial steps, packets are taken from live network traffic and unrelated packets are truncated by packet filter to minimize the processing time and operating cost of the packets. Assume output of packet filter module as Panalyse. Later, features relevant to P2P hosts are extracted by feature extraction. Ten minutes time window is set for phase 1 and all the packets except Panalyse are examined if source or destination IPs match with any host IP. If matched packets are found, then 14D feature vector (FP2P) is retrieved for every host. Now, feature selection chooses optimal features (F0P2P) out of FP2P and ignores the remaining. F0P2P is given as an input to the classifier. Using P2P traffic and non – traffic samples Decision Tree classifier is trained to predict on F0P2P. If the classifier detects F0P2P as non-P2P then it is rechecked using same phase in successive time window else host IP is separated from the list of P2P detection and included to P2P botnet detection list.

If a host is identified as P2P host Phase 2 starts and a time window of 60 minutes is used here. Flow – based features are retrieved for all the IPs in P2P botnet detection list. Single flow has IPs and port number of source and destination and extracts 18D feature vector (Fbot) for time window of 60 minutes. Fbot as input to the feature selection 14 optimal features (F0bot) are extracted which helps for the classification. Classifier receives F0bot as input.

If F0bot is detected as P2P bot by botnet classifier, it is eliminated from the respective list and included to P2P botnet detected file else it restored for further checks. The IPs which are listed in P2P botnet detections are marked as Bot IPs.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

