HTTP Botnet Detection using Data Mining Approach

Bhautik Trivedi[#1], Zishan Noorani[#2]
[#] *CE Dept (M.E.)-LDCE, Ahmedabad ,*CE Dept-LDCE, Ahmedabad*
Gujarat Technological University, Ahmedabad

***Abstract*— Among the diverse forms of malware, Botnet is the most widespread and serious threat which occurs commonly in today's cyber-attacks. A botnet is a group of compromised computers which are remotely controlled by hackers to launch various network attacks, such as DDoS attack, spam, click fraud, identity theft and information phishing. Botnet has become a popular and productive tool behind many cyber-attacks. The defining characteristic of botnets is the use of command and control channels through which they can be updated and directed. Recently malicious botnets evolve into HTTP botnets out of typical IRC botnets. Data mining algorithms allow us to automate detecting characteristics from large amount of data, which the conventional heuristics and signature based methods could not apply.**

***Keywords*— Botnet, Botnet Detection, Apriori, HTTP Botnet, DDOS**

I. Introduction

The improvement and advancement in network bandwidth and computing, parallel and distributed computing are widely accepted. So they have been the obviously targeted by hackers [1]. Botnet is a collection of internet-connected computers whose security defences have been breached and control ceded to a malicious party, blackhat community. The groups of compromised computers are controlled by one or group of attacker known as "Botmaster" [2]. Botnet operators can use the aggregated power of many bots to exponentially raise the impact of those dangerous activities. A single bot might not be a danger for the Internet, but a network of bots certainly is able to create huge malfunctioning. A study shows that, on a typical day, about 40% of the 800 million computers connected to the Internet in a botnet in year 2008 [3]. Communication, resource sharing and curiosity have been great motivators for underground research and hacking. The major attacks under Botnet are, DDos, Scanning, Phishing, Click fraud, spamming [4].

II. Background and Related work

The untraceable feature of coordinated attacks is just what hackers/attackers demand to compromise a computer or a network for their illegal activities. Once an attack is initiated by a group of computer nodes having different locations controlled by a malicious individual or controller, it may be very hard to trace back to the origin due to the complexity of the Internet [9]. Because of these reasons it has become very serious problem now a day.

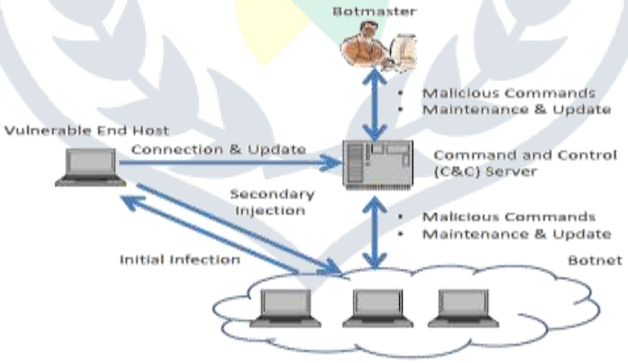


Fig. 1 Botnet Life Cycle [11]

Botnet Life Cycle basically invokes 5 Steps. 1) Initial infection 2) Secondary Injection 3) Connection 4) Command & Control 5) Update and Maintenance [20] .
Botnet Basically characterized in 3 forms based on Communication protocol they are using as C&C Server
1) IRC based 2) HTTP based 3) P2P based.

IRC and HTTP are known as Centralized C&C where as P2P is Decentralized C&C. For Centralized C&C, if it is mitigated then whole Botnet will be closed and BotMaster has to create new C&C Server and Botnet.

