Applying Data Mining Techniques in Cyber Crimes

Mohiuddin Ali Khan[1] , Sateesh Kumar Pradhan[2], Huda Fatima[3]
[1 & 2]Dept. of Computer Science, Utkal University, Bhubaneshwar, India
[3]Dept. of Computer Science, Sambalpur University, Orissa, India
moinkku@gmail.com[1], sateesh1960@gmail.com[2], hudafatima@gmail.com[3]

***Abstract*— Globally the internet is been accessed by enormous people within their restricted domains. When the client and server exchange messages among each other, there is an activity that can be observed in log files. Log files give a detailed description of the activities that occur in a network that shows the IP address, login and logout durations, the user's behavior etc. There are several types of attacks occurring from the internet. Our focus of research in this paper is Denial of Service (DoS) attacks with the help of pattern recognition techniques in data mining. Through which the Denial of Service attack is identified. Denial of service is a very dangerous attack that jeopardizes the IT resources of an organization by overloading with imitation messages or multiple requests from unauthorized users.**

Keywords: Denial of Service, Log File, Cyber Crimes, Data mining, outliers, Association rules.

I. CYBER SECURITY

Cyber Security is that branch of Computer Technology that deals with security in cyberspace. Cyberspace refers to the description of policies regarding the networks and computer systems. The policies laid out in the Cyber security are for the reason of avoiding the malicious activity or unauthorized access to secured information. Since the emergence of high structured networks [1], there arises a concern about how intelligently these networks are secured. These issues are major concerns in the internet era. Cyber security [3] is concerned with protecting IT resources like server; network etc. from performing illegal activities or fradulent acts. Data mining is also applicable to problem solving or network intrusions. Therefore in this paper we focus the applications of data mining for cyber security applications.

II. CYBER CRIMES

Cyber refers to something that can be done on internet. Crime refers to something that is done illegally or without authorization. All those crimes that are done on the internet in order to gain access to secured information or authorization rights is termed as ''Cyber Crime''. Globally the cyber-crime hindrance is spread across abundantly.

In our research paper we are focusing the detection of the Denial of Service (DoS) attacks using Data Mining techniques. This will jeopardize the network and IT resources by artificially increasing the network traffic and load on the server by sending imitation requests.

III. TYPES OF CYBER CRIMES:

- Hacking: It is the unauthorized access to a computer system that usually modifies the computer hardware and software configurations.
- Malware: Refers to malicious software, which means gaining information without prior permission.
- Virus and Worms: Small programs that are attached to the files and spread to the computers whereas worms make copies of themselves to corrupt the computers.
- Denial of service: When a website takes too long to function effectively.
- Spam: Unintended or unwanted emails that reach in bulks to email accounts is Spam
- Software Piracy: Making copies or software without prior permission and performing commercial acts.

IV. DENIAL OF SERVICE

In network architecture, the networking strategies should be made prone to identify intruders that attack the system (or cause a denial of service attack). DoS attacks [6] are some of the oldest Internet threats and continue to be the top risk to networks around the world. DoS events making it difficult to detect them. DoS attack remains a serious problem that increasingly affects company resources, in most cases a DoS attack caused the services to be completely unavailable impacting their business directly which is a potential financial loss to businesses[5].

According to [4], the number of DoS attacks durations by the end of the year 2015 became shorter and more discreet. The figure 1 below shows the ups and downs of DoS attacks.

978-1-5090-5814-3/17/$31.00 ©2017 IEEE

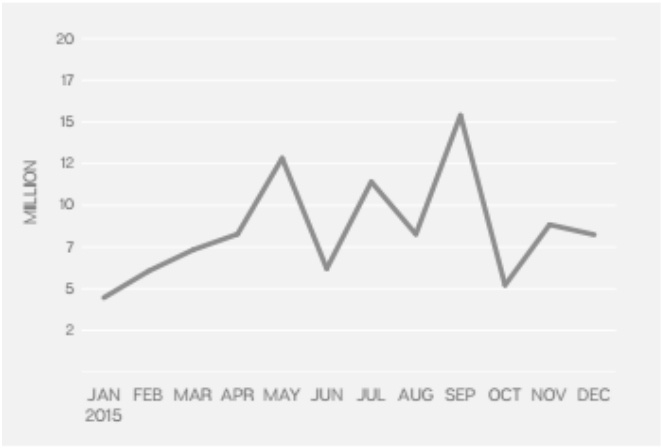


Fig. 1 Ups and downs of the Denial of Service attacks.

A. *Vulnerabilities*

Vulnerabilities refers to the weaknesses that exist in network architecture due to poor infrastructure mechanisms. Cyber attackers are solely grabbing opportunities to look for vulnerabilities in network architecture. The existing websites, approximately 75% are prone to vulnerabilities that are unpatched.

