Anomaly Extraction and Mitigation using Efficient-Web Miner Algorithm

Gargi Joshi
Department of computer Engineering
Dr. D. Y. Patil College of Engineering
University of Pune, Ambi, Pune – 410506

A. K. Bongale
Department of Computer Engineering
Dr. D. Y. Patil College of Engineering
University of Pune, Ambi, Pune - 410506

ABSTRACT

Today network security, uptime and performance of network are important and serious issues in computer network. Anomaly is deviation from normal behavior affecting network security. Anomaly Extraction is identification of unusual flow from network, which is need of network operator. Anomaly extraction aims to automatically find the inconsistencies in large set of data observed during an anomalous time interval. Extracted anomalies will be important for root cause analysis, network forensics, attack mitigation and anomaly modeling. Frequent pattern mining technique namely Efficient-Web Miner Algorithm will be used to generate the set of association rules applied on metadata. Using network traffic log data, algorithms effectively finds the flow associated with the anomalous event(s). Efficient-Web Miner Algorithm triggers a very small number of false positives. Efficient- Web Miner has much better performance in terms of time and space complexity than Apriori Algorithm and its variations like Apriori All algorithm.for large data sets This anomaly extraction method significantly reduces the time needed for analyzing alarms, making anomaly detection systems more practical, simple and realistic. System makes an effort to mitigate the anomaly so detected without human intervention. Proposed system provides human overrides in mitigation process and inculcates self-learning approach which is advantageous.

General Terms

Data Mining, Network Security Algorithms

Keywords

Anomaly Extraction, Association rule mining, data mining, detection algorithms, Efficient-Web Miner Algorithm

1. INTRODUCTION

Anomaly detection techniques are the last line of defense when other approaches fail to detect security threats. Anomaly detection techniques have been extensively studied since they pose a number of interesting research problems, involving statistics, modeling, and efficient data structures. Nevertheless, they have not yet gained widespread adaptation, as number of challenges, like reducing the number of false positives or simplifying training and calibration, remain to be solved.

An anomaly detection system provides meta-data relevant to narrow down the set of candidate anomalous flows. For example, histogram bins generated using Histogram based detection technique [4] [5] [6] [7], indicates affected range of IP addresses or port numbers. Such meta-data can be used to restrict the candidate anomalous flows to affected network node. To extract anomalous flows, one could build a model describing normal flow characteristics and use the model for identifies deviating flows. However, building such a microscopic model is very challenging due to the wide variability of network flow characteristics. Similarly, one could compare flows during an interval with flows from normal or past intervals and search for deviations, like new flows that were not previously observed or flows with significant increase/decrease in their volume[8][9]. Such kind of approaches essentially performs anomaly detection at the level of individual flows and could be used to identify anomalous flows.

Proposed system aims to identify an anomaly from the network traffic. We aim to find the flows associated with the event(s) that triggered an observed anomaly. Beginning with network traffic data logs observation for time interval t We then filter this data to eliminate large fraction of normal flows. A summary report of frequent item-sets from the set of suspicious flows is generated by applying association rule mining techniques. System uses Efficient-Web Miner algorithm for anomaly detection. Comparative study shows that the Efficient-Web Miner algorithm works better than standard association rule mining algorithms i.e. Apriori in terms of space and time by reducing the number of database scans and candidate set pruning is reduced in stages.

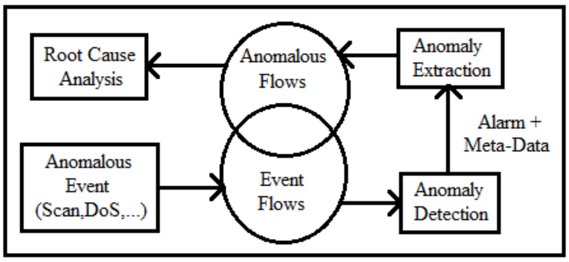

Figure 1: Goal of Anomaly Extraction

2. LITERATURE SURVEY

F, Silveira and Diot [3] introduced a tool called URCA that searches for anomalous flows by iteratively eliminating subsets of normal flows. URCA also classifies the type of a detected anomaly. Nevertheless, it requires to repeatedly evaluating an anomaly detector on different flow subsets, which can be costly. In comparison with this work, simply computing frequent item-sets on pre filtered flows is sufficient to identify anomalous flows. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

Dewaele et al.[11] use sketches to create multiple random projections of a traffic trace, then model the marginal's of the sub traces using Gamma laws and identify deviations in the parameters of the models as anomalies. In addition, their method finds possible anomalous source or destination IP addresses by taking the intersection of the addresses hashing into anomalous sub traces. Compared to this work, paper introduce and validates the techniques to address the more challenging problem of finding anomalous flows rather than IP addresses.