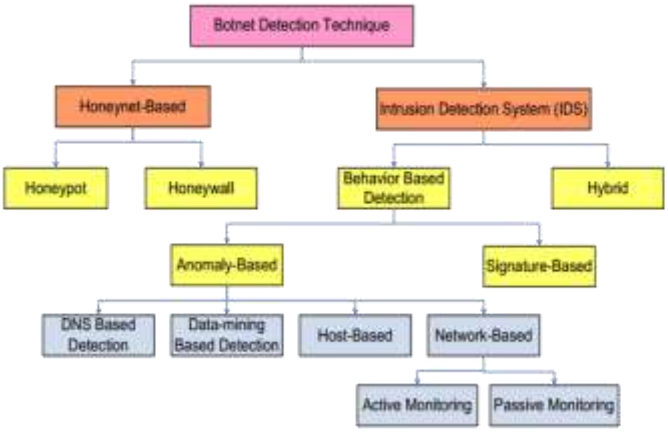


Fig. 2 Botnet Detection Techniques [21]

Botnet connection has specific network flow characteristics in which frequent communication happens between C&C and infected machine.

TABLE 1: COMPARISON OF RELATED WORK ON BOTNET DETECTION

Authors	Proposed Method	Advantages	Disadvantage
Jae-Seo Lee, HyunCheol Jeong, Jun-Hyung Park et. al. [7]	Relationship of HTTP Client to HTTP server via analysis of periodic repeatability. Standard deviation(DPR) will be low for potential Bot.	Overcome existing DNS filtering method.	High false positive ratio as no separation of normal traffic and bot traffic
Roberto Perdiscia et al [13]	network-level behavioural malware clustering system, structural similarities among malicious HTTP traffic	HTTP method like, GET, POST analyzed	HTTP request and response encryption is major limitation.
Masayuki Ohrui [10]	Apriori and PrefixSpan algorithms	Apriori deals with subset of events without considering the order of events	High false positive ratio. Exclusion of expected output
Sajjad Arshad [12]	Cluster Based Data Mining Tech.	Better with normal data flow	Hard to work with large database and if large number of whitelist domains are included.
Zang [16]	Hierarchical and K mean clustering	Operate on Flow level internet traffic.	Need to handle large level of flow data.
Mazzariello [18]	Classification based data mining algorithm. Uses Support Vector Machine (SVM) and J48 decision trees	perfect separation of botnet C&C traffic from normal one	dependence on the predefined IRC models, limits the effective detection among different types of botnets and new botnets

From the previous studies and research work it can be concluded that Botnet connection has specific network flow characteristics in which frequent communication happens between C&C and infected machine. To detect anomaly attack, network flow analysis is the best approach [13]. Any tool or method cannot detect a Botnet Communication as real time processing as there is no specific signature.

Data mining and machine learning techniques are easily applicable on network flow information. Flow data have a structured and related nature, which do not require massive preprocessing tasks. Besides, flow information implies patterns inside, which makes data mining algorithms convenient and effective for analysis. Clustering, Classification, Fuzzy Logic and AI related approach already been implemented for botnet detection.

III. PROPOSED WORK AND FLOWCHART

Here is the proposed solution to detect HTTP based C&C center from the Botnet. The target would be to analyze and mining the logs from infected machine. Proposed Flow chart involves 4 basic steps.

1) Network Traffic: Traffic monitoring is responsible to detect the group of hosts that have similar behaviour and communication pattern by inspecting all network traffic. Each flow record has following information: Source IP address, Destination IP address, Source Port, Destination Port, Number of packets transferred in both directions, Time of packet received and transferred to particular IP address.

2) Filtering: Filtering is responsible to filter out irrelevant traffic flows. The main objective of this part is to reduce the traffic workload and makes the rest of the system perform more efficiently. This step will reduce bot irrelevant traffic. If this is not

performed well then it can raise false positive rates. Some common and useful Filtering criteria likes, eliminate all port-scan activities & Filter out based on black lists and white lists.