Intruders are yet continuing to take advantages of the underlying vulnerabilities of the network architectures, exploiting weaknesses in framed encryption systems that ideally allow intruders in all connections. This consequently affects the entire network architecture of the organization. The fig. 2 below [4] shows the decrease of vulnerabilities in the consequent years from 2006 to 2015.

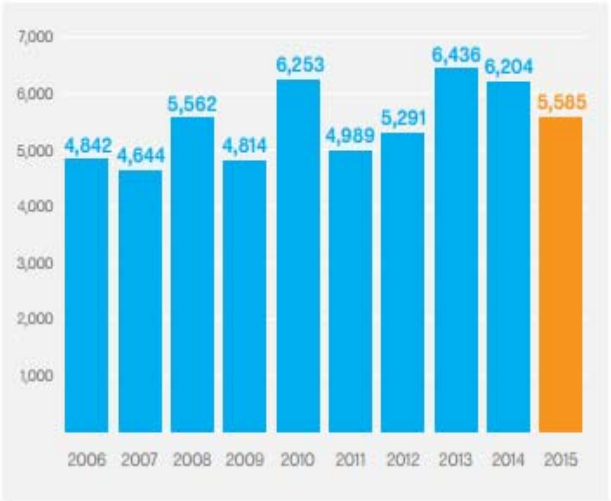


Fig. 2 Existence of Vulnerabilities

B. *Better protection*

Network architecture must look in for better security mechanisms rather than the traditional network architecture. Secured techniques should be applied in a network architecture using the updated technologies to make the network free from attacks. For this reason, the focus of attention should be drawn towards the IT departments of organizations to keep them active by monitoring the upcoming risks and threats. Security needs to start digging through the data proactively during non-breach response time.

V. DATA MINING

Data Mining [7] emphasizes the extraction of data from databases and various patterns can be concluded for deriving association rules. Although Data Mining is eventually gaining a wider scope in different areas, its research has made remarkable significance in Cyber Crimes.

Data mining, which is defined as the process mining[8] or extracting data into productive information. Based on this data, significant patterns are formed.

A. *Association Rules*

The term association means, "connectivity" or "together" or frequently that appears. Therefore association rules are related to those conditions where the values in a data set are frequently appearing and this appearance will show relationship or "connectivity" among the values. For this reason, in order to show the relationship, we assign a support or threshold value and henceforth the association rules are generated based on the algorithm adopted, like "Apriori" or "K-means" etc.

B. *Cluster Analysis*

The term "cluster" refers to "groups" or categorizing the data into separate labels. These labeled classes hold similar type of data. Henceforth it implies that data in different class labels differ from each other in terms of their features. When this data is analyzed in different classes or labels then using different analysis techniques, the data is extracted. This process is called as "Cluster Analysis". In fig. 3, the cluster analysis is shown along with outliers.

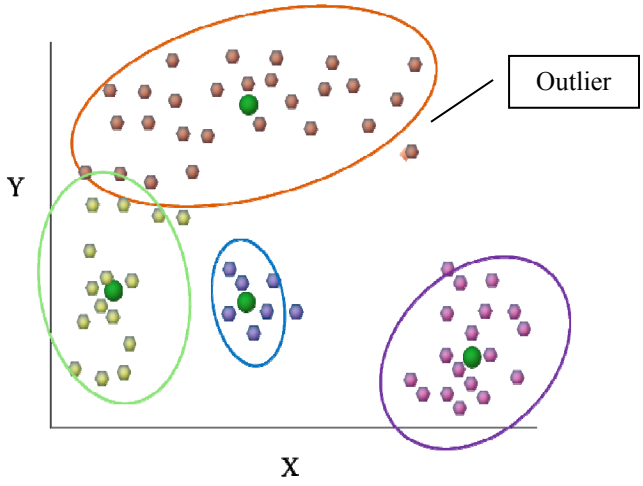


Fig. 3 Cluster Analysis

VI. OUTLIERS

The term outlier refers to those readings in a data that are comparatively different from the rest of the data. Data mining techniques for detecting outliers will be extensively used for spotting abnormal behavior.

When the outliers are discovered, their existence can signify important information. This information has helped many crime agencies, banks with unpredicted results. Outlier detection has been extensively studied in the recent years. Outlier is that concept that its existence is comparatively different from the remaining set of data. According to [2], outliers have been classified into two categories:

A. Classic Outlier approach

In this approach, whatever outliers are observed in a dataset on the basis of transactions. Nevertheless, problems persist to exist to apply mining techniques on data such as repetition of large data sets, availability of vague information etc.

B. Spatial Outlier approach

The term "spatial" means or refers to the objects that are present in space or have a geographical existence. Spatial Outlier refers to spatially referred object whose non spatial values are comparatively different in the same region.