Lakhinaet al.[12] Use SNMP data to detect network-wide volume anomalies and to pinpoint the origin-destination (OD) flow along which an anomaly existed. In contrast, proposed approach takes as input a large number of flow records, e.g., standard 5-tuple flows, and extracts anomalous flows. An OD flow may include millions of both normal and anomalous 5- tuple flows and, therefore, can form the input to proposed methodology.

Li et al. [5], use sketches to randomly aggregate flows as an alternative to OD aggregation. The authors show that random aggregation can detect more anomalies than OD aggregation in the PCA subspace anomaly detection method. In addition, the authors discuss how their method can be used for anomaly extraction. However, the work and evaluation focus primarily on anomaly detection.

Lee and Stolfo [13] show how association rules can be used to extract interesting intrusion patterns from system calls and tcp dump logs.

Vaarandi [14] introduces a tool called LogHound that provides an optimized implementation of Apriori and demonstrates how LogHound can be used to summarize traffic flow records. Yoshida et al.[15] also use frequent item-set mining to identify interesting events in traces from the MAWI traffic archive.

Li and Deng [16] outline a variant of the Eclat frequent item-set mining algorithm that operates in a sliding window fashion and evaluate it using traffic flow traces from a Chinese university.

Chandola and Kumar [17] describe heuristics for finding a minimal set of frequent item-sets that summarizes a large set of flows.

Mahoney and Chan [18] use association rule mining to find rare events that are suspected to represent anomalies in packet payload data. They evaluate their method on the 1999 DARPA/Lincoln Laboratory traces. Their approach targets edge networks where mining rare events is possible. In massive backbone data, however, this approach is less promising. Another application of rule mining i edge networks is eXpose, which learns fine-grained communication rules by exploiting the temporal correlation between flows within very short time windows. Compared to these studies, association rule mining can be combined with anomaly detection to effectively extract anomalous flows. Hierarchical heavy-hitter detection methods [19] [7] group traffic into hierarchical clusters of high resource consumption and focus primarily on optimizing computational performance for summarizing normal traffic. For example, they have been used to identify clusters of Web servers in hosting farms. Hierarchical heavy- hitter detection is similar to frequent item-set mining in that both approaches find different forms of multidimensional heavy hitters. Compared to these studies, intelligently combining multidimensional heavy-hitters with anomaly detection enables us to extract anomalous flows. In addition, frequent item-set mining scales to higher dimensions much better than existing hierarchical heavy-hitter detection methods. Finally, substantial work has focused on dimensionality reduction for anomaly detection in backbone network. These papers investigate techniques and appropriate metrics for detecting traffic anomalies, but do not focus on the anomaly extraction problem which are addressed in this project.

For web log analysis MahendraPratapYadav [2] presents an efficient web mining algorithm for web log analysis and applied the results obtained on this web log analysis to a class of problems for finding out the contexts of website design of a E- commerce web portal which demands security. In this paper the authors compared the algorithm with Improved Apriori All Algorithm which is its other earlier incarnation . The proposed algorithm, Efficient Web Miner or E-Web Miner can be verified by computational comparative performance analysis and can be traced for its valid results and. This paper intends to show that the E-Web Miner has lower complexity of time and space than Improved Apriori All Algorithm and confirms the correctness of result obtained by providing a trace back route for candidate set pruning for both the algorithms. E-Web Miner is the proposed web mining algorithm that removes the flaws of Improved Apriori All algorithm and improve upon the time complexity of the earlier Apriori All algorithm. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines correct result of candidate set where as the Improved Apriori All algorithm fails to deliver the correct result. The algorithm has been designed independent of Apriori All algorithm. In this paper authors work shows that when Improved Apriori All Algorithm fail to deliver the desired result, the proposed algorithm of web mining in web log analysis presents a cost effective valid result having reduced candidate set pruning of correct order. So in this way the authors proved that E-Web miner is proved to be more time effective than other algorithm which provides a strong base to select efficient web miner algorithm over Apriori algorithm and its veriations like Apriori all algorithm.

