I J C T A, 9(34) 2016, pp. 175-187
© International Science Press

Detecting Botnets based on Statistical Feature Analysis in a Local Enterprise Network

Harpinder Kaur[*], Sanjeev Kumar[**] and Neeraj Sharma[***]

ABSTRACT

In the internet community, the number of users are increasing exponentially by adopting the internet as their daily resources. More and more number of devices are connected to the internet which are so called Internet of Things (IoT) in current terminology, thereby the security is becoming the crucial factor in the internet community to secure the applications, data of the users hosted on the internet. Among the current malwares spreading on internet, botnet is a still a major threat in cyber security. The botnet malware is different than the other class of the malwares as they are remotely controlled by the remote attacker known as botmaster.

Here in this paper, a botnet detection framework is presented to detect the botnet through statistical features and behavior of the network flow. The statistical properties are combined with the network flow analysis to detect the botnets in a network. The developed model is validated on honeypot traces and it is applicable to detect the botnets in a single network. The statistical engine is deployed in a network as SPAN port of the switch and it is giving quite promising results in detection of botnets. As botnet family has specific behavior and there exist adequate similarities in each botnet that separates the behavior of it from the benign traffic which is key feature to distinguish the botnet traffic from normal traffic of a network. Here in this paper a subset of features are selected based on wide experimentations and behavior analysis of botnet families. In the end the evaluation of the developed model is presented which depict that it is possible to detect botnet through statistical feature analyses in a single enterprise network.

Keywords: Botnets, Cyber Security, IoT, Traffic Analysis, Honey Pot, Malware

I. INTRODUCTION

In today's computer network, the size of connected devices in the form of IoTs, ADSL(s), and PLC is growing in an exponential rate and so are threats to them. The current IT security or security of Internet connected devices is becoming the prime concern to protect the internet from cyber-attacks. In today's scenario, the end users or devices connected to internet are more prone to the cyber-attacks as 1) end user is less aware about the security techniques 2) connected devices has less in-built security 3) High bandwidth availabilities. The integration of latest IoT devices to internet keep the user more connected to the internet and thereby more prone to the attacks. The security of these internet connected devices is becoming the critical for researchers as due to less secure implementation, these Internet-connected embedded devices, including Smart TVs, Refrigerators, Microwaves, Set-top boxes, Security Cameras and printers, are routinely being hacked and used as weapons in cyber-attacks [30-46].

The bot malwares are class of threats spreading on internet which are controlled, monitored and updated by the remote attackers known as botmaster. In botnet infections, there are two known things so called bot and botmaster. The bot is a malware is planted on a vulnerable system by botmaster [4]. Bot – it is a new

[*] M.E. Scholar (CSE), Chandigarh University, Gharuan, Punjab, India, *E-mail: Harpinder.kaur1011@gmail.com*

[**] Cyber Security Technology Division, CDAC, Mohali, India, *E-mail: sanjeev@cdac.in*

[***] Department of CSE, Chandigarh University, Gharuan, Punjab, India, *E-mail: neeraj.kirti@gmail.com*

type of malware which is installed into the compromised systems and then controlled by the botmaster for executing some commands which are ordered by the botmaster. [7].

The rest of the paper is organized as follow, Section II. presents the overview of related work. Section III. describes statistical feature analyses. Section IV describes the methodology of implementation. Section V presented the experimental results and finally in section VI concluded the paper and outline the future work.

II. RELATED WORK

Saad et al. [19], the features are divided into groups- flow based and host-based, 17 such features are applied to determine the botnets. Beigi et al. [1] presented the network traffic features of bot malwares and the subset of most appropriate features are selected specifically for bots.

Basam Sayad et.al. [2], discuss the botnet detection based on network traffic behavior analysis and flow analysis. In this paper, they propose a new approach to detect the botnet activities based on traffic behavior and by applying machine learning model.

The authors in [4-11], [13-17] addressed the network based botnet detection method, spam botnet detection, some of them are independent on protocol used in botnet communications.

Nguyen et al. [18] presented different techniques to recognizing statistical pattern by externally observed traffic attributes without any deep inspection of the packet payload that can be encrypted. The classification of network traffic are performed based on different set of attributes such as:

1) Port based attributes are targeted TCP or UDP port number.
2) Statistical based features relate to refer the traffic flow attributes like duration, packet length, and average time and so on. These attributes defines uniqueness against different application from another ones.

Author in [23] addressed the botnet detection problem through correlation of IDS events, which give the promising results for centralized botnet detection. The network dialogs of IDS events are statistically correlated to determine the botnet infections in a network.

G. Gu et. al [24] presented the BotMiner which is protocol independent framework for botnet detection. In this they applied the clustering techniques to group the network traffic into different set of botnets.

Sanjeev Kumar et.al [25-27], [12] discuss the honeynet based malware capturing and then analysis of captured traffic analysis for botnet signatures. The major restrictions of these techniques is that it may not be able to detect the encrypted behavior of network traffic and for those malwares which uses the evasion mechanism as their stealth behavior.

During the research study performed, it was observed that no accurate set of statistical features of network traffic of botnet families were available for public research and which are less resources intensive for detection of botnet malware in a single enterprise network. The hypothesis in this that a high resources intensive attribute cannot be taken as botnet classifier in an enterprise network. Therefore the set of statistical properties which are relevant for a single network are selected by performing analysis on captured bot malwares on honeynet.

III. BOTNET DETECTION FRAMEWORK

In this section, we discuss the process of botnet related statistical feature extraction and building database for selected stats.

3.1. Creation of Base line environment for botnet detection

To build the baseline environment for botnet detection framework, firstly network traces corresponding Bot malwares captured on honeynet are processed by applying deep packet inspections and payload analysis. The statistical behavior properties are identified which determine the botnets. The process of analysis is 1) First, the complete TCP-handshake connections is identified in a network trace 2) then following traffic were inspected to see the malware drop as occurred in a botnet infection life cycle 3) In the continuation of it, the traffic stats were observed corresponding to the IP address identified in a malware drop as shown in the following snapshot figure 2. 4) The traffic behavior were observed of particular IP address to detect the anomalous behavior of it as shown in snapshot figure 3.

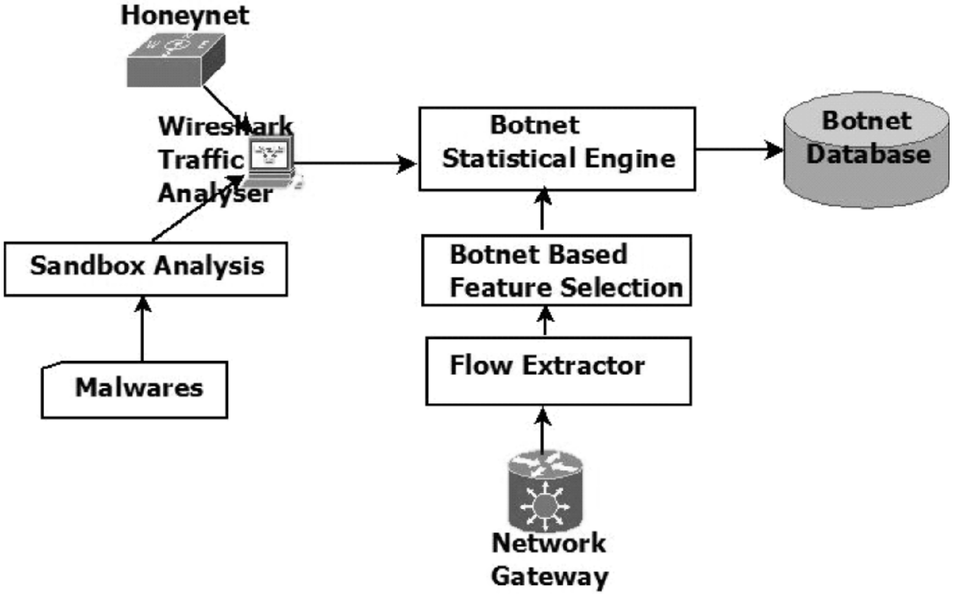

Honeynet
Wireshark Traffic Analyser
Sandbox Analysis
Malwares
Botnet Statistical Engine
Botnet Database
Botnet Based Feature Selection
Flow Extractor
Network Gateway

Figure 1: Botnet Base lining Framework

Figure 1 describe the botnet baseline framework, it has four major modules: bot traffic extractor, statistical engine (Program), database, Flow based feature extractor. Malware execution module.