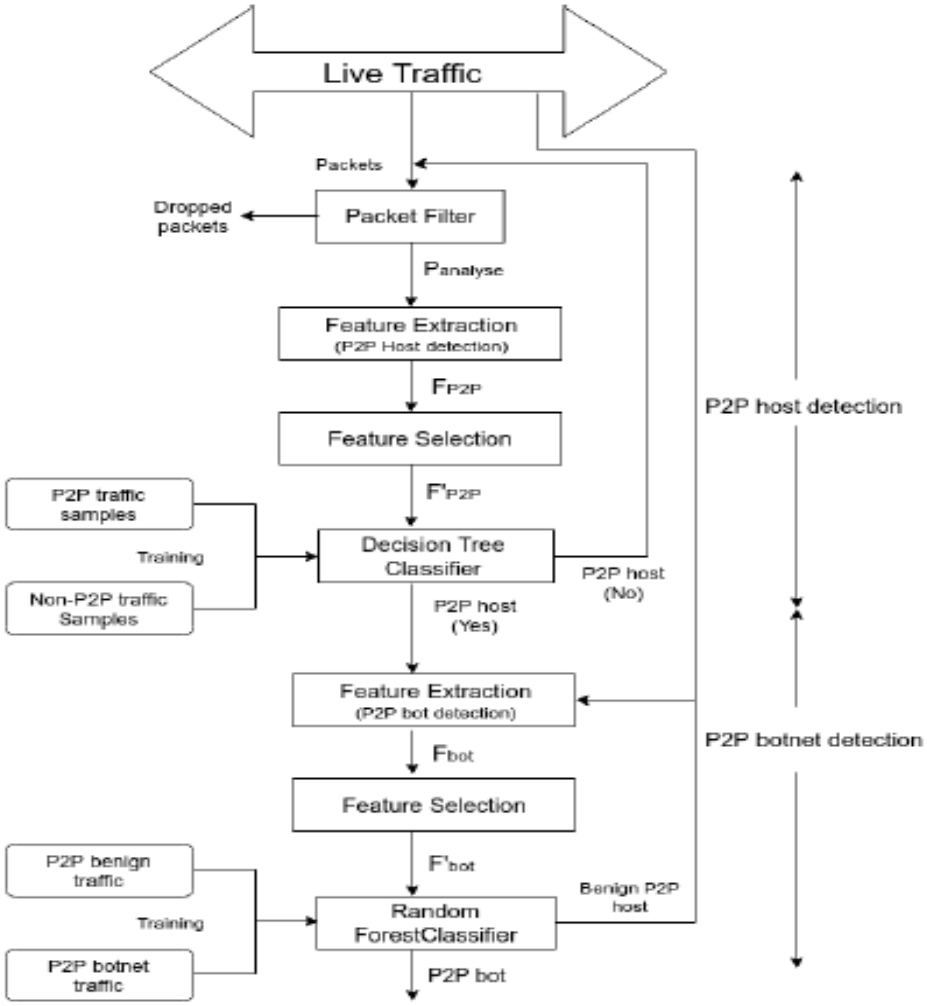

Figure 2: Block diagram of the proposed system

If a host is identified as P2P host Phase 2 starts and a time window of 60 minutes is used here. Flow – based features are retrieved for all the IPs in P2P botnet detection list. Single flow has IPs and port number of source and destination and extracts 18D feature vector (Fbot) for time window of 60 minutes. Fbot as input to the feature selection 14 optimal features (F0bot) are extracted which helps for the classification. Classifier receives F0bot as input.

If F0bot is detected as P2P bot by botnet classifier, it is eliminated from the respective list and included to P2P botnet detected file else it restored for further checks. The IPs which are listed in P2P botnet detections are marked as Bot IPs.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

3.1 How P2P Host Detection is carried?

The phase targets in detecting all the hosts involved in P2P activity with 4 modules namely Packet filter, Feature extraction, Feature selection and Classifier. Unwanted packets are streamed in packet filter module to minimize processing time and cost incurred for packet monitoring. To attain this target IPs list is retained settled by DNS queries. If at all DNS query results to a sniffed packet, we can retrieve from the packets the determined IPs, include them in the DNS list with the timestamp and the expired IPs are also deleted. If the packet is not of DNS type, then its source and destination IPs are compared with DNS list else packet is discarded and posted to feature extraction module.

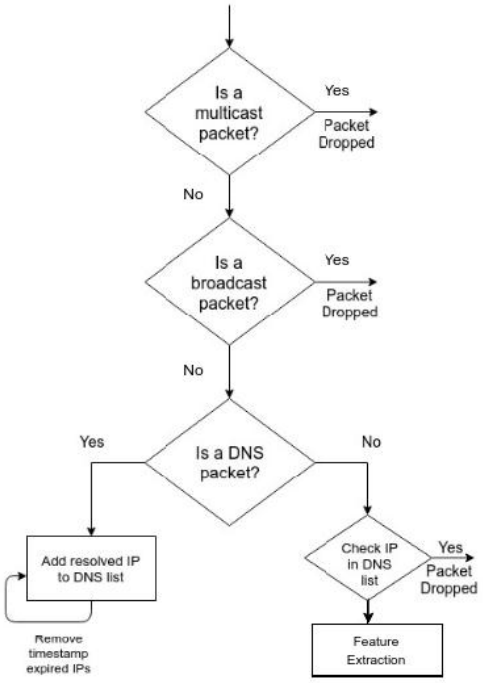

Figure 3: The process for Packet filtering

The diagrammatic representation for the process to explain the packet filtering is given in figure 3. Once packet filtering is performed, P2P host detection is an essential task that need to be done with the help of the features by identifying distinctive properties exhibited by hosts in P2P movement for instant peer – churn present in P2P network [10] as it exhibits good number of failed connections significantly. Few properties like failed connections, DNS filter, destination diversity which are forming the foundations of features selected are listed in Table 1.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

Table 1: Preferred network traffic features for P2P Host Detection

S. No	Feature	Description
1	F_1	Packet count of the retransmitted
2	F_2	Destination Diversity
3	F_3	Destination Diversity Ratio
4	F_4	Number of attempts made on diverse ports for connection
5	F_5	Number of contacted IPs (unique)
6	F_6	Number of reset packets
7	F_7	Count of packets (out of order)
8	F_8	Count of ICMP destination packets (unreachable)
9	F_9	Quantity of sent and received packets
10	F_{10}	(Number of Bytes)/packet in forward direction
11	F_{11}	(Number of Bytes)/packet in backward direction
12	F_{12}	Number of packets averagely retransmitted/host
13	F_{13}	Number of duplicate acknowledgement packets
14	F_{14}	Total quantity of sent and received control packets (packet without data)