3. METHODOLOGY

3.1 Assumption and Dependencies

Here, idea is to form a network between n number of computers or laptops. Our system will depend on the multiple machines connected with each other in the local area network. We are assuming server as a router, which observes and keep logs of all the traffic in the network. We will form network traffic for certain interval of time only. We require minimum 4 machines for better results but can be scaled further for additional ones based on requirement.

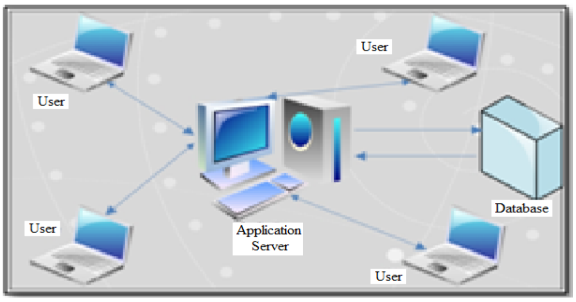

Figure 2: System Architecture

3.2 Process Summary

1) Form network between computers or laptops LAN connection may be wired/wireless.
2) One of the systems will be the server and router which will monitor the network traffic, monitoring program logs the network traffic information in a flat file/ database and the others are designated as clients for network communication.
3) Data is exchanged between the clients and server in the form of network packets. Server being the central point of control blocks the communication from suspicious node. Misbehavior of a node is decided on basis of history maintained by earlier observation windows.
4) Following tuples and parameters are considered as an input to the system along with network traffic log data

   {Source IP, Destination IP, Source Port Number, Destination Port Number, Number of Packets sent}.
5) Observe network for certain interval of time t.
6) Apply Efficient-Web miner algorithm to the set of suspicious flows with somepredefined threshold value set by the administrator satisfying the minimum support indicating normal behavior of the node.
7) Find frequent item sets from the set of suspicious flows which will be the IP address of the misbehaving node.
8) As mitigation, preserve the IP address of misbehaving node. This information will be used to block this IP from network communication from next observation window.

3.3 Mathematical Model

1) U is main set of users (ATM Holders)
   like u1,u2, u3…. So U = {u1, u2, u3…….}
2) A is main set of Administrators
   like a1, a2,a3…. So A = {a1, a2, a3…….}
3) C is the main set of histogram clones
   like c1, c2, c3.... So C = {c1, c2, c3......}
4) Identify the processes as P.
   P = {Set of processes} i.e. P = {P1, P2, P3……}
5) If (anomaly is detected in the network)
   Then
   P1 = {e1, e2, e3, e4}, Where
   {e1=i|i is to build c number of clones}
   {e2=j|j is to find anomalous bins from histogram}
   {e3=k|k is to filter suspicious data}
   {e4=l|l is to find frequent item sets from given suspicious data}
   {e4=m|m is to mitigate the observed anomaly by blocking the anomalous node IP address from accessing the network}
6) Else
   P1 = {e1, e2}, Where
   {e1=i|i is to observe traffic during time interval t}
   {e2=j|j is to check whether anomaly detects or not}

.

3.4 Proposed Efficient-Web Miner Algorithm

E-Web Miner is the proposed web mining algorithm that removes the flaws of Improved Apriori All algorithm and improve upon the time complexity of the earlier Aprioriall algorithm. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines correct result of candidate set whereas the Improved Apriori All algorithm fails to deliver the correct result. The algorithm has been designed independent of Apriori All algorithm

E Web Miner algorithm is built to work upon a single input parameter at one pass as proposed by MahendraPratapYadav[2] but over here paper proposes the same algorithm with a different view by making it able to work for multiple input parameters by considering a single parameter at one go this can be done by adding the self-learn ability to the algorithm as it can learn from past knowledge from the database which is exactly contrast with the classic Apriori Algorithm which can work on bunch of parameters at an instance and do not have learnability this factor makes the E Web Miner algorithm to outstand the Apriori Algorithm by providing more accuracy, scalability, flexibility, learnability and better performance in terms of complexity than the Apriori Algorithm.