Firstly, the statistical model of botnet traffic is developed by extracting the behavioral features of known botnet traffic captured on a honeynet repository. The large set of network dumps corresponding to the botnet families are processed to build the statistical engine model. Then the unknown class of malware samples which has some stealth behavior are analyzed through commercial sandbox solution known as Threat analyzer. The statistical properties are combined of the both class of malwares in the form a database.

3.1.1. Bot Traffic Extractor

The network traces of honeynet are processed by applying a deep packet inspection to extract the bot relevant traffic as per the matching criteria of below mentioned feature [22] set in table 1, table 2. The set of 919 network dumps of size more than 100Kb size are processed to build the statistical analysis engine. The malware samples which do not have network traces are also processed through sandbox and extracted the complete behavior of the malware. The traffic dump of indicated Bot malwares are processed to give the data feed to statistical analysis engine.

3.1.2. Flow Extractor

Flow extractor is a program which extract the network flow of related bot malwares and submit this data set as feed to model which apply the analysis corresponding to the extracted botnet features. The extracted outcome in the form of bot malwares behavior properties are appended into database. To validate the hypothesis in this module to check the botnet behavior in the traffic, the payload based inspection technique is applied. The payload inspection technique is resource intensive and very time consuming, thereby this is only applied to validate the extracted traffic behavior of the bot malware.

3.1.3. Botnet Feature Selection

In the case of botnet based features selection, we find those features which help us to show the flow is malicious or not. It means feature based selection of network flow can be extracted by analyzing benign and non- malicious traffic flow. By analyzing all types of flow behavior, we have extracted some features that show in figure 1 flow is malicious.

3.1.4. Database

In order to evaluate the performance of the proposed detection method, database is built which include the statistical properties of bot malwares as well normal network traces captured on an enterprise LAN network [30] in cyber security laboratory of CDAC. Some of the traces were recorded in sandbox by re-executing the malwares, others in real networks.

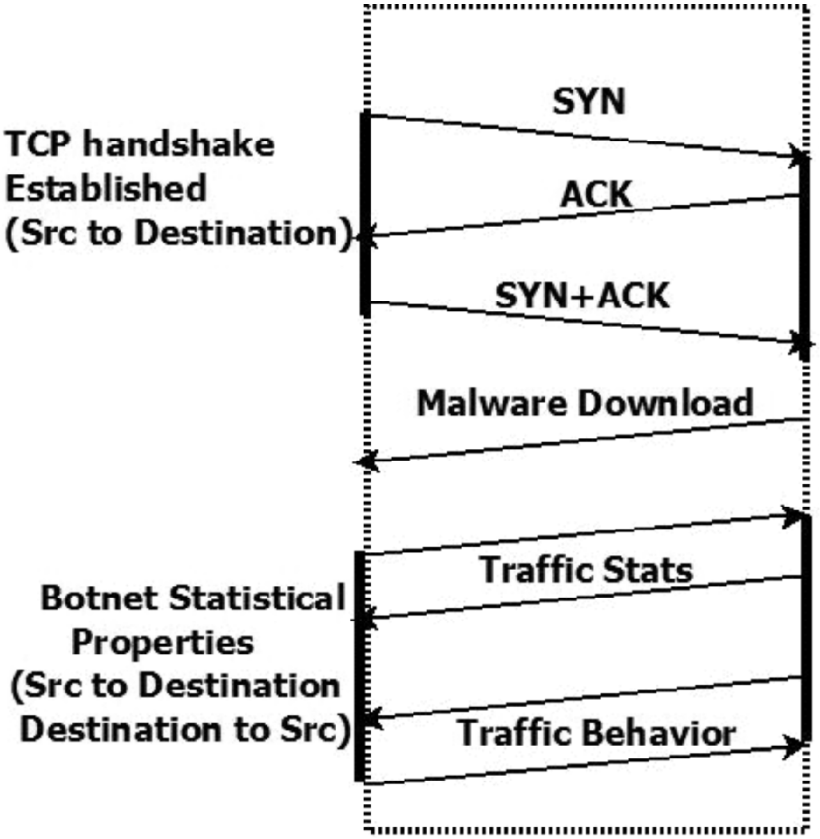
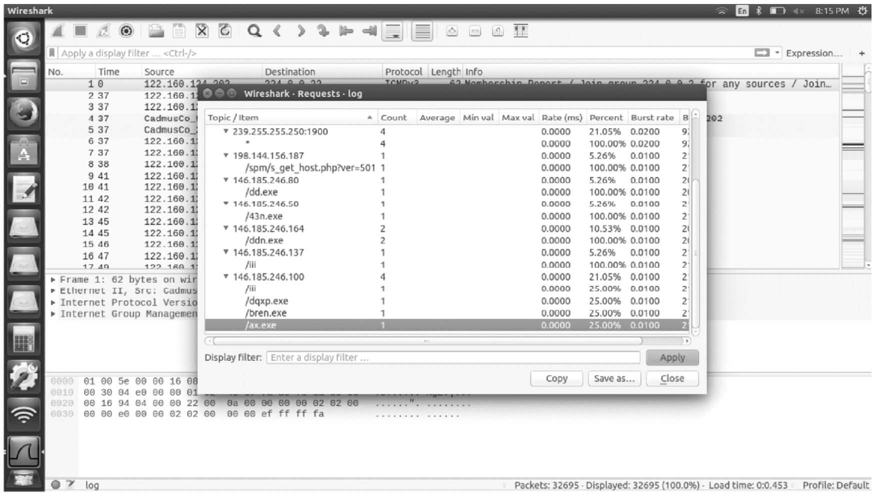

SYN
TCP handshake Established (Src to Destination)
ACK
SYN+ACK
Malware Download
Traffic Stats
Botnet Statistical Properties (Src to Destination Destination to Src)
Traffic Behavior

Figure 2: Flow of Extracting Botnet features to create a baseline data sets

By applying the multiple iterations of above processes, the database has created to record the statistical results of the botnet families. The detailed behavior analyses in the form of stats of individual bot malwares are stored in a database to feed it as model engine. The following snapshots depict the database view of the botnet database. Figure 5 depict the database schema design which is further used for as a direct feed to intelligent algorithms for detection of botnets.