To implement the process of extraction **tshark** is used from static pcap files which allows to capture or read or print packet data lively from the network. For instance, to observe the packets which are TCP transmitted from the file the tshark command is given as

tshark -r <<filename.pcap>> -Y <<NameOfTheDisplayFilter>>

Assume that we have applied 'tcp.analysis.retransmission' as display filter the feature 'F_1' becomes incremented and should have a 'label' as a feature vector which holds either '0' for non – P2P and '1' for P2P. Three distinct time windows of 5, 10 and 20 minutes are applied to extract the features and to locate ideal time window for P2P host detection. The format is represented as given below:

<F_1, F2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, label>

Assuming that among 14 employed features few are not efficient during the classifier training, few may not affect or worsen the performance of the classification. In our

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

proposal, we have considered and studied 'feature reduction technique' to diminish the vector size. In this proposal, Information Gain algorithm is utilized as a feature reduction measure for implementation of decision classifier where all becomes part of the construction of Decision tree. These measures Entropy, Information Gain are the top features selected for classification. We've using Decision Tree classifier for P2P host detection and efficiency is evaluated for our approach using python's Scikit – learn, a machine learning.

3.2 How P2P Botnet Detection is carried?

This segment aims to detect host similar to P2P bots from the recognized P2P hosts uses flow – based approach and the segment has 3 modules namely feature extraction, feature selection and classification techniques. To differentiate among gentle P2P traffic and botnet P2P traffic flow management is driven factor. As soon as botnet contaminates a host and to sustain the connection with other hosts, continuous control packets are sent by botnet which provide insights for communication. The protocol design decides the flow management and the data flow varies on the user usually legalizes interaction with P2P applications by the user. The separation of management flows from data flows is expected as these flows are usually implanted. Few parameters like inter packet time and duration of flow are believed for separation. But the features like host access features and flow size features are used to for P2P botnet detection. The former captures host accessing patterns of the botnets similar to packets minimum, maximum inter – arrival time and the latter captures the distribution of host's incoming and outgoing flows. The selected features are listed in Table 2.

Table 2: 18 preferred network traffic features for P2P Botnet Detection by using Pyshark

S.No	Feature	Description
1	F_1	Mean of the inter – arrival time among packets
2	F_2	Sent Count of packets in flow
3	F_3	Received Count of packets in flow
4	F_4	Sent count of bytes in flow
5	F_5	Received count of bytes in flow
6	F_6	Sum of count of data sent and received in flow including headers
7	F_7	Smallest packet in flow
8	F_8	Largest packet in flow

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

9	F_9	Highest inter-arrival time between any two packets in flow
10	F_{10}	Lowest inter-arrival time between any two packets in flow
11	F_{11}	Total duration of flow
12	F_{12}	Packet frequency (flow duration/number of packets in flow)
13	F_{13}	Mean inter-time between packets sent in forward direction
14	F_{14}	Mean inter-time between packets sent in backward direction
15	F_{15}	Highest inter-time between packets sent in forward direction
16	F_{16}	Lowest inter-time between packets sent in forward direction
17	F_{17}	Highest inter-time between packets sent in backward direction
18	F_{18}	Lowest inter-time between packets sent in backward direction

It is to mention that Information gain feature selection algorithm is used here also to diminish the feature vector's dimensionality. As Random forest classifier fits good amount of decision trees and hence considered an appropriate classifier for our proposal.

3.3 Random forest approach, a Machine Learning based Classification Technique

Random forest classifier a supervised learning algorithm is used to operate and to construct a variety of decision trees during training and classifying the output. As random forest can be applied for classification as well as regression problems, it brings additional unpredictability into the model during the process of building trees as it results in best subset of features [2].