Input-

1) Log data file having n number of records each following below format {Source IP, Number of Packets}
2) Log data base repository
3) Support/ threshold value indicating normal behaviour of node set by the administrator.

Output-

1) IP address of the misbehaving node is extracted as the anomalous node in the network.

Algorithm-

```
1) Arrange the packet data set of users in increasing
   order.
2) Store all web packet data set of user in string array A.
3) Frequency =0, MAX=0;
4) FOR i=1 to n
     FOR j=0 to (n-1)
       IF substring (A[i], A[j])
         Frequency=frequency+1;
       END IF
       B[i] =Frequency;
     END FOR
     IF Max <= Frequency
       Max=Frequency;
     END IF
   END FOR
5) Find all position in Array B where value is equal to
   Max and select the corresponding substring from A.
6) Produce output of all substrings with their position
   which is the desired output.
```

3.5 Algorithm Execution Steps

1) Run server monitoring window to record network traffic.
2) From this monitored data prepare input data set i.e. - {Source IP, Number of Packets}
3) Run E-Web Miner Algorithm on the input data set of Step 2 and the log data base repository.
4) Outcome of above step is partial output i.e. – Potentially misbehaving node.
5) Use support/confidence threshold values on outcome of step 4 to get Malicious misbehaving node IP address

3.6 Mitigation Process

Anomaly is the deviation from normal behavior. Here anomalous node is the node flooding the network and hence causing the attack. Anomaly is detected by network observation in fix time window. Efficient Web Miner algorithm is used for this purpose. Efforts are required to act on this identified anomaly. As mitigation the anomalous node is restricted from sending network data packets in next observation window onwards.

In proposed experimental setup network router act as a server. It has a complete control over the communication within the network. As per the basic working router principal of router client sends out the data packets to router/server and then router/server forward the same to intended node. After completing the observation window, server applies efficient web miner algorithm to identify the malicious node. This anomaly is then preserved on permanent storage like databasefor future use. Different possible approaches for blocking the anomalous node are detailed below.

Approach 1: Client is allowed to sends data packet to server. Server then checks whether the packet sender is anomalous as per the historical data available. If sender is already identified as anomalous thendata packets will be dropped and further network congestion will be prevented.

Approach 2: For communication, client needs to choose another network node as destination. Interested sender node communicates with the server to get knowledge about other nodes connected in network for communication. Once server receives such request, it checks if the requesting node is anomalous or not. If sender is already identified as anomalous then it will not be allowed to send data packets to any other network node including router/server. Hence the objective of blocking the node from communication is achieved.

Approach 2 is clearly more effective than approach 1. As per the first approach, it is still possible for the attacker to flood the router/server. This is not the case for later approach.

As the anomalous node is blocked from flooding the network again, mitigation is achieved. Analogous node if detected at the end of observation window is preserved on the router/server and so network administrator has a full control on it. If required administrator can opt to omit a detected node and hence unblocking it. This provision is advantageous to have human override on the system implementation.

4. RESULTS

Data Set: Given the number of item-sets, find frequent subsets which are common to at least a minimum numbers of item-sets. Our item-set consists of 5-tuples, namely {Source IP address, Destination IP address, Source Port, Destination Port, #Packets}.

Input:

1) { traffic log data}
2) Log database repository
3) Support/ threshold value indicating normal behavior of node.
4) List of anomalous nodes identified as per previous observation windows.

Result Set: A summary report of frequent item-sets in the set of suspicious flows is generated by association rule mining.

Output:

1) {Frequent Item Sets}
2) IP address of the misbehaving node is extracted as the anomalous node in the network
   Success: {if anomaly is detected}
   Failure: {if anomaly not detected}

4.1 Comparison with Existing System

In order to justify the selection of efficient web miner algorithm and to gauge its efficiency it is subjected to a comparison against a well proven association rule mining algorithm, Apriori All algorithm is used for this purpose. Both the algorithms are run over a number of experimental transaction sets. Result records the execution time among the proposed E Web Miner algorithm and the Apriori All algorithm.