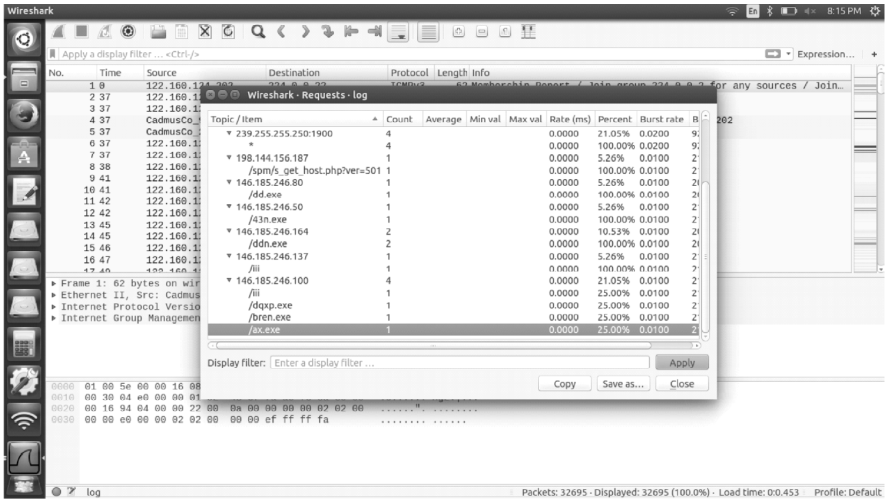
Wireshark · Requests · log

Topic / Item	Count	Average	Min val	Max val	Rate (ms)	Percent	Burst rate	B
▾ 239.255.255.250:1900	4				0.0000	21.05%	0.0200	9
*	4				0.0000	100.00%	0.0200	9
▾ 198.144.156.187	1				0.0000	5.26%	0.0100	2
/spm/s_get_host.php?ver=501	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.80	1				0.0000	5.26%	0.0100	2
/dd.exe	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.50	1				0.0000	5.26%	0.0100	2
/43n.exe	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.164	2				0.0000	10.53%	0.0100	2
/ddn.exe	2				0.0000	100.00%	0.0100	2
▾ 146.185.246.137	1				0.0000	5.26%	0.0100	2
/iii	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.100	4				0.0000	21.05%	0.0100	2
/iii	1				0.0000	25.00%	0.0100	2
/dqxp.exe	1				0.0000	25.00%	0.0100	2
/bren.exe	1				0.0000	25.00%	0.0100	2
/ax.exe	1				0.0000	25.00%	0.0100	2

Display filter: Enter a display filter ...

Packets: 32695 · Displayed: 32695 (100.0%) · Load time: 0:0.453 Profile: Default

Figure 3: Traffic stats of IP observed in malware drop



Wireshark · Requests · log

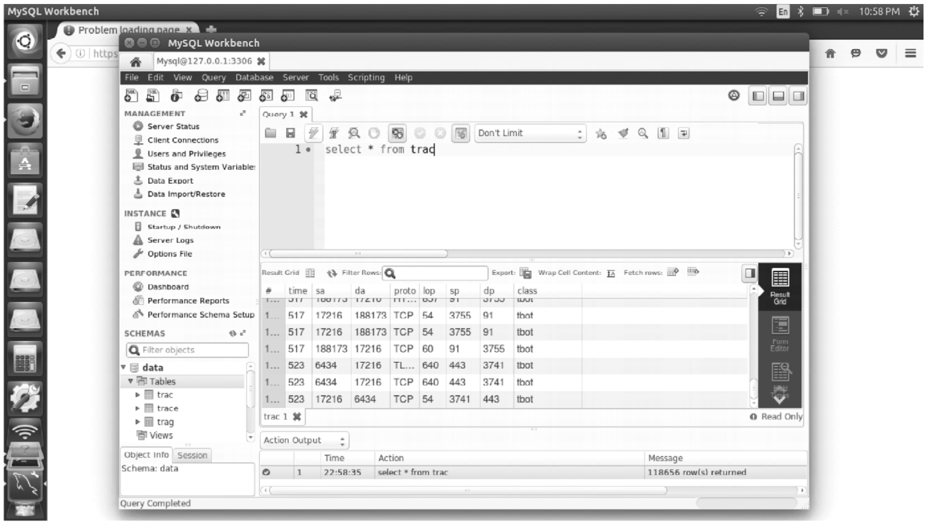
Topic / Item	Count	Average	Min val	Max val	Rate (ms)	Percent	Burst rate	B
▾ 239.255.255.250:1900	4				0.0000	21.05%	0.0200	9
*	4				0.0000	100.00%	0.0200	9
▾ 198.144.156.187	1				0.0000	5.26%	0.0100	2
/spm/s_get_host.php?ver=501	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.80	1				0.0000	5.26%	0.0100	2
/dd.exe	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.50	1				0.0000	5.26%	0.0100	2
/43n.exe	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.164	2				0.0000	10.53%	0.0100	2
/ddn.exe	2				0.0000	100.00%	0.0100	2
▾ 146.185.246.137	1				0.0000	5.26%	0.0100	2
/iii	1				0.0000	100.00%	0.0100	2
▾ 146.185.246.100	4				0.0000	21.05%	0.0100	2
/iii	1				0.0000	25.00%	0.0100	2
/dqxp.exe	1				0.0000	25.00%	0.0100	2
/bren.exe	1				0.0000	25.00%	0.0100	2
/ax.exe	1				0.0000	25.00%	0.0100	2

Display filter: Enter a display filter ...

Packets: 32695 · Displayed: 32695 (100.0%) · Load time: 0:0.453 Profile: Default

Figure 4: Traffic volume observed in malware drop

By applying traffic analysis on network dumps captured on a honeynet as vulnerable set of machines, the database is built to



```
select * from trac
```

#	time	sa	da	proto	lop	sp	dp	class
1...	517	17216	188173	TCP	54	3755	91	tbot
1...	517	17216	188173	TCP	54	3755	91	tbot
1...	517	188173	17216	TCP	60	91	3755	tbot
1...	523	6434	17216	TL...	640	443	3741	tbot
1...	523	6434	17216	TCP	640	443	3741	tbot
1...	523	17216	6434	TCP	54	3741	443	tbot

		Time	Action	Message
●	1	22:58:35	select * from trac	118656 row(s) returned

Schema: data

Query Completed

Figure 5: Database with bot stats

3.2. Statistical Botnet features

In literatures, there are number of features which often repetitive in various studies. But also there are number of features which are unique and can be retained as unique feature of botnet. Moreover the new set of botnet features are observed and incorporated in statistical model which are identified by executing the real bot samples in controlled environment.

The certain behavior of bot malware is: a) a bot may also inject random number of packets into the C&C communication to throw off different size packets. This technique, allows to take advantage of the fact that bots will generate more connections and more flows rather than non-malicious traffic b) the bots through C&C communication also measure the reconnects and connections through the communication. In the case of bots, they randomly reconnects again and again for tasks and carries out these tasks, through this random disconnect a bot will perform some mitigated against the bot activities.

3.3. Feature Selection

For the initial building of the model with available data sets of honeynet, the botnet features are obtained from literatures. Elaheh Biglar Beigi et.al [1], David Zhao et.al [13] and Matija Stevanovic [12] describe the behavior features [Table 1] of the bot malware which lead to machine learning based botnet detection. There are some features which are mostly extracted for every TCP/UDP header flow like source IP, destination IP, source port, destination port, and protocol. These features are the part of feature vector which capture the characteristics of single flow.

