Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Anomaly or Misbehaviour Node Extraction Using Efficient-Web Miner Algorithm

Ms. Gargi Joshi*, Prof. Anupkumar Bongale

Dept of Computer Engg,
Dr.D.Y Patil College of Engineering,
Ambi, Pune, University of Pune, India

***Abstract**— Today network security, uptime and performance of network are important and serious issue in computer network. Anomaly is deviation from normal behaviour 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 modelling. We use meta data provided by several histogram based detectors to identify suspicious flows, and then apply association rule mining to find and summarize anomalous flows. Frequent pattern mining techniques namely Apriori Algorithm and 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 for large data sets. This anomaly extraction method significantly reduces the time needed for analyzing alarms, making anomaly detection systems more practical, simple.*

***Keywords** Anomaly Extraction, Apriori Algorithm, Association rule mining, data mining, detection algorithms, Efficient-Web Miner Algorithm, histogram cloning.*

I. 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, modelling, 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 proposed solution then applies histogram detection techniques for anomaly detection. Upon detection of anomaly, we build the clones of histogram detector and find suspicious flows. 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 Apriori and Efficient-Web Miner algorithm for anomaly detection. Comparative study shows that Efficient-Web Miner algorithm works better than standard association rule mining algorithms i.e. Apriori in terms of space and time.

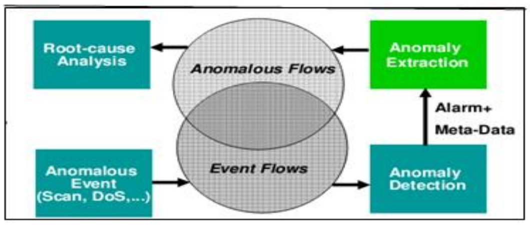

Fig. 1 Goal of Anomaly Extraction

II. 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. Compared to this work, we 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, we introduce and validate techniques to address the more challenging problem of finding anomalous flows rather than IP addresses.

Lakhina et 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, our 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 our 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 we are addressing in this project.

2.1 Contribution Work:

We are implementing Efficient-Web Miner algorithm [2] to find out frequent item sets. The main strength of Efficient-Web Miner is its simplicity. Efficient-Web Miner is the proposed web mining algorithm that removes the flaws of Apriori algorithm and improves upon the time complexity. It provides an improved candidate set pruning as well. In fact, it has been successfully proven that it mines correct result of candidate set whereas the Improved Apriori algorithm fails to deliver the correct result. The algorithm has been designed independent of Apriori algorithm. Comparative study is done to indicate that Efficient-Web Miner algorithm is efficient than Apriori algorithm. Our project comes into the NP complete category. Solution of all decision problems can be obtained in polynomial time using proposed algorithm.

III Method and Procedure

Proposed system contains three different phases –

- Histogram Detector
- Histogram Cloning and Voting
- Association Rule Mining

Histogram detector will observe the network traffic and alert the system upon anomaly detection. Histogram cloning assures the anomaly detection and finds the suspicious flows from network traffic. Finally association rule mining algorithm i.e. Efficient-Web Miner Algorithm finds the frequent item sets.

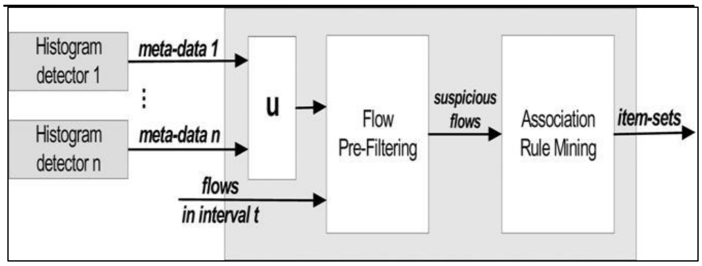

Fig. 2 Process of Anomaly Extraction

- *Flow Pre-filtering*

Filtering usually eliminate large fraction of normal flows from network data captured during time interval t. Filtering thus results into reduction of workable dataset and improves association rule mining algorithm by removing flows that result in false-positive item-sets.