Table 1: Time to detect anomaly Vs Number of Packets Sent in the network simulation results

Sr No	Number of Packets	Apriori Time	E-Web Miner Time	Performance Improvement %
1	5000	1.32	0.785	40.53
2	7000	1.851	1.147	38.03
3	8000	2.104	1.845	12.31
4	10000	3.214	2.951	8.18
5	11000	3.938	3.153	19.93
6	12000	4.17	3.759	9.86
7	13000	4.385	3.986	9.10
8	16000	6.129	4.391	28.36
9	17000	7.161	5.971	16.62
10	19000	8.124	6.971	14.19

The above results clearly show that proposed E Web Miner Algorithm performs better than the Apriori All Algorithm which is indicated by the performance graph-.

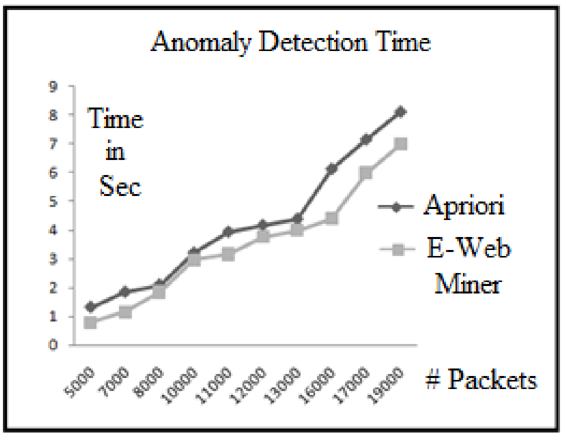

Figure 3 Anomaly detection time comparison graph

The main objective of this study is to reduce the number of elements in every candidate set without any repetition. Below are the findings of this comparison -

1) The algorithm provides more accurate results with better performance with respect to time complexity. This proves it that proposed algorithm is better against the classic Apriori All Algorithm as candidate set

pruning is reduced in steps and number of database scans are reduced which proves E Web Miner to be more effective.

2) The proposed algorithm is self learnable. It uses a log repository built over a period. This repository is used to match the results of real time mined logs to identify anomaly. Also real time network logs are preserved in this repository. Hence algorithm keeps on learning at every instance.
3) For all the above transaction sets proposed Efficient Web Miner Algorithm shows a noticeable average performance improvement of 19.71%.

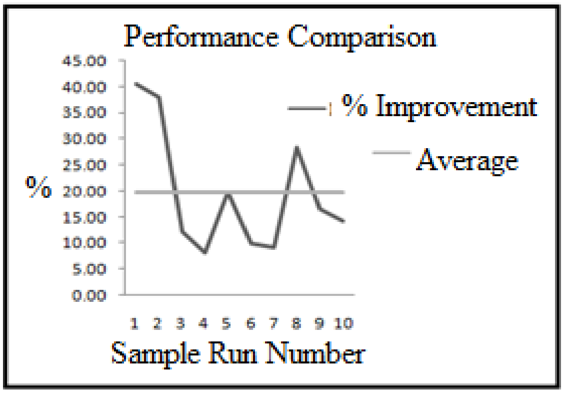

Figure 4 Performance graph of proposed E Web Miner Algorithm

4) Due to better efficiency of E-Web miner algorithm, it is proved to be scalable in comparison with Apriori All algorithm.

5. CONCLUSION

The proposed methodology is very useful for finding suspicious network data flows. Anomalies so detected help in anomaly mitigation, network forensics and anomaly modeling.

Histogram detection technique is used to provide metadata for filtering anomalous network data flows. Apriori All algorithm was used for frequent item-set mining. As an alternative to improve on Apriori algorithm Efficient-Web Mining Algorithm is implemented. E Web Miner Algorithm takes into consideration support and confidence of any sequential pattern It proves that Efficient-Web Mining algorithm is simple and provides better space and time complexity than Apriori is scalable flexible provide accurate results These advantages of Efficient-Web Mining Algorithm are mainly because of reduction in data set scans and improved candidate set pruning thus removing loop holes of Apriori Algorithm. E Web miner can deal with single input parameter at one time this scope of the algorithm is further extended to work on multiple input parameters, self-learnability is introduced as it can learn from past knowledgeall these additional features enhance the beauty and scope of the algorithm. Among all these the latest technology used is E-Web miner. But in the E-web miner[2], although it has reduced the problem of candidate set generation by providing an improved candidate set pruning but still it cannot remove this problem completely which has again further scope of research and improvement Possible future extension exists in optimizing the scalability and efficiency of frequent item set mining for dealing with huge data, mining on top k item sets, mining on multilevel and multidimensional or quantitative features of for network traffic monitoring.