Over the above mentioned feature set, the more number of statistical features are added into the model which are extracted from deep investigation of bot malwares captured on honeynet. The complete list of these features are depicted in table 2. These features are selected by through research experimentations and analysis of bot traffic captured through implementation of honeypot sensors. The traffic behavior of such bot samples were analyzed to finalize these features of the botnet traffic. Presently the auto-learning capabilities are added into the engine by writing a small python program which automatically extract the flow level features of botnet traffic and append the database. It was also observed that during the research analysis, few class of malwares were not having network traces on honeypot sensors, thereby those were re-executed in sandbox environment to study the behavior of them and extract the relevant botnet features. The network dumps created by sandbox are further processed through the python program to automatically append the database.

Table 1
Features List of Botnet traffic

Attribute Name	*Description*
Src IP	Source IP address
Dest IP	Destination IP
Src port	source port
Dest port	destination port
Name of protocol	Transport layer protocol(TCP)
Payload average length	Avg. pkt payload length on the network for a particular time interval
Variance of payload packet length	Variance of payload pkt length on the network for that particular time interval
Total no of packets exchanged	No. of pkt's exchanged for particular time interval
Total no of Pkts exchanges/sec	No. of pkt's exchanged /sec for time window T
First packet's size	Size of 1st pkt's in the traffic flow.

contd. table 1

Attribute Name	*Description*
Average time interval between pkts	Avg. time b/w pkt's for the time window T
Reconnects attempts	No. of reconnects for a traffic flow
Flow count for a single address	No. of network traffic from a address over the total no. of flow generated / hour
Count of NULL pkts exchanged	Count of null pkts exchanged in the network
Count of small pkts exchanged in network	Count of small pkts exchanged in the network
%age of small packet exchanged on the network	%age of small pkts exchanged
Ratio between the number of incoming over outgoing packets exchanged	Ratio b/w the no. of incoming pkt's over outgoing pkt's exchanged
Time Duration	Count of time duration
Count of bytes exchanged	Count of bytes exchanged in the network
Number of packets exchanged with the same length	Total no. of packets exchanged of same size
S.D of packet payload length	S.D of pkt's payload length
Average number of bits / sec	Avg. no. of bits/sec in the flow
Average number of packets / sec	Avg. number of pkts/sec
Average inter-arrival time of packets	Avg. inter- arrival time of pkt's
Mean of bytes/ sec	Mean of total bytes/ sec
Standard of number of bytes/ sec	S.D of the total number of bytes/sec
TCP 3 –way handshake	TCP 3-way handshake communication
Number of tear down connections	No. of tear down connections in the flow
average of duration period	Avg. of duration period of the flow
%age of TCP SYN packets	%age of TCP SYN packets
%age of TCP SYN ACK packets	%age of TCP SYN ACK pkts
%age of TCP ACK packets	%age of TCP ACK pkts
% age of ACK PUSH packets	%age of ACK PUSH pkt's
Src IP	Source IP address
Dest IP	Destination IP
Src port	source port
Dest port	destination port

Table 2
Added features of botnet traffic

Features	*Type*
Total number of packets	Numerical
Layer 7 protocol	Numerical
Duration (seconds)	Numerical
Average packet payload size	Numerical
Total number of bytes	Numerical
Average number of bytes/sec	Numerical
Average number of bit/sec	Numerical
Mean of number of bytes/packets	Numerical
Percentage of TCP connections	Numerical
Number of null packet exchange	Numerical

contd. table 2

Features	*Type*
Number of small packet exchanged	Numerical
Number of large packet exchanged	Numerical
Percentage of HTTP response packets	Numerical
Percentage of HTTP request packets	Numerical
Percentage of GET request	Numerical
Percentage of Post request	Numerical
Total number of ports	Numerical
Total number of DNS packets	Numerical
Number of conversation	Numerical
Total number of resolved IPs	Numerical
Number of average packet exchanged	Numerical
Percentage of small packet exchanged	Numerical
Percentage of large packet exchanged	Numerical

The above mentioned features in table 2, are selected through in depth experimentations and analysis. For example, the layer 7 based features depict the communication behavior of botnets in application layer and this is observed in most of the financial botnets to steal the sensitive information from the user. The user browse the websites which exploit the client side applications and drop a malware into the user’s machine. This feature is observed in financial botnets.

IV. EXPERIMENTAL RESULTS

Initially the set of 5 botnet families known as Tbot, Kelihos cutwail, Blackhole are processed to build the statistical engine for botnet detection. Following tables indicate the network stats of the botnet families.

Table 3
Botnet families analysed

botnet traces	*number of packets*	*number of TCP connections*	*number of reconnections*	*durations (in sec)*
Tbot	6908	5290	256	1729
Tbot	5004	3966	181	523
Tbot	8000	6224	352	533
Tbot	4787	3907	208	312
Tbot	13050	10503	1556	5403
kelihos	133	37	6	69
kelihos	13961	13924	697	1091
cutwail	8612	8041	70	75
blackhole	3128	3057	0	105
cutwail	133	13	5	68

Table 4
Traffic stats of botnet families

botnet traces	*number of packets*	*TCP connections*	*number of reconnections*	*durations*
1	26166	5316	54	63882
2	52332	10636	208	86400
3	60646	12564	229	83240
4	55545	11545	199	81452
5	45489	95245	165	80125

Table 4 indicate the number of packets in network dumps corresponding to the analyzed botnet malwares. The number of TCP connections recorded in communications and number of reconnection attempts are extracted from the network traces.