VII. METHODOLOGY

In our research paper, we have shown the concept of data mining techniques to identify cyber-attacks. Our focus of attention would be on "finding patterns" in a log file *(records that occur in the system)* which shows the sequence of events. From this log file we identify patterns.

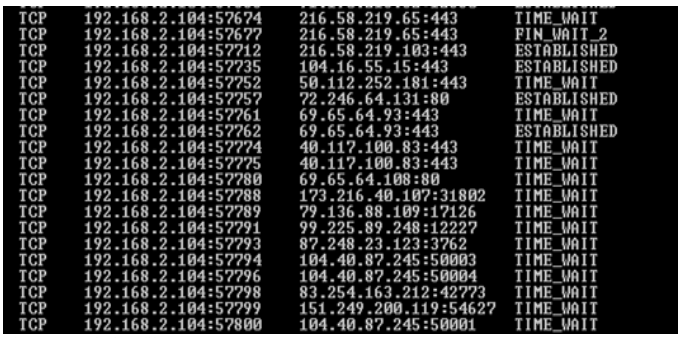
To start with, we use the clustering technique to discover the type of cyber-crime, Denial of service (DoS) attacks. As we know that clustering is grouping of data that has similar features. So this grouping helps to discover similar patterns of data that occur constantly in the log file.

Step 1: Evaluate the log file.
Step 2: Mine the date with time
Step 3: Scan the data
Step 4: Add the found data in the main file.

When the above procedure is carried out, we will record that data which contains normal patterns and also abnormal patterns (malicious). By using the clustering technique we identify the data that occur repeatedly [9].

System Configuration: In order to run our obtained data, we use the Windows Server to maintain the database. Initially we run the data that contains zero attacks and then add them to the master file or log file. The ICMP (Internet Control Message Protocol) will make the system inactive by sending voluminous amount of "ping" command. Now the data that contains the normal activities and the data that contains attacks are passed through the technique that we have proposed. If the observations of the log file show normal behavior then they will be ignored. If the observations show multiple requests of the same transaction, then this data will be directed through our algorithm "Apriori" and will be shown in the attack logs.

This algorithm will detect if similar patterns of requests exist in the normal records prior to consider it as attack. If the algorithm finds out the pattern and or finds the number of request for the same transaction more than the threshold value it is considered as an attack and it sends signal or message to the administrator about the suspected attack.



TCP	192.168.2.104:57674	216.58.219.65:443	TIME_WAIT
TCP	192.168.2.104:57677	216.58.219.65:443	FIN_WAIT_2
TCP	192.168.2.104:57712	216.58.219.103:443	ESTABLISHED
TCP	192.168.2.104:57735	104.16.55.15:443	ESTABLISHED
TCP	192.168.2.104:57752	50.112.252.181:443	TIME_WAIT
TCP	192.168.2.104:57757	72.246.64.131:80	ESTABLISHED
TCP	192.168.2.104:57761	69.65.64.93:443	TIME_WAIT
TCP	192.168.2.104:57762	69.65.64.93:443	ESTABLISHED
TCP	192.168.2.104:57774	40.117.100.83:443	TIME_WAIT
TCP	192.168.2.104:57775	40.117.100.83:443	TIME_WAIT
TCP	192.168.2.104:57780	69.65.64.108:80	TIME_WAIT
TCP	192.168.2.104:57788	173.216.40.107:31802	TIME_WAIT
TCP	192.168.2.104:57789	79.136.88.109:17126	TIME_WAIT
TCP	192.168.2.104:57791	99.225.89.248:12227	TIME_WAIT
TCP	192.168.2.104:57793	87.248.23.123:3762	TIME_WAIT
TCP	192.168.2.104:57794	104.40.87.245:50003	TIME_WAIT
TCP	192.168.2.104:57796	104.40.87.245:50004	TIME_WAIT
TCP	192.168.2.104:57798	83.254.163.212:42773	TIME_WAIT
TCP	192.168.2.104:57799	151.249.200.119:54627	TIME_WAIT
TCP	192.168.2.104:57800	104.40.87.245:50001	TIME_WAIT

Fig 4. Sample log file

In the fig. 4 as shown above, we could see the DoS attack that has been made by the anonymous user (intruder) initially by gaining the access to the system (server) by posing as a authenticated user. In denial of service attack, the attacker gains the access through the vulnerabilities present in the system and copies the message sent by an authenticated user and makes multiple copies of the same request or query and sends it to the server. So, the server will process the same query or the request sent by a user for multiple times. In this way, the server is kept busy by processing the same request multiple times. This is called as denial of service attack. Another example is the "ping" attack where multiple ping requests will be sent from one user or multiple users and the server is again overloaded with processing the same request. This type of attack is severe. We apply data mining techniques to identify these types of attacks by finding similar patterns or request from the users. In our approach, we define a threshold of minimum support (5). If the same request is received to the server more than the threshold value, it assumes it as an attack and notifies the administrator. In some cases, based on the working environment, the threshold value could be set accordingly.