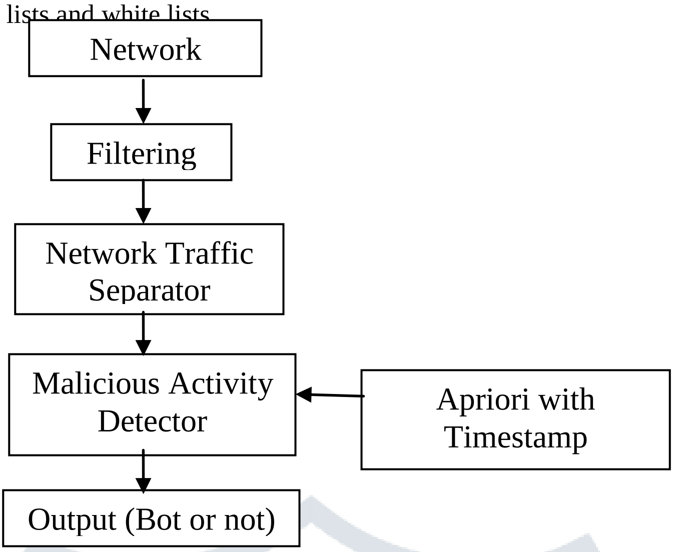


Fig. 3 Proposed Botnet Detection Flow Chart

3) Network Traffic Separator: After Filtering network traffic contains many different types of packets, like, TCP, DHCP, Broadcast, Local Network packets etc. Some packets from this traffic are not part of Botnet so they should be taken off from the network traffic. Network Traffic Separator is responsible to separate HTTP traffic from the rest of traffic and sends them to centralized part. Like most network protocols, HTTP uses the client-server model and HTTP protocol.

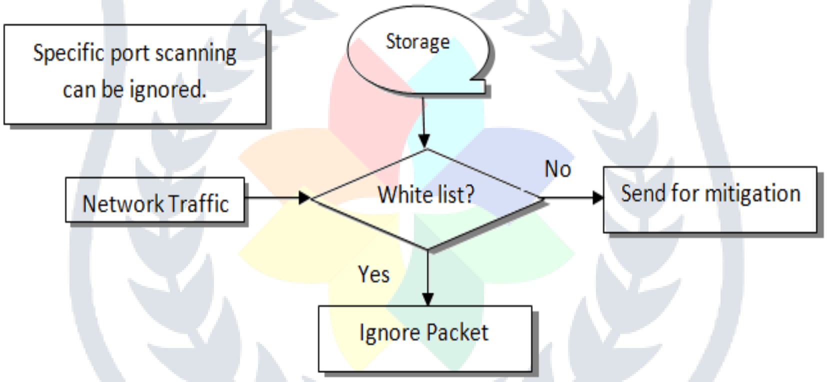


Fig. 4 Network Filtering Flowchart

4) Malicious Activity Detector: This part will detect malicious activity generated by Botmaster. Data mining technique is used for extracting suspicious activity. Apriori is a well-known algorithm for association rule discovery. The Apriori can be used to detect the association rule for the botnet detection. It was designed to detect significant correlation of set of items for extracting rules of items with high support (a fraction of the subset of items). The support is useful feature for detecting all possible coordinated behaviours among servers. However, since Apriori deals with subset of events without considering the order of events, it has high false positive ratio. For instance, a sequence of events x and then y is equivalent to one of y and x in Apriori. The detected patterns in Apriori contain some false coordination that two independent servers happened to work at almost same time by chance. Hence, its confidence is not so high. So, Timestamp is used to find order of event for association rule generated by Apriori.

How Apriori Works

1) Find all frequent itemsets: a) Get frequent items: Items whose occurrence in database is greater than or equal to the min.support threshold. b) Get frequent itemsets: Generate candidates from frequent items. Prune the results to find the frequent itemsets.
2) Generate strong association rules from frequent itemsets: Rules which satisfy the min.support and min.confidence threshold.

IV. Implementation of Problem and Experimental results

The Botnet DDOS attack has been implemented using Bonesi 0.2.0 [22] DDOS attack simulator. Initially for testing method only 4 IP addresses has been taken to form the DDOS attack from 4 distinct IP addresses. The data log has been generated using wireshark network simulator.

After network filtering process these data has been passed through apriori method to detect the frequent pattern sets. Below it shows the result with min support 20 and confidence 50. Training dataset is shown in Table 2.

```
192.168.104.14 <-  (100, 94)
192.168.104.14 <- 253.99.92.111 (22, 100)
```

192.168.104.14 <- 244.174.48.40 (24, 100)
192.168.104.14 <- 222.109.217.68 (24, 100)
192.168.104.14 <- 214.221.36.5 (24, 100)

These data indicates that IP address 253.99.92.111 frequently communicate with 192.168.104.14 which is infected machine. Same applies to other IP addresses as well.