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

Table 3: Dataset for P2P Host Detection

S.No	Time window	P2P samples	Non – P2P samples
1	5 min	21387	37394
2	10 min	10695	18701
3	20 min	5062	9376

Table 4: Dataset for P2P Botnet Detection

S.No	Application	Size	Flows	Training	Evaluation
1	Storm	5.1GB	660970	528776	132194
2	Zeus	109.8MB	17634	14107	3527
3	Waledac	1.1GB	43667	34934	8733
4	Vinchuca	622MB	1423	1138	285
5	P2P benign	122GB	815659	652527	163132

The dataset is gathered by running P2P application, Vinchuca botnet in lab and considering network traffic dump. Table 3 and Table 4 are representing the datasets for P2P Host Detection and P2P Botnet Detection respectively.

4. EXPERIMENTATION AND EVALUATION

We will discuss the tools and metrics in this approach and also the results for the proposed. Wireshark 2.4.4, Tshark 2.4.4, Tcpdump 4.9.2, Pyshark 3.7.11, Scikit – learn 0.19.1 are the tools used for implementation and true positive, true negative, false positive, false negative are the metrics employed by Confusion matrix for visualization of the performance of the chosen algorithm. Additional metrics like true positive rate, false positive rate, precision, error rate, accuracy are the adornments for the proposed evaluation.

4.1 Training and Testing

The proposed dataset's training and testing are performed on Ubuntu 16.04, 16 GB

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

RAM, Intel – i7 Octa core processor LTS machine. The training set has 70 percent of occurrences, 30 percent of occurrences for testing and 10-fold cross – validation is calculated to lessen the over – fitting hazard and to safeguard the model that simplifies so.

4.1.1 P2P Host Detection Results

4.1.1.1 Feature selection results

As information gain feature selection algorithm is used for decision tree classifier in P2P host detection, the following figures represent information gain with the used features by ignoring the unaffecting features and accuracy versus reducing features respectively and the final feature vector is used for the classification.

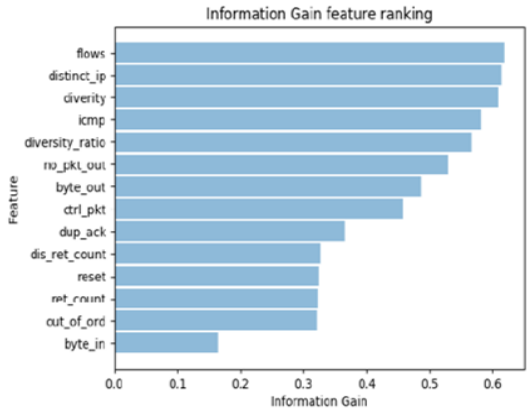

Figure 4: Information Gain

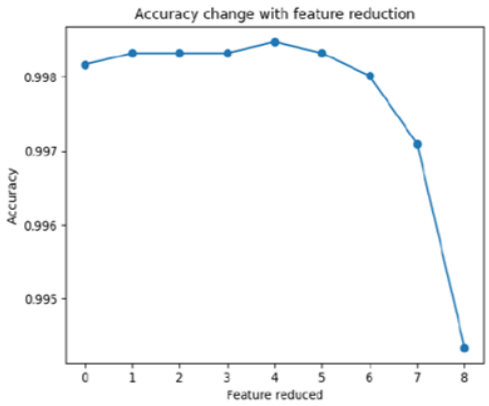

Figure 5: Variation of accuracy on reducing features

4.1.1.2 Detection Results

During P2P host detection, the extracted data is at 3 different time windows as mentioned earlier. As we are implementing Decision Tree Classifier for classification results didn't yield properly on applying the decision tree completely deprived of pruning. Hence, we will vary the depth from 2 to 20 and track the changes in error rate and we can notice that the error rate turns out to be stable after a particular depth which is represented in Figure 6 on adjusting depth for 10 minutes time window.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