Procedures:

Step 1: Start

Step 2: Let the Count=0, set the threshold value. The threshold value can be set based on the working environment.

Step 3: Check if the counts of matched rules have crossed the threshold value.

- If true, intimate the administrator assuming as an attack.
- If false, continue.

Step 4: Check whether new event is recorded in log file.

- If no new event found, wait
- If event_found, go to step 2

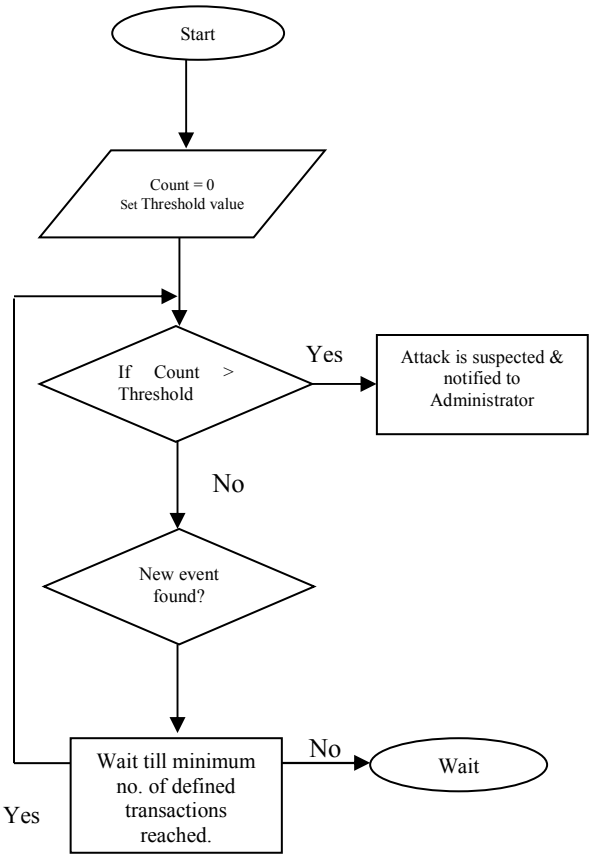


Fig. 5. Flowchart for detecting DoS attacks.

VIII. CONCLUSION

In this paper we have applied the data mining techniques for identifying the Denial of Service attack. This type of attack is very dangerous as it jeopardizes the IT resources. It makes the server busy by imitation messages and repeated queries. The server is congested by traffic packets, in order to mitigate the server performance. In this research paper, we have discussed about Cyber security, cyber-crimes their types, clustering, outliers and pattern recognition. We have applied the famous data mining technique called as pattern recognition on the log file. We set a threshold value. If the number of similar requests are received at the server, which is greater than the threshold value, we assume this as an attack and the administrator is been informed. By this approach we can identify the denial of service attack easily as in DoS attack, the attacker or the hacker sends same multiple requests in order to mitigate the server performance.

REFERENCES

[1] Know Your Enemy: Learning About Security Threats, 2nd Edition.ISBN: 0321166469. The Honeypot Project 2004.

[2] M.Khan , S.K.Pradhan, M.A.Khaleel, "Outlier Detection for Business Intelligence using data mining techniques", International journal of Computer Applications (0975 -8887), Volume 106- No. 2, November 2014.

[3] Masud, M.M, Gao,J.Khan, "Peer to Peer Botnet Detection for Cyber Security: A Data Mining Approach". In proceedings: Cyber-security and information Intelligence research workshop. Oakridge national Laboratory, Oakridge May 2008.

[4] Internet Security Threat Report, Volume 21, April 2016, Symantec Crime Report.

[5] Ibrahim Salim, T.A.Razzack,"A study on IDS for Preventing denial of service attack using outliers techniques", 2nd IEEE international conference on Engineering and technology, March 2016.

[6] S.S Rao, SANS Institute Infosec Reading Room.,"Denial of service Attack and mitigation techniques: Real time implementation with detailed analysis", 2011.

[7] Data Mining:Concepts and Techniques, Third Edition, Jiawei Han and Micheline Kamber, ISBN-13, 9780123814791.

[8] Mining of Massive Data Sets, Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman,2014

[9] A. Klein, F. Ishikawa, and S. Honiden. Efficient heuristic approach with improved time complexity for qos-aware service composition. In ICWS, pages 436–443. IEEE, 2011.

[10] Tripathy, M.Khan, M.R.Patra, H.Fatima, P.Swain, " Dynamic web service composition with QoS clustering" IEEE , International Conference on Web services, 2014.