6. FUTURE ENHANCEMENTS

E-Web miner. But in the E-web miner [2], although it has reduced the problem of candidate set generation by providing an improved candidate set pruning but still it cannot remove this problem completely which has again further scope of research and improvement Possible future extension exists in optimizing the scalability and efficiency of frequent item set mining for dealing with huge data, mining on top k item sets, mining on multilevel and multidimensional or quantitative features of for network traffic monitoring. The algorithm is efficient but may be further improved using suitable data compression techniques

7. REFERENCES

[1] D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K Salamatian, "Anomaly extraction in backbone networks using association rules,"inProc.IEEE ACM TRANSACTION ON NETWORKING, VOL.20. NO 6, DECEMBER 2012.

[2] M. Yadav,P. Keserwani, S. Samaddar "An efficient web mining algorithm for web log analysis: E Web Miner" RAIT 2012.

[3] F. Silveira and C. Diot, "URCA: Pulling out anomalies by their root causes," in Proc. IEEE INFOCOM, Mar. 2010, pp. 1-9.

[4] A. Kind, M. P. Stoecklin, and X. Dimitropoulos, "Histogram-based traffic anomaly detection," IEEE Trans. Netw. Service Manage., vol. 6, no. 2, pp. 110-121, Jun. 2009.

[5] M. P. Stoecklin, J.-Y. L. Boudec, and A. Kind, "A two-layered anomaly detection technique based on multi-modal flow behavior models," in Proc. 9th PAM, 2008, Lecture Notes in Computer Science, pp. 212-221.

[6] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone,and A. Lakhina, "Detection and identification of network anomaliesusing sketch subspaces," in Proc. 6th ACM SIGCOMM IMC, 2006,pp. 147 -152.

[7] K. H. Ramah, K. Salamatian, and F. Kamoun, "Scan surveillance in Internet networks," in Proc. Netw., 2009, pp. 614-625

[8] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: Methods, evaluation, and applications," in Proc. 3rdACM SIGCOMM IMC, 2003, pp. 234-247.

[9] G. Cormode and S. Muthukrishnan, "What's new: Finding significant differences in network data streams," IEEE/ACM Trans. Netw., vol. 13, no. 6, pp. 1219-1232, Dec. 2005.

[10] Y. Gu, A. McCallum, and D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," in Proc. 5th ACM SIGCOMM IMC, 2005, pp. 32-32.

[11] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, andK. Cho,"Extractinghidden anomalies using sketch and non Gaussian multi resolution statistical detection procedures," in Proc. LSAD, 2007, pp. 145-152.

[12] A. Lakhina, M. Crovella, and C. Diot,“Diagnosing network-wide traffic anomalies,” in Proc. ACM SIGCOMM, 2004, pp. 219-230.

[13] W. Lee and S. J. Stolfo, “Data mining approaches for intrusion detection,” in Proc. 7th USENIX Security Symp., 1998, vol. 7, p. 6.

[14] R. Vaarandi, “Mining event logs with SLCT and LogHound,” in Proc.IEEE NOMS, Apr. 2008, pp. 1071-1074.

[15] K. Yoshida, Y. Shomura, and Y. Watanabe “Visualizing networkstatus,” in Proc. Int. Conf. Mach. Learning Cybern., Aug. 2007, vol.4, pp. 2094-2099.

[16] X. Li and Z.-H. Deng, “Mining frequent patterns from network flowsfor monitoring network,” Expert Syst. Appl. vol. 37, no. 12, pp.8850-8860, 2010.

[17] V. Chandola and V. Kumar, “Summarization—Compressing data intoan informative representation,” Knowl. Inf. Syst., vol. 12, pp. 355-378,2007.

[18] M. V.Mahoney and P. K. Chan, “Learning rules for anomaly detection of hostile network traffic,” in Proc. 3rd IEEE ICDM, 2003, pp.601-6

[19] G. Cormode and S. Muthukrishnan, “An improved data stream sum- mary: The count-min sketch and its applications,” J. Algor., vol. 55, no. 1, pp. 58-75, 2005.

[20] Tong, Wang and Pi-lian, He, Web Log Mining by an ImprovedAprioriAll Algorithm World Academy of Science,EngineeringandTechnology, Vol 4 2005 pp 97-100.