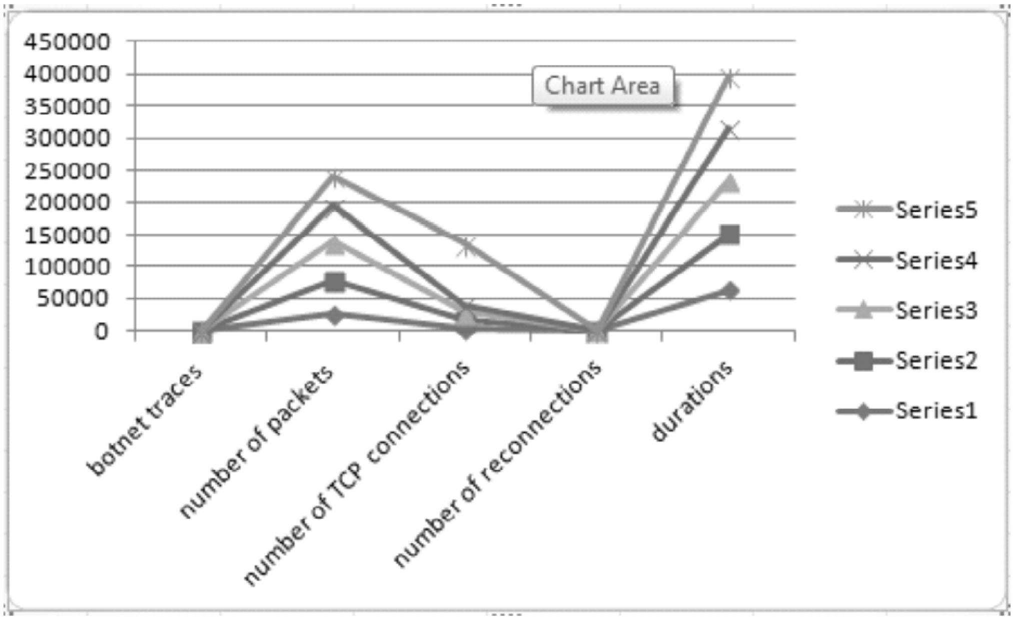


Figure 6: Traces of Botnet traffic w.r.t selected features

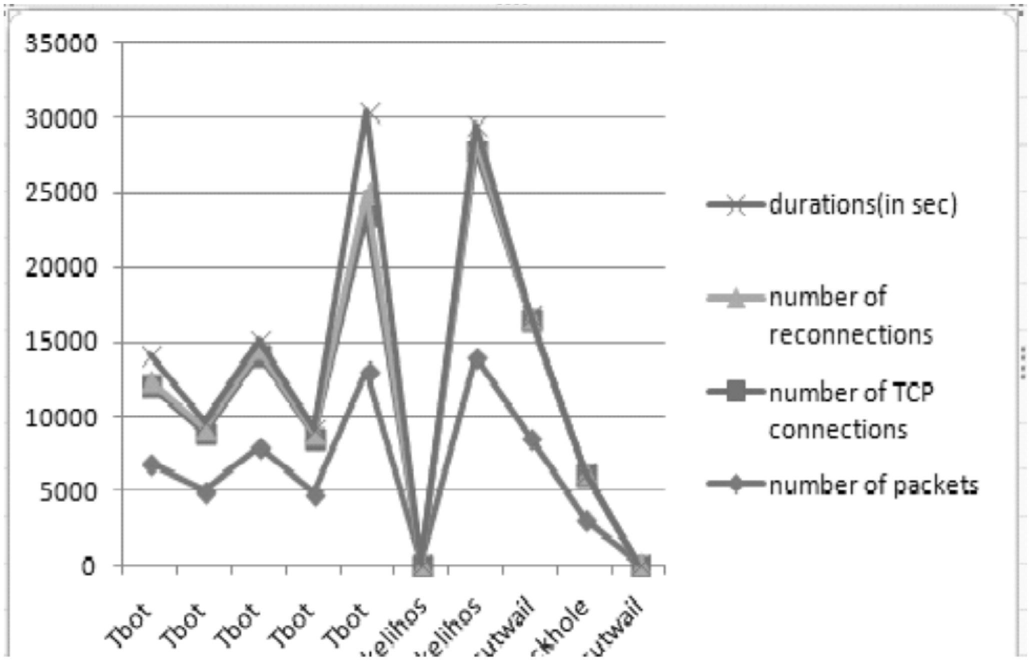
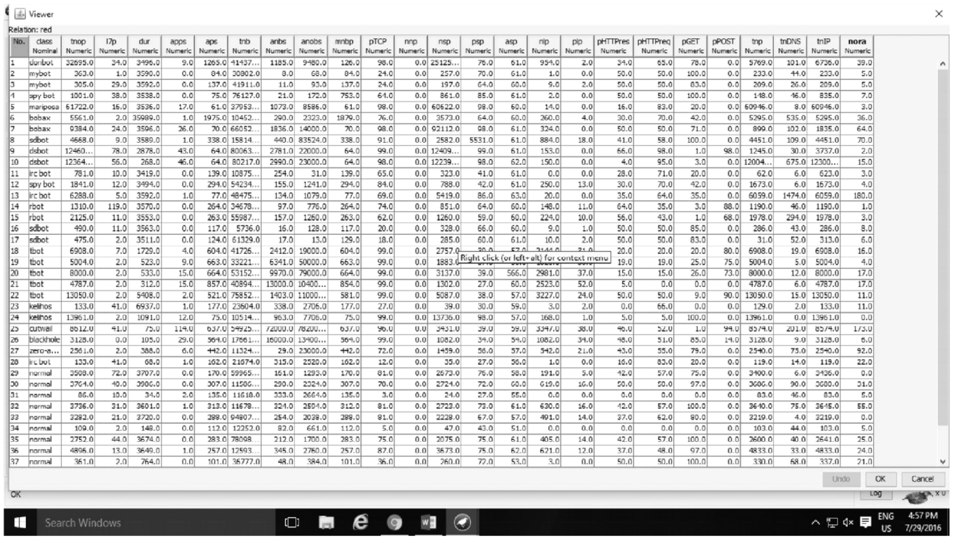


Figure 7: Botnet families w.r.t botnet features

The statistical analytical techniques are applied on extracted data set by R tool. Below figure depict the attributes of the data set formulated through above mentioned methodology.

The data set is tested for experimentations and loaded into R tool as CSV file and the certain statistical algorithms were applied to get the accuracy of the data set. The following figure depict the principle component analysis results produced by the R engine [47].