- *Frequent Item Set Mining*

Here, we apply the first step of association rule mining, i.e., we find frequent item-sets to extract suspicious flows from a large set of network data flows observed during a time interval t. The standard algorithm for discovering frequent item-sets is the Apriori algorithm. We are using Efficient-Web Mining Algorithm. Efficient-Web Mining is simple and it provides an improved candidate set pruning. It has better time complexity than Apriori algorithm. Due to these advantages it overcomes the drawbacks of Apriori Algorithm.

- *Histogram based Detection*

We build a histogram-based detector for our evaluation that uses the Kullback–Leibler (KL) distance to detect anomalies [7] [10]. Each histogram detector monitors a flow feature distribution, like the distribution of source ports or destination IP addresses. We assume histogram-based detectors that correspond to n different traffic features and have each m histogram bins.

- *Histogram cloning and voting*

As an alternative to arbitrary binning, we introduce histogram cloning [3]. With histogram cloning, different clones provide alternative ways to group feature values into a desired number of bins/groups creating effectively additional views along which an anomaly may be visible. The cloning mechanism is coupled with a simple voting scheme that controls the sensitivity of the detector and eventually affects a tradeoff between false positives and negatives.

- *Process Summary*

1. Form network between computers or laptops.
2. Histogram detector will observe network for certain interval.
3. On anomaly detection form clones of histogram and find suspicious flows in network.
4. Apply Efficient-Web miner algorithm to these suspicious flows.
5. Find frequent item sets from the set of suspicious flows

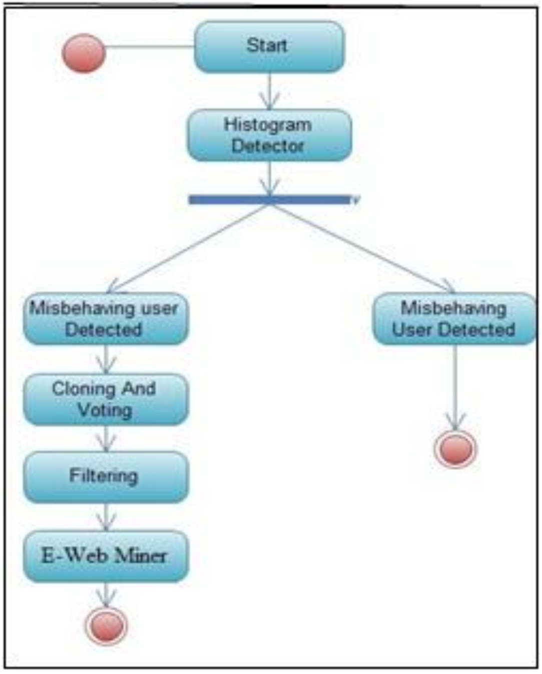

Fig. 3 System State Diagram

A. 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….
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}
P = {P1, P2, P3……}

```
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}
Else
    P1 = {e1, e2}, Where
{e1=i|i is to observe network traffic during time interval t}
{e2=j|j is to check whether anomaly detects or not}
```