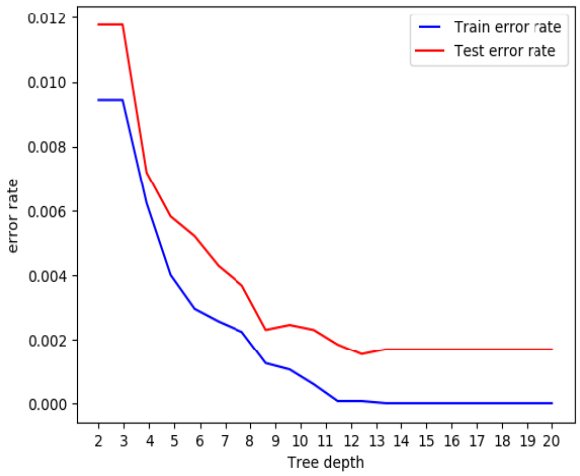

Figure 6: Variation in error with depth (10 minutes)

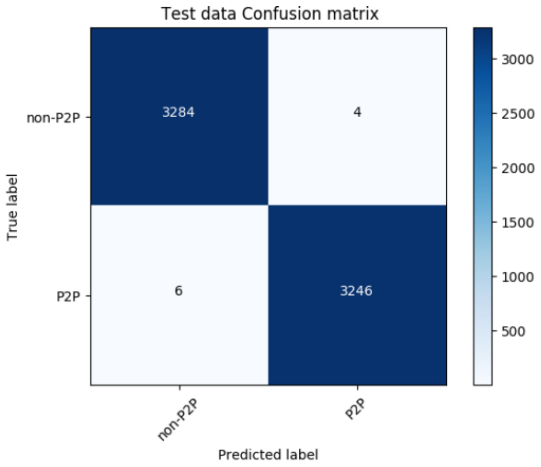

Figure 7: P2P Host Detection Confusion Matrix

Figure 7 summarizes the results in terms of TPR, FPR, Precision and accuracy. It is observed that at 20 minutes time window best results are observed with 99.81 percent TPR, 0.1 percent FPR.

Table 5: Results for P2P Host Detection Results

S.No	Time window	TPR	FPR	Precision	Accuracy
1	5 min	96.93%	2.67%	97.39%	97.13%
2	10 min	99.39%	0.55%	99.45%	99.42%
3	20 min	99.82%	0.13%	99.88%	99.85%

4.1.2 P2P Botnet Detection Results

4.1.2.1 Results using Feature selection

In P2P Botnet detection, Information Gain feature selection is used for Random Forest Classifier represented in Figure 8. The variation of accuracy after eliminating non – performing features is represented in Figure 9 which are shown below.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