TABLE 2: SAMPLE DATA

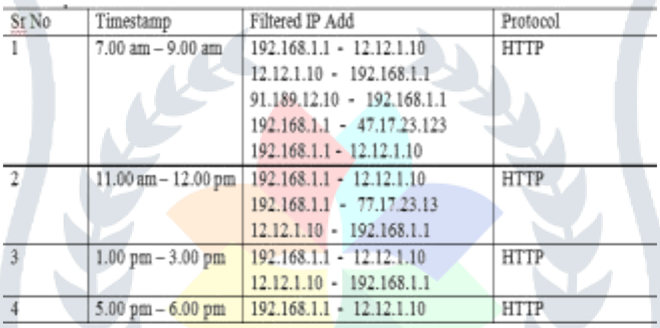
The same method applies to check the scalability of the approach and to enhance the detection technique's accuracy and decrease false positive ratio Source Port and Destination port has been also consider as items to apriori algorithm. Here is the experimental result. This was tested and analyzed to mitigate DDOS attack on a host machine.

Normally C&C server communicates to malicious host frequently and may be for specific duration and on specific time. So to detect such kind of C&C following method can be applied.

Source	Destination	Protocol
214.221.36.5	192.168.104.14	TCP
192.168.101.28	192.168.255.255	NBNS
253.99.92.111	192.168.104.14	TCP
244.174.48.40	192.168.104.14	TCP

TABLE 3: EXPERIMENTAL RESULT PROVIDES SCALABILITY FOR DDOS ATTACK

For the Botnet C&C server detection this approach can be applicable. To fulfill this approach, data has been generated manually as no dataset repository provides such kind of data. The data has been prepared for consecutive 3days and around 4000 records have been created.



Sr No	Timestamp	Filtered IP Add	Protocol
1	7.00 am – 9.00 am	192.168.1.1 - 12.12.1.10 12.12.1.10 - 192.168.1.1 91.189.12.10 - 192.168.1.1 192.168.1.1 - 47.17.23.123 192.168.1.1 - 12.12.1.10	HTTP
2	11.00 am – 12.00 pm	192.168.1.1 - 12.12.1.10 192.168.1.1 - 77.17.23.13 12.12.1.10 - 192.168.1.1	HTTP
3	1.00 pm – 3.00 pm	192.168.1.1 - 12.12.1.10 12.12.1.10 - 192.168.1.1	HTTP
4	5.00 pm – 6.00 pm	192.168.1.1 - 12.12.1.10	HTTP

Fig. 5 Training Dataset based on timestamp

Apriori is used to generate frequent itemsets for request machine address and timestamp is used for sequence of patterns. Left side of rule represents request machine IP address and right side of rule represents local machine address, here in our case it is 192.168.1.1. For 3 consecutive days data has been captured and Apriori detects following patterns for botnet detection as shown in Figure 6. Rules with highest support and confidence are kept on particular timestamp.

The experimental result in Table 4 shows that IP address 12.12.1.10--192.168.1.1 and 45.17.44.123-- 192.168.1.1 have high support and confidence in all three days with minimum support 20 and minimum confidence 20. So, Request machine IP addresses 12.12.1.10 and 45.17.44.123 could be malicious C&C botnet server.