B *Data Independence and Data Flow Architecture*

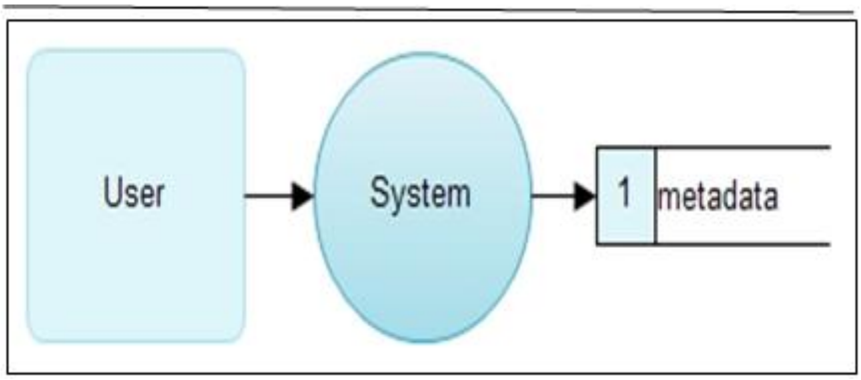

Fig. 4 State 0 Data Flow Architecture

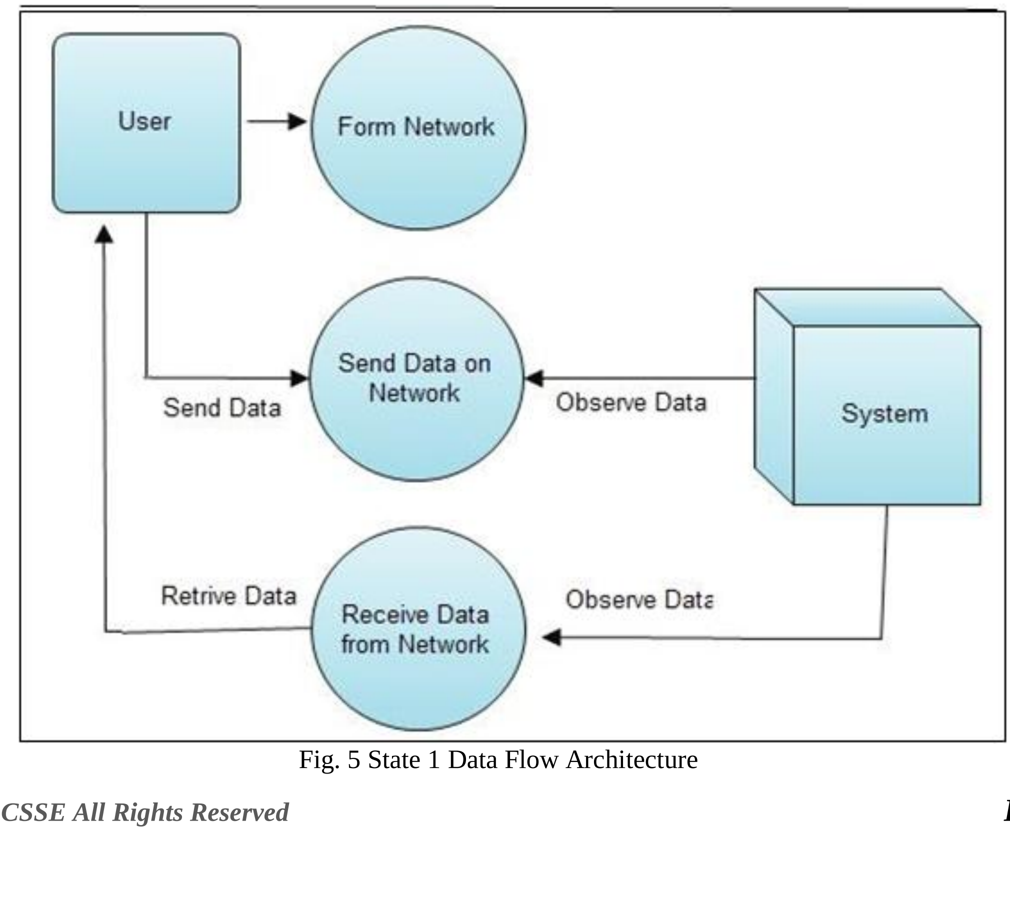

Fig. 5 State 1 Data Flow Architecture

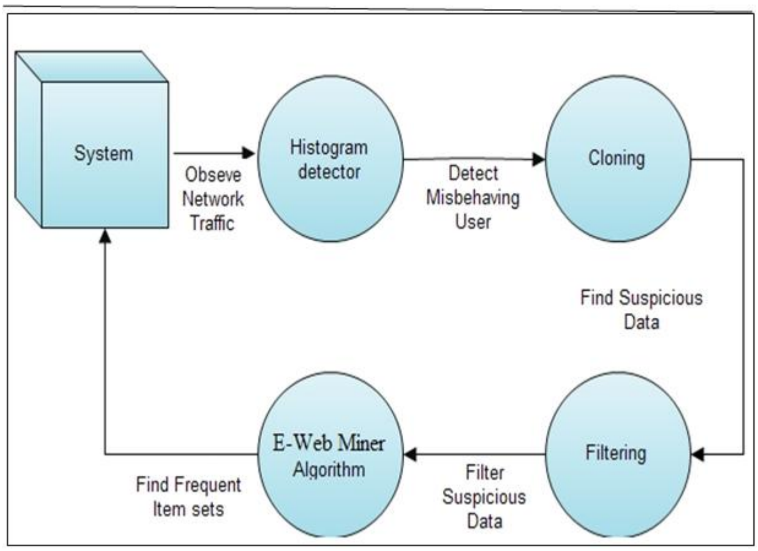

Fig. 6 State 2 Data Flow Architecture

C Proposed Efficient-Web Miner 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.
```

B. Project Setup

1) Software Specification:

Operating System	: Windows 7.
Development End	: JAVA [JDK 1.6]
IDE	: Eclipse Helios
Tool	: JCreator

2) Hardware Specification:

Processor	: PIV– 500 MHz to 3.0 GHz
RAM	: 1GB
Disk	: 20 GB
Monitor	: Any Color Display
Key Board	: Standard Windows Keyboard

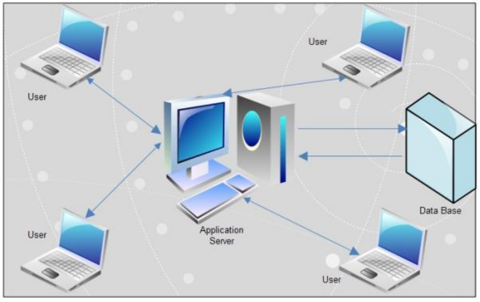

Fig. 7 System Architecture

C. Module Information

1) Module 1: Design of the Graphical User Interface (GUI) for our system with client and server.
2) Module 2: Build histogram detector to observe the traffic on the network and detect anomalies.
3) Module 3: Find suspicious flows from the network traffic that causes anomaly in the network.
4) Module 4: Implement Apriori and Efficient-Web Miner algorithm to find frequent item sets.

D. Assumption and Dependencies

Here, we 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 in our system 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.

III. RESULTS AND DISCUSSION

A. Data Set

Given the number of item-sets we find frequent subsets which are common to at least a minimum numbers of item-sets. Our item-set consists of 7-tuples, namely {Source IP address, Destination IP address, Source Port, Destination Port, Protocol, #Packets, #Bytes}.

Input: {Network traffic data}

B. Result Set

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

Output: {Frequent Item Sets}
Success: {if anomaly is detected}
Failure: {if anomaly not detected}

IV. CONCLUSION

The proposed methodology is very useful for finding suspicious network data flows. Anomalies so detected help in anomaly mitigation, network forensics and anomaly modelling. Histogram detection technique is used to provide metadata for filtering anomalous network data flows. Apriori algorithm was used for frequent item-set mining. As an alternative to Apriori algorithm Efficient-Web Mining Algorithm is implemented. It proves that Efficient-Web Mining algorithm is simple and provides better space and time complexity than Apriori. 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. 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 for network traffic monitoring.

ACKNOWLEDGEMENT

This work was supported by Prof. Anupkumar Bongale, Department of Computer Engineering, Dr. D.Y.Patil College of Engineering Ambi, Pune.

REFERENCES

[1] D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K Salamatian, "Anomaly extraction in backbone networks using association rules,"in Proc.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, and K. 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.