No.	class	tnop	l7p	dur	apps	aps	tnb	anbs	anobs	mnbp	pTCP	nnp	nsp	psp	asp	nip	pip	pHTTPres	pHTTPreq	pGET	pPOST	tnp	tnDNS	tnIP	nora
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	donbot	32695.0	34.0	3496.0	9.0	1265.0	41437...	1185.0	9480.0	126.0	98.0	0.0	25125...	76.0	61.0	954.0	2.0	34.0	65.0	78.0	0.0	5769.0	101.0	6736.0	39.0
2	mybot	363.0	1.0	3590.0	0.0	84.0	30802.0	8.0	68.0	84.0	24.0	0.0	257.0	70.0	61.0	1.0	0.0	50.0	50.0	100.0	0.0	233.0	44.0	233.0	5.0
3	mybot	305.0	29.0	3592.0	0.0	137.0	41911.0	11.0	93.0	137.0	24.0	0.0	197.0	64.0	60.0	9.0	2.0	50.0	50.0	83.0	0.0	209.0	26.0	209.0	5.0
4	spy bot	1001.0	38.0	3538.0	0.0	75.0	76127.0	21.0	172.0	753.0	64.0	0.0	861.0	85.0	61.0	2.0	0.0	50.0	50.0	100.0	0.0	148.0	46.0	835.0	7.0
5	mariposa	61722.0	16.0	3536.0	17.0	61.0	37953...	1073.0	8586.0	61.0	98.0	0.0	60622.0	98.0	60.0	14.0	0.0	16.0	83.0	20.0	0.0	60946.0	8.0	60946.0	3.0
6	bobax	5561.0	2.0	35989.0	1.0	1975.0	10452...	290.0	2323.0	1879.0	76.0	0.0	3573.0	64.0	60.0	260.0	4.0	30.0	70.0	42.0	0.0	5295.0	535.0	5295.0	36.0
7	bobax	9384.0	24.0	3596.0	26.0	70.0	66052...	1836.0	14000.0	70.0	98.0	0.0	92112.0	98.0	61.0	324.0	0.0	50.0	50.0	71.0	0.0	899.0	102.0	1835.0	64.0
8	sdbot	4668.0	9.0	3589.0	1.0	338.0	15814...	440.0	83524.0	338.0	91.0	0.0	2582.0	5531.0	61.0	884.0	18.0	41.0	58.0	100.0	0.0	4451.0	109.0	4451.0	70.0
9	dsbot	12460...	78.0	2878.0	43.0	64.0	80063...	2781.0	22000.0	64.0	99.0	0.0	12409...	99.0	61.0	153.0	0.0	66.0	98.0	1.0	98.0	1245.0	30.0	3737.0	2.0
10	dsbot	12364...	56.0	268.0	46.0	64.0	80217.0	2990.0	23000.0	64.0	98.0	0.0	12239...	98.0	62.0	150.0	0.0	4.0	95.0	3.0	0.0	12004...	675.0	12300...	15.0
11	irc bot	781.0	10.0	3419.0	0.0	139.0	10875...	254.0	31.0	139.0	65.0	0.0	323.0	41.0	61.0	0.0	0.0	28.0	71.0	20.0	0.0	62.0	6.0	623.0	3.0
12	spy bot	1841.0	12.0	3494.0	0.0	294.0	54234...	155.0	1241.0	294.0	84.0	0.0	788.0	42.0	61.0	250.0	13.0	30.0	70.0	42.0	0.0	1673.0	6.0	1673.0	4.0
13	irc bot	6288.0	5.0	3592.0	1.0	77.0	48475...	134.0	1079.0	77.0	69.0	0.0	5419.0	86.0	63.0	20.0	0.0	35.0	64.0	35.0	0.0	6059.0	1474.0	6059.0	180.0
14	rbot	1310.0	119.0	3570.0	0.0	264.0	34678...	97.0	776.0	264.0	74.0	0.0	851.0	64.0	60.0	148.0	11.0	64.0	35.0	3.0	88.0	1190.0	46.0	1190.0	1.0
15	rbot	2125.0	11.0	3553.0	0.0	263.0	55987...	157.0	1260.0	263.0	62.0	0.0	1260.0	59.0	60.0	224.0	10.0	56.0	43.0	1.0	68.0	1978.0	294.0	1978.0	3.0
16	sdbot	490.0	11.0	3563.0	0.0	117.0	5736.0	16.0	128.0	117.0	20.0	0.0	328.0	66.0	60.0	9.0	1.0	50.0	50.0	85.0	0.0	286.0	43.0	286.0	8.0
17	sdbot	475.0	2.0	3511.0	0.0	124.0	61329.0	17.0	13.0	129.0	18.0	0.0	285.0	60.0	61.0	10.0	2.0	50.0	50.0	83.0	0.0	31.0	52.0	313.0	6.0
18	tbot	6908.0	7.0	1729.0	4.0	604.0	41726...	2412.0	19000.0	604.0	99.0	0.0	2757.0	30.0	57.0	2144.0	31.0	20.0	20.0	20.0	80.0	6908.0	19.0	6908.0	16.0
19	tbot	5004.0	2.0	523.0	9.0	663.0	33221...	6341.0	50000.0	663.0	99.0	0.0	1883.0	[illegible]	[illegible]	[illegible]	[illegible]	19.0	19.0	25.0	75.0	5004.0	5.0	5004.0	4.0
20	tbot	8000.0	2.0	533.0	15.0	664.0	53152...	9970.0	79000.0	664.0	99.0	0.0	3137.0	39.0	566.0	2981.0	37.0	15.0	15.0	26.0	73.0	8000.0	12.0	8000.0	17.0
21	tbot	4787.0	2.0	312.0	15.0	857.0	40894...	13000.0	10400...	854.0	99.0	0.0	1302.0	27.0	60.0	2523.0	52.0	5.0	0.0	0.0	0.0	4787.0	6.0	4787.0	17.0
22	tbot	13050.0	2.0	5408.0	2.0	521.0	75852...	1403.0	11000...	581.0	99.0	0.0	5087.0	38.0	57.0	3227.0	24.0	50.0	50.0	9.0	90.0	13050.0	15.0	13050.0	11.0
23	kelihos	133.0	41.0	6937.0	1.0	177.0	23604.0	338.0	2706.0	177.0	27.0	0.0	39.0	30.0	59.0	3.0	2.0	0.0	66.0	0.0	0.0	129.0	2.0	133.0	11.0
24	kelihos	13961.0	2.0	1091.0	12.0	75.0	10514...	963.0	7706.0	75.0	99.0	0.0	13736.0	98.0	57.0	168.0	1.0	5.0	5.0	100.0	0.0	13961.0	0.0	13961.0	0.0
25	cutwail	8612.0	41.0	75.0	114.0	637.0	54925...	72000.0	78200...	637.0	96.0	0.0	3431.0	39.0	59.0	3347.0	38.0	46.0	52.0	1.0	94.0	8574.0	201.0	8574.0	173.0
26	blackhole	3128.0	0.0	105.0	29.0	564.0	17661...	16000.0	13400...	564.0	99.0	0.0	1082.0	34.0	54.0	1082.0	34.0	48.0	51.0	85.0	14.0	3128.0	9.0	3128.0	6.0
27	zero-a...	2561.0	2.0	388.0	6.0	442.0	11324...	29.0	23000.0	442.0	72.0	0.0	1459.0	56.0	57.0	542.0	21.0	43.0	55.0	79.0	0.0	2540.0	75.0	2540.0	92.0
28	irc bot	133.0	41.0	68.0	1.0	162.0	21674.0	315.0	2520.0	162.0	12.0	0.0	35.0	27.0	56.0	1.0	0.0	16.0	83.0	20.0	0.0	119.0	14.0	119.0	22.0
29	normal	3508.0	72.0	3707.0	0.0	170.0	59965...	161.0	1293.0	170.0	81.0	0.0	2673.0	76.0	58.0	191.0	5.0	42.0	57.0	75.0	0.0	3400.0	6.0	3436.0	0.0
30	normal	3764.0	40.0	3986.0	0.0	307.0	11586...	290.0	2324.0	307.0	70.0	0.0	2724.0	72.0	60.0	619.0	16.0	50.0	50.0	97.0	0.0	3606.0	90.0	3680.0	31.0
31	normal	86.0	10.0	34.0	2.0	135.0	11618.0	333.0	2664.0	135.0	3.0	0.0	24.0	27.0	55.0	0.0	0.0	0.0	0.0	0.0	0.0	83.0	46.0	83.0	5.0
32	normal	3736.0	31.0	3601.0	1.0	313.0	11678...	324.0	2594.0	312.0	81.0	0.0	2723.0	73.0	61.0	630.0	16.0	42.0	57.0	100.0	0.0	3640.0	75.0	3645.0	55.0
33	normal	3282.0	21.0	3720.0	0.0	288.0	94807...	254.0	2038.0	288.0	81.0	0.0	2228.0	67.0	57.0	491.0	14.0	37.0	62.0	80.0	0.0	3219.0	4.0	3219.0	0.0
34	normal	109.0	2.0	148.0	0.0	112.0	12252.0	82.0	661.0	112.0	5.0	0.0	47.0	43.0	51.0	0.0	0.0	0.0	0.0	0.0	0.0	103.0	44.0	103.0	5.0
35	normal	2752.0	44.0	3674.0	0.0	283.0	78098...	212.0	1700.0	283.0	75.0	0.0	2075.0	75.0	61.0	405.0	14.0	42.0	57.0	100.0	0.0	2600.0	40.0	2641.0	25.0
36	normal	4896.0	13.0	3649.0	1.0	257.0	12593...	345.0	2760.0	257.0	87.0	0.0	3673.0	75.0	62.0	621.0	12.0	37.0	48.0	97.0	0.0	4833.0	33.0	4833.0	24.0
37	normal	361.0	2.0	764.0	0.0	101.0	36777.0	48.0	384.0	101.0	36.0	0.0	260.0	72.0	53.0	3.0	0.0	50.0	50.0	100.0	0.0	330.0	68.0	337.0	21.0

Right click (or left+alt) for context menu

Figure 8: Formulated Data set

Figure 10 depicts the classifier results when the data is loaded with Naïve Bays algorithm. The total TP rate observed is 88.8 % which is initially good enough but it also require to incorporate data of other bot malwares.