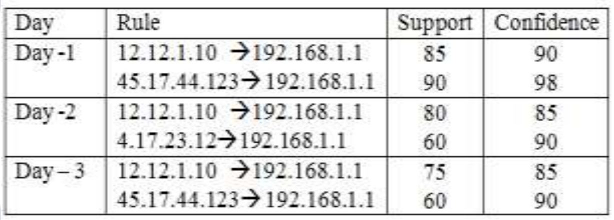
	Frequent Item	Frequent Item	Support	Confidence
253.99.92.111	192.168.104.14	253.99.92.111	17.2456	100
	253.99.92.111	192.168.104.14	98.4275	17.5211
	80	253.99.92.111	17.2456	100
	253.99.92.111	80	99.9942	17.2465
	32554	253.99.92.111	17.2456	100
	253.99.92.111	32554	99.9942	17.2465
244.174.48.40	192.168.104.14	244.174.48.40	17.2571	100
	244.174.48.40	192.168.104.14	98.4275	17.5328
	80	244.174.48.40	17.2571	100
	244.174.48.40	80	99.9942	17.2581
	32554	244.174.48.40	17.2571	100
	244.174.48.40	32554	99.9942	17.2581

214.221.36.5	192.168.104.14	214.221.36.5	17.2571	100
	214.221.36.5	192.168.104.14	98.4275	17.5328
	80	214.221.36.5	17.2571	100
	214.221.36.5	80	99.9942	17.2581
	32554	214.221.36.5	17.2571	100
	214.221.36.5	32554	99.9942	17.2581

TABLE 4: EXPERIMENTAL RESULT BASED ON TIMESTAMP

Day	Rule	Support	Confidence
Day -1	121.211.12.14 --192.168.1.1	75	25
	12.12.1.10 --192.168.1.1	90	85
	217.77.223.12--192.168.1.1	45	56
	66.220.1.1--192.168.1.1	55	60
	116.12.31.11-- 192.168.1.1	30	45
	45.17.44.123-- 192.168.1.1	90	95
	91.189.33.112-- 192.168.1.1	45	25
Day -2	146.21.34.122--192.168.1.1	25	50
	12.12.1.10 --192.168.1.1	80	85
	195.145.43.23 --192.168.1.1	60	30
	76.71.12.14--192.168.1.1	50	25
	4.17.23.12--192.168.1.1	60	90
	73.224.3.12--192.168.1.1	45	21
Day – 3	12.12.1.10 --192.168.1.1	75	85
	45.17.44.123-- 192.168.1.1	60	90
	76.71.12.14--192.168.1.1	91	77
	56.67.12.56--192.168.1.1	50	50
	66.220.1.1--192.168.1.1	40	20

The experimental results in Figure 6 shows that pattern 12.12.1.10 --192.168.1.1 has high support and confidence in all three days with minimum support 60 and minimum confidence 60. So, Request machine IP address 12.12.1.10 could be malicious C&C botnet server.



Day	Rule	Support	Confidence
Day -1	12.12.1.10 →192.168.1.1	85	90
	45.17.44.123→192.168.1.1	90	98
Day -2	12.12.1.10 →192.168.1.1	80	85
	4.17.23.12→192.168.1.1	60	90
Day – 3	12.12.1.10 →192.168.1.1	75	85
	45.17.44.123→192.168.1.1	60	90

Fig. 6 Final Result with support & confidence 60

V. CONCLUSION AND FUTURE WORK

In recent era, there are so many research work has been done for P2P and IRC botnet. The motivations for using the HTTP protocol are multiple. Developing a web-based C&C application is typically easier than implementing customized C&C communication protocols (e.g., peer-to-peer protocols), and there is evidence that web-based "reusable" kits (or platforms) for botnet C&C are available for sale on the Internet.

Data mining techniques are easily applicable on network flow information. Flow data have a structured and related nature, which do not require massive preprocessing tasks. Besides, flow information implies patterns inside, which makes data mining algorithms convenient and effective for analysis. Association rule based data mining approach can be an effective solution to botnet detection as it can generate frequent pattern sets from the flow information. The high support and confidence implies the string association rule i.e strong correlation between the item i.e IP addresses.

In future research work can be carried out on the other flow parameters like, response time, particular time on which request has been made, and more work on source port and destination port. The filtering process can be implemented dynamically so overhead of manual task can be eliminated.

REFERENCES

[1] Chung-Huang Yang , Kuang-Li Ting. *Fast Deployment of Botnet Detection with Traffic Monitoring,* Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pages 856-860, 2009.