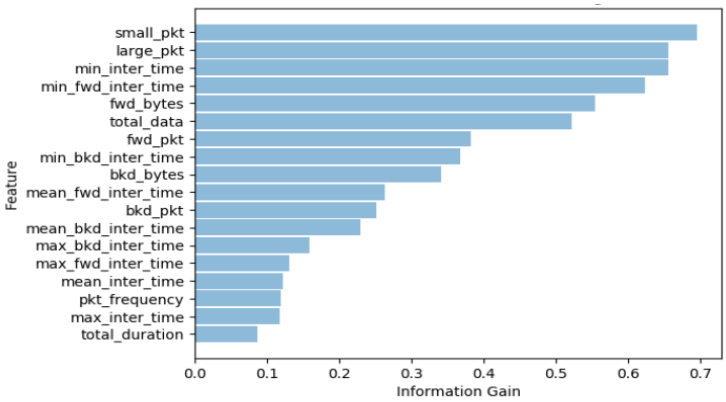

Figure 8: Information Gain feature ranking

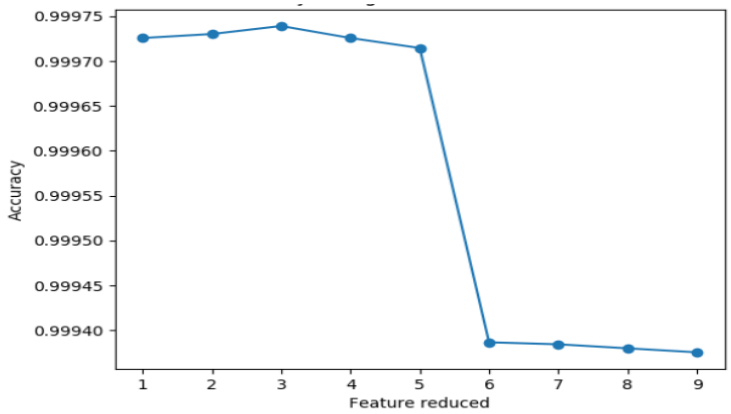

Figure 9: Variation of accuracy on reducing features

4.1.2.2 Detection Results

During P2P Botnet Detection, flow – based data is extracted for a duration of 60 minutes time window using Random Forest Classifier is mentioned earlier.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

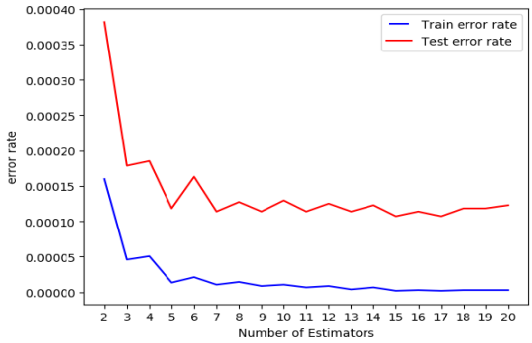

Figure 10: Variation of error rate with number of estimators

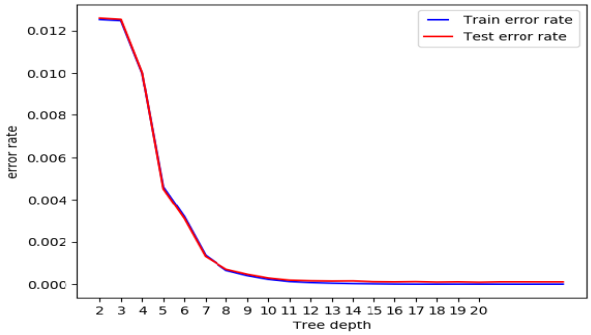

Figure 11: Variation of error rate with depth

The results are fairly good with default parameter settings of Python sklearn library and can be further improved by tuning number of estimators needed to construct Random Forest and depth of the trees with 99.9 percent TPR, 0.03 percent FPR for the data captured in 60-minute time window. Figure 10 represents Confusion matrix for test data and Table 6 represents P2P Botnet Detection results using classifier.