V. CONCLUTION AND FUTURE WORK

In this paper, the botnet detection based on statistical feature analysis in presented by inclusion of exiting features of botnet families and adding more botnet network traffic features by performing in depth experimentations and analysis applied on honeynet data. The statistical engine is developed which is placed in a network to sniff the traffic and extract the botnet traffic features from a network dumps in near time. The developed system is producing quite promising results for set of botnet families.

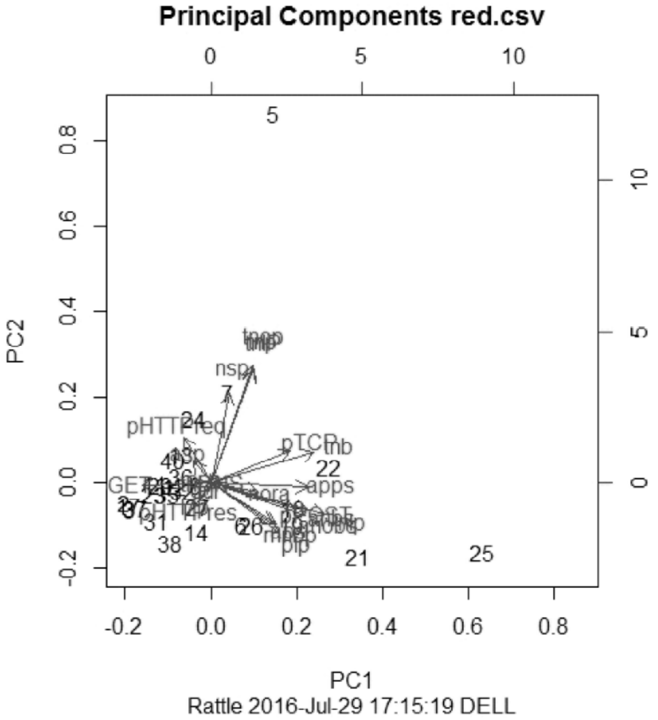
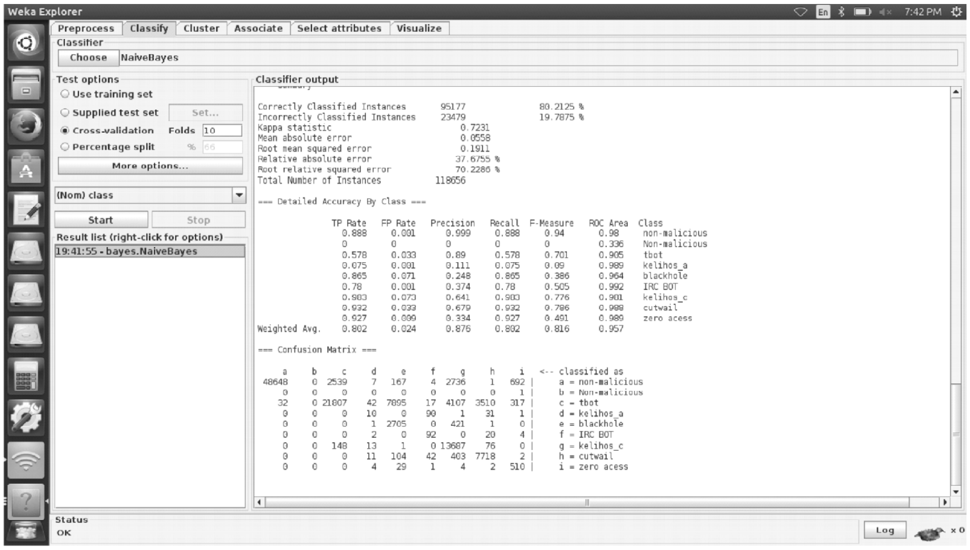


Figure 9: Principle component analysis



Classifier: NaiveBayes

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options): 19:41:55 - bayes.NaiveBayes

Status: OK

Classifier output

```
Correctly Classified Instances       95177               80.2125 %
Incorrectly Classified Instances     23479               19.7875 %
Kappa statistic                          0.7231
Mean absolute error                      0.0558
Root mean squared error                  0.1911
Relative absolute error                 37.6755 %
Root relative squared error             70.2286 %
Total Number of Instances           118656

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.888     0.001      0.999     0.888     0.94       0.98     non-malicious
                 0         0          0         0         0          0.336    Non-malicious
                 0.578     0.033      0.89      0.578     0.701      0.905    tbot
                 0.075     0.001      0.111     0.075     0.09       0.989    kelihos_a
                 0.865     0.071      0.248     0.865     0.386      0.964    blackhole
                 0.78      0.001      0.374     0.78      0.505      0.992    IRC BOT
                 0.983     0.073      0.641     0.983     0.776      0.981    kelihos_c
                 0.932     0.033      0.679     0.932     0.786      0.988    cutwail
                 0.927     0.009      0.334     0.927     0.491      0.989    zero acess
Weighted Avg.    0.802     0.024      0.876     0.802     0.816      0.957

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
 48648     0  2539     7   167     4  2736     1   692 |     a = non-malicious
     0     0     0     0     0     0     0     0     1 |     b = Non-malicious
    32     0 21807    42  7895    17  4107  3510   317 |     c = tbot
     0     0     0    10     0    90     1    31     1 |     d = kelihos_a
     0     0     0     1  2705     0   421     1     0 |     e = blackhole
     0     0     0     2     0    92     0    20     4 |     f = IRC BOT
     0     0   148    13     1     0 13687    76     0 |     g = kelihos_c
     0     0     0    11   104    42   403  7718     2 |     h = cutwail
     0     0     0     4    29     1     4     2   510 |     i = zero acess
```

Figure 10: Classification results of Naïve Bays Classifier

In this research, it is observed that by applying a single technique and approach cannot solve the problem of botnet detection as in this implementation, there is a scope to add the features of DGA and fast flux class of botnet families. The advance botnets are more encrypted and stealth in nature, thereby detection technique for these class of bot malware need to be incorporated into this framework. Another limitation of this research is that more advance set of malwares such APT are not analysed during. The consolidation of the research data to visualize them as GUI interface is also remained developmental work.

ACKNOWLEDGMENT

A grateful thank to Director, CDAC-Mohali for his continuous encouragement and support in this work, the entire Cyber Security Technology Division (CSTD), CDAC-Mohali which provide the timely support and access of threat repository of malwares.

REFERENCES

[1] Elaheh Biglar Beigi, Hossein Hadian Jazi, et.al “Towards effective features selection in machine learning based botnet detection approaches” at IEEE conference on communication and network security, 2014.

[2] Basam Sayad, wei lu,et. Al “Botnet detection based on traffic behaviour analysis and flow intervals” computer and security technology, nov 2013.

[3] P.Narang, J.M.Reddy, and C.Hota, “Features selection for detection of peer to peer botnet traffic” in proceeding of the 6th ACM India computing convention. ACM , 2013, p.16.

[4] M.Stevanovic and J.M.Pederson, “Machine learning for identifying botnet network traffic”, Networking and security, Department of electronic Systems, Aalborg university, tech. Rep. 2013.

[5] L.Bilge, D.B Alzarotto, W.R obbertson, E .Kirda”Disclosure: Detecting botnet command and control servers through large scale net flow analysis” in Proceeding of the 28th Annual Computer Security Application Conference. ACM, 2012, PP, 129-128.

[6] Gregory Fedynyshyn, Mooi Choo Chuah, et. Al “Detection and classification of different botnet C&C channels” Lehign university, Bethlehem, PA, USA, 2012.