[2] Haritha S. Nair, Vinodh Ewards S E *A Study on Botnet Detection Techniques*, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012

[3] *Botnet scams are exploding*, 2008 http://usatoday30.usatoday.com/tech/news/computersecurity/2008-03-16-computer-botnets_n.htm.

[4] Nicholas Ianelli, Aaron Hackworth. *Botnets as a Vehicle for Online Crime* - CERT and CERT Coordination Center are registeredin the U.S. Patent and Trademark Office. 2005

[5] *Oregon Man Cops Plea in eBay DDOS Attack*, http://www.internetnews.com/security/article.php/3574101

[6] *Worm strikes down Windows 2000* systems,http://www.cnn.com/2005/TECH/internet/08/16/computer.worm/index.html

[7] Kraken botnet, Wikipedia, http://en.wikipedia.orglwikilKraken_botnet, 2008.

[8] Zeus botnet steals $47M from European bank customers,2012. http://news.cnet.com/8301-1009_3-57557434-83/zeus-botnet-steals-$47m-from-european-bank-customers/

[9] Erdem Alparslan, Adem Karahoca and Dilek Karahoca. *BotNet Detection: Enhancing Analysis by Using Data Mining Techniques*, Downloaded from http://dx.doi.org/10.5772/48804 (BOOK)

[10] Sonal P.Patil, Swatantra Kumar. *Botnet-A Network Threat*, International Conference on Recent Trends in Information Technology and Computer Science (IRCTITCS), Pages 29-35, 2011.

[11] Xiaonan Zang, Athichart Tangpong, George Kesidis and David J. Miller. *Botnet Detection Through Fine Flow Classification*. Departments of CS&E and EE, The Pennsylvania State University, University Park, PA, 16802. CSE Dept Technical Report No. CSE11-001, Jan. 31, 2011

[12] Alireza Shahrestani, Maryam Feily, Rodina Ahmad, Sureswaran Ramadass. *ARCHITECTURE FOR APPLYING DATA MINING AND VISUALIZATION ON NETWORK FLOW FOR BOTNET TRAFFIC DETECTION*, International Conference on Computer Technology and Development,IEEE, Pages 33-37 2009.

[13] Zhang yanyan,Yao Yuan, *Study of Database Intrusion Detection Based on Improved Association Rule Algorithm*, IEEE. Pages 673-676, 2010.

[14] Sajjad Arshad, Maghsoud Abbaspour, Mehdi Kharrazi, Hooman Sanatkar. *An Anomaly-based Botnet Detection Approach for Identifying Stealthy Botnets*, Presented in International Conference on Computer Application Industrial Electronics, IEEE, Pages 564-569, 2011

[15] Wang Zilong, Wang Jinsong, Huang Wenyi, Xia Chengyi. *The Detection of IRC Botnet Based on Abnormal Behavior*. Second International Conference on MultiMedia and Information Technology, IEEE, Pages 146-149, 2010.

[16] J. Goebel and T. Holz. *Rishi: Identify bot contaminated hosts by irc nickname evaluation* In USENIX Workshop on Hot Topics in Understanding Botnets (HotBots 07), 2007.

[17] Claudio Mazzariello. *IRC traffic analysis for botnet detection*, The Fourth International Conference on Information Assurance and Security, IEEE, Pages 318-323, 2008.

[18] Hossein Rouhani Zeidanloo, Mohammad Jorjor Zadeh shooshtari, Payam Vahdani Amoli, M. Safari, Mazdak Zamani. *A Taxonomy of Botnet Detection Techniques*, IEEE, Pages 158-162, 2010

[19] Roberto Perdiscia, Wenke Leea, and Nick Feamstera, *Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces*, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA, USENIX, 2010

[20] Jignesh Vania, Arvind Meniya, H.B. Jethva, “A Review on Botnet and Detection Technique”, International Journal of Computer Trends and Technology- volume4Issue1- 2013, Pages 23-29

[21] Raihana Syahirah Abdullah et al., “Revealing the Criterion on Botnet Detection Technique”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 3, March 2013, Pages 208-215

[22] Bonesi the DDOS Botnet Simulator available from https://code.google.com/p/bonesi/