TPR	FPR	Precision	Accuracy
99.99%	0.04%	99.99%	99.99%

Table 6: Results for P2P Botnet Detection

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

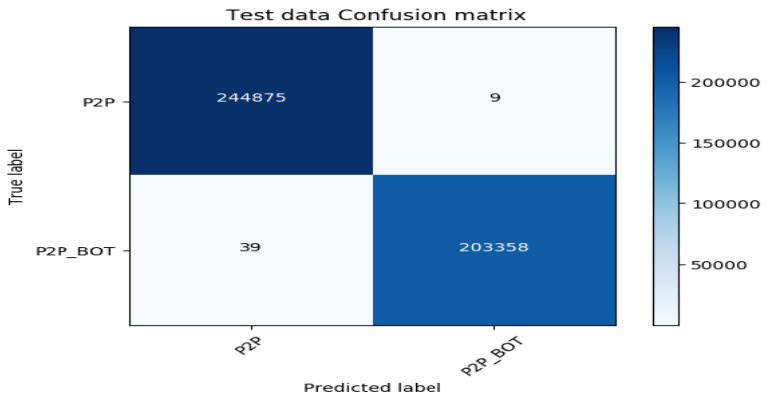

Figure 10: Confusion matrix for P2P Botnet Detection

5. CONCLUSION AND FUTURE WORK

Among botnets P2P botnets are considered as a key threat in cyber security field. Our proposal has developed an approach for this key threat. The proposal has 2 phases, P2P host detection using host – based approach is implemented in first phase and the detection of P2P bots from the known bots is done in second phase. The host – based features for a time window of 10 minutes on properties shown by P2P is extracted in first phase. Flow – based approach is implemented by flow – based features for bot detection on the patterns of host access and data exchange for window of 1 hour. It can be observed that when the compared an accuracy of 99.6% with false – positive rate is lower than 0.028%.

Based on the work done in our proposal many other enhancements can be worked to improve the work. Few of them can be implemented by (i) training the model with the live traffic (ii) selecting or using different feature selection algorithm (iii) adding or appending more features which are informative (iv)using advanced GUIs.

REFERENCES

1) The Hunt for the Financial Industry's Most-Wanted Hacker, Bloomberg Business. Retrieved 2 March 2016.
2) (Accessed April 24, 2018). https://towardsdatascience.com/the- random- forest- algorithm-d457d499ffcd/.
3) (Accessed March 10, 2018). https://cuckoosandbox.org/.
4) (Accessed May 17, 2018). http://www.tmttlt.com/archives/5289/.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/UE28D

5) H. CHOI, H. LEE, AND H. KIM, Botnet detection by monitoring group activities in DNS traffic, 7th IEEE International Conference on Computer and Information Technology, (2007).
6) M. EGELE, T. SCHOLTE, E. KIRDA, AND C. KRUEGEL, A survey on automated dynamic malware-analysis techniques and tools, ACM Computing Surveys (CSUR) Volume 44 issue 2, (Feb. 2012).
7) N. M. D. FEIS AND P. C. PATTERSON, 'botnets' and battle against cyber-crime, New York Law Journal, (Apr. 2015).
8) E. K. LUA AND M. PIAS, A survey and comparison of peer-to-peer overlay network schemes, University of Cambridge, (2005).
9) NUMMIPURO, Detecting p2p-controlled bots on the host, TKK T-110.5290, Seminar on Network Security, Aalto University, Espoo, Helsinki, (Oct. 2007).
10) VDANIEL STUTZBACH AND R. REJAIE, Understanding churn in peer-to-peer networks, 6th ACM SIGCOMM conference on Internet measurement, (2006).
11) S. T. VUONG AND M. S. ALAM, Advanced methods for botnet intrusion detection systems, University of British Columbia, Canada, (2011).
12) S. ZHANG, Conversation-based P2P Botnet Detection with Decision Fusion, PhD dissertation, The University of Brunswick, 2010.

Authors

Gadde Ramesh received the B.C.A. and M.Sc. degree from Kakatiya University, Warangal, India, in 2001 and 2004, respectively, M.Tech. Degree from Jawaharlal Nehru Technological University, Hyderabad, India in 2010. He is currently working toward the Ph.D. degree with the Faculty of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, India. He is presently an Associate Professor of Computer Science and Engineering with Vaagdevi Engineering College, affiliated to JNTU, Warangal, Telangana State, India. He is member of ISTE and CSI. He is published 13 journals and 3 conference papers at international level. He is research interest include network traffic analysis, Botnet detection, Machine learning, Peer to peer bots.

Suresh Pabboju received the Ph.D. degree in Computer Science and Engineering from Osmania University, Hyderabad, India. He is currently a Full Professor of Information Technology with Chaitanya Bharathi Institute of Technology, Affiliated to Osmania University, Hyderabad, India. He is member of IEEE, ISTE, ISME and IEEE-CIS. He is published 54 journals and 23 conference papers at international level. He is research interest include Image Processing, Data Mining, Soft Computing, Artificial Intelligence, Deep Learning, and Data Science.