[7] Peter Ekstrand Berg, Katrin Franke, et. Al “Generic feature selection measure for botnet malware detection” computer science and media technology, Gjovik university, Norway, 2013.

[8] Matija Stevanovic and Jens Myrup Pedersen “ An analysis of network traffic classification for botnet traffic” Aalborg university, 2013.

[9] David Zhao, Issa Traore, et. Al “Botnet detection based on traffic behaviour analysis and flow intervals” at computer and security technology, 2016.

[10] Yao Zhao, Fang Yu, et. Al "Botgraph: Large scale spamming botnet detection" Microsoft research silicon valley, northwestern university, 2013.

[11] Raihana Syahirah Abdullah, Mohd Faizal Abdollah, et. Al "Preliminary study of host and network analysis on P2P botnet detection" on international conference on technology, informatics, management, engineering & environment, banding, Indonesia, june 23-26, 2016.

[12] Constantinn Musca, Emu Mrica, Razvan Deaconescu, et. Al " Detecting and analysing zero-day attacks using honeypots" on 19th international conference on control system and computer science,2013.

[13] Gregory Fedynyshyn, Mooi Choo Chuah , et. Al " Detection and classification of different botnet C&C channels" at Lehigh university. Bethlehem, PA 18015, USA,2010.

[14] Yao Zhao, Yinglian Xie, et. al "Botgraph: Large scale spamming botnet detection" at northwest university, Microsoft research silicon valley, Microsoft corporation, 2010.

[15] Guofei gu, Vinod yegeswaran et. Al "Active botnet probing to identify obscure command and control channels" at annual computer security application conference, 2009.

[16] Hossein Rouhani Zeidanloo, Parnian Najafi Borazjani " Botnet detection based on common network based behaviour by utilizing artificial immune system (AIS) " at 2nd international conference on software technology and engineering (ICSTE), vol 21-25,2010.

[17] T.T.Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning ", IEEE communication Syrveys & Tutorials, vol. 11 , no. 3, pp, 37-52,2009.

[18] S.Saad, I.Trore,A. Ghorbani, et al. "Detecting p2p botnets through network behaviour analysis and machine learning " Privacy Security and Trust , ninth annual international conference on IEEE, 2011, pp 174-180.

[19] Xiaonan Zang, Athichart Tangpong, et.al "Botnet Detection Through Fine Flow Classification", The Pennsylvania State University, jan 2011.

[20] Udaya Wijesinghe, Udaya Tupakula, Vijay Varadharajan," An Enhanced Model for Network Flow Based Botnet Detection", 38th Australasian Computer Science Conference (ACSC 2015), Sydney, Australia,27 - 30 January 2015.

[21] Jean-Pierre Charles, Anders Furuskar, et.al." Refined Statistical Analysis of Evolution Approaches for Wireless Networks",on IEEE Transation on Wireless Communication, nov 2015.

[22] Wireshark, http://www.wireshark.org.

[23] Guofei Gu et.al, BotHunter: Detecting Malware Infection Through IDS-Driven Dialog Correlation

[24] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "Botminer: Clustering analysis of network traffic for protocol- and structure independent Botnet detection," in Proc. 17th USENIX Security Symposium, 2008

[25] J.S.Bhatia, Rakesh Sehgal and Sanjeev Kumar, "Botnet Command Detection using Virtual Honeynet", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.5, Se p 2011,pp. 177-189 DOI :10.5121/ijnsa.2011.3514

[26] Sanjeev Kumar, Rakesh Sehgal and J.S. Bhatia, "Hybrid Honeypot Framework for Malware Collection and Analysis", 7th International Conference on Industrial and Information Systems (ICIIS-2012), August 6-9, 2012, IIT Chennai, Published in IEEE Xplore.

[27] Sanjeev Kumar, Paramdeep Singh, Rakesh Sehgal, and J.S.Bhatia, "Distributed Honeynet System using Gen III Virtual Honeynet" International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012, IJCTE 2012 Vol.4(4): 537-541 ISSN: 1793-821X

[28] Brumley, D., Hartwig, C., Liang, Z., Newsome, J., Poosankam, P., Song, D., and Yin, H. 2007: Automatically identifying trigger-based behavior in malware.In: Book chapter in Botnet Analysis and Defense

[29] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "Botminer: Clustering analysis of network traffic for protocol- and structure independent Botnet detection," in Proc. 17th USENIX Security Symposium, 2008

[30] "Malware targeting SOHO devices" [Online]. Available at: http://www.teamcymru.com/ReadingRoom/Whitepapers/2013/TeamCymruSOHOPharming.pdf

[31] "Attacks on wireless routers " [Online]. Available at: http://arstechnica.com/security/2014/03/hackers-hijack-300000-plus-wireless-routers-make-malicious-changes/

[32] "Top 10 IoT Vulnerabilities (2014) Project". [Online]. Available at: *https://www.owasp.org/index.php/Top_10_IoT_Vulnerabilities_(2014)*

[33] http://thehackernews.com/2016/06/cctv-camera-hacking.html

[34] ČELEDA, Pavel, Radek KREJČÍ, Jan VYKOPAL a Martin DRAŠAR. Embedded Malware - An Analysis of the Chuck Norris Botnet. In European Conference on Computer Network Defense. 1. vyd. Los Alamitos, CA: IEEE Computer Society, 2010. s. 3 -10, 8 s. ISBN 978 -1 -4244 -9377 -7. doi:10.1109/EC2ND.2010.15

[35] "Lizard stresser runs on hacked CCTV cameras — Krebs on Security." [Online]. Available:http://krebsonsecurity.com/2015/01/lizard-stresserruns-on-hacked-home-routers/

[36] "Daily Tech - Hackers Use Refrigerator, Other Devices to Send 750,000 Spam Emails." [Online] Available: http://www.dailytech.com/Hackers+Use+Refrigerator+Oter+Devices+to+Send+750000+Spam+Emails+/articl e34161.htm.

[37] "PSYB0T Information Page," 2009, [Online]. Available at: http:// baume.id.au/psyb0t.

[38] "Network Bluepill - stealth router-based botnet has been DDoSing dronebl for the last couple of weeks," [Online]. Available: http://www.dronebl.org/blog/8

[39] "Botnet targeting IoT devices"[Online]. Available at: http://www.bluecoat.com/security-blog/2015-01-09/botnet-internet-things

[40] "Malware targeting SOHO devices" [Online]. Available at: http://www.teamcymru.com/ReadingRoom/Whitepapers/2013/TeamCymruSOHOPharming.pdf

[41] "Attacks on wireless routers " [Online]. Available at: http://arstechnica.com/security/2014/03/hackers-hijacssk-300000-plus-wireless-routers-make-malicious-changes/

[42] "Top 10 IoT Vulnerabilities (2014) Project". [Online]. Available at: https://www.owasp.org/index.php/Top_10_IoT_Vulnerabilities_(2014)

[43] "Attacks targeting internet of things [Online]. Available at: "https://www.bluecoat.com/security-blog/2015-01-09/botnet-internet-things

[44] "CCTV DDoS botnet attacks" [Online]. Available: https://www.incapsula.com/blog/cctv-ddos-botnet-back-yard.html

[45] "Malware targeting broadband routers" [Online]. Available at: http://www.pcworld.com/article/2098160/worm-themoon-infects-linksys-routers.html

[46] "The moon malware targeting broadband routers" [Online]. Available at: http://arstechnica.com/security/2014/02/bizarre-attack-infects-linksys-routers-with-self replicating-malware/

[47] https://www.r-project.org/