Anomaly Extraction Using Histogram-Based Detector

Mayur Devidas Khairnar[1]
Student of BE Information Technology
BVCOE & RI, Nasik, Maharashtra, India
University of Pune
mayur.khairnar81290@gmail.com

Sandesh Ramesh shinde[2]
Student of BE Information Technology
BVCOE & RI, Nasik, Maharashtra, India
University of Pune
ciit.sandeshshinde@gmail.com

Swapnil Bajirao Suryawanshi[3]
Student of BE Information Technology
BVCOE & RI, Nasik, Maharashtra, India
University of Pune
suryawanshiswapnil603@gmail.com

Girish Sunil Patil[4]
Student of BE Information Technology
BVCOE & RI, Nasik, Maharashtra, India
University of Pune
grshpatil4@gmail.com

Prof. Kavita S. Kumavat[5]
ME Computer Engineering
BVCOE & RI, Nasik, Maharashtra, India
University of Pune
kavitakumavat26@gmail.com

Abstract **-** Now a day's network traffic monitoring and performance of the network are more important aspect in the computer science. Anomaly Extraction is a method of detecting in large set of flow observed during an anomalous time interval, the flows associated with the one or more anomalous event. Anomaly extraction is important problem that essential for application ranging from root cause analysis and attack mitigation and anomaly extraction is also important problem for several application of testing anomaly detector. In this paper, use a meta-data provided by histogram detector for detect and identify the suspicious flow after successfully detection suspicious flow then applying the association rule mining for finding the anomalous flow. By using the rich traffic data from the meta-data of the histogram-based detector we can reduce the classification cost. In this paper, Anomaly extraction method reduce the working time which is required for analyzing alarm, its make system more practically.

Keywords- *Anomaly Extraction, Histogram-based detector, Apriori Algorithm, Association rule .*

*****

I. INTRODUCTION

An anomaly detection technique may provide meta-data. In anomaly detection system meta-data useful to an alarm that minimizes the set of candidate anomalous flows. Anomaly detection system produced histograms and it uses histogram bins to show flow will be anomaly affected. For e.g. range of Source Port numbers, Source IP address. When other methods that failed to detect security thread or other problem that time anomaly extraction is most useful technique to detect the all problems that occurred when they have been studied and introduces number of interesting problems likes statistic, modeling and efficient data structure. However they have not yet obtained the widespread adaptation as the number of challenge, like reducing the number of false positive or simplifying training and calibration, remains to be solved. System observed the network traffic time interval t and identifying the traffic flows associated with an anomaly during time interval with an alarm. In general term detecting the network flow associated with anomaly we call these anomalous flow is anomaly extraction or extraction problem.

Anomaly Extraction: Anomaly extraction is a method for finding the network flow which is affected by the anomaly or simply finding the anomalous flow associated with the traffic flow.

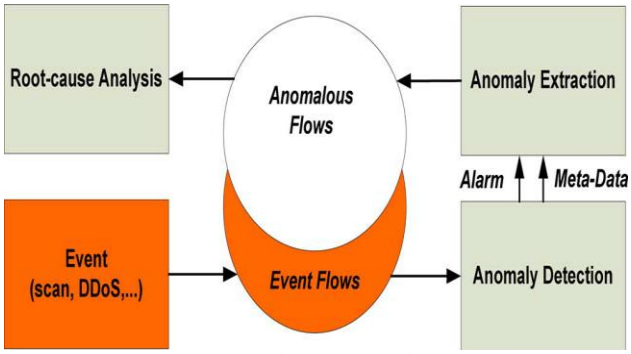


Figure 1. Anomaly extraction for anomalous flow and associated with network event.

Figure 1 shows that the high level of anomaly extraction. In the figure shows the network trigger events flow or attacks, like Denial of services or scan. After the finding the suspicious anomalous flow or event anomaly detector may raise an alarm and provide the meta-data that is useful for

applying the association rule. Then by using meta-data provided by anomaly detection system it will used in anomaly extraction system. In the anomaly detection meta-data provide the candidate anomalous flows. By using the candidate anomalous flow anomaly extraction system builds the histogram. And histogram bins show the flow which are affected by the anomalous flow.

In figure 2 {F1, F2 and F3} that show the candidate suspicious flow system produced the FA i.e. union of the suspicious flow F2 and F3. Anomaly detection system provides the meta-data for the candidate anomalous flow. In the table 1 show the outline useful meta-data that are provided by the different well-known anomaly detection system.

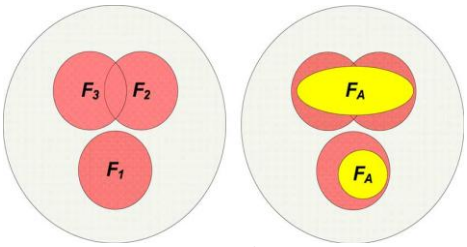


Figure 2 applying association rule and filtering the mete-data using union.

Table 1 meta-data provided by different anomaly detectors

Meta-data	Anomaly detection technique
Protocol	Maximum-Entropy [11] Histogram [14], [30]
IP range	Defeat [20] MR-Gaussian [9] DoWitcher [27] Histogram [14], [30]
Port range	Maximum-Entropy [11] Histogram [14], [30] DoWitcher [27]
TCP flags	Maximum-Entropy [11] Histogram [14], [30]
Flow size	DoWitcher [27]
Packet size	Histogram [14], [30]
Flow duration	Histogram [14], [30]

This system use the multiple histogram-based detector that provide the alternative approch to identify the anomalous flow. The Instrusion Detection System (IDS) used the past data for normal models. By using the histogram-based anomaly detection system does not depends on any past data for the noramal models. histogram-based detector does not used any past data so it will provides an new and addition view of network traffic. In the figure 2, detector produced the set FA and it is represent the candidate flows. After finding the candidate flow system use association rules to extract from the union of F3 and F2 and FA is the summary of the anomalous flow.

Existing System:

In existing system to finding the anomaly is critical for timely mitigation of event like attack, failure & to provide the security and performance for network is critical to timely mitigation of event. The IDS (Intrusion Detection System) was used for identify known attack only it can't identify the new attack that are newly introduced in network. Signature-based detection finds most known attacks pattern, it fails to identify new attacks pattern.

Proposed System:

In system consist the four phase

1) Histogram based detector
2) Histogram cloning & voting
3) Flow profiteering
4) Association rule mining i.e. Apriori algorithm.

In the first phase Histogram based detector is used for observed the network traffic flow And when anomaly is detected it will raised them system alarm .Histogram based detector detecting anomalous behavior and changes in traffic network distributions. In the second phase Histogram cloning & voting find out suspicious flow in network traffic. In the third phase flow prefiltering is used for union the all Meta data in given time interval *'i'*. In the fourth phase association rule mining i.e. Apriori algorithm is applied for finding the frequent item set

.

II. LITERATURE SURVEY

The Unsupervised Root Cause Analysis (URCA) for isolating anomalous traffic and classify alarms with high accuracy [1]. It used to reduce anomalous space and also eliminate the normal traffic which is provided by the anomaly detection method. URCA accurately diagnose large range of anomaly types (e.g. D0DoS attacks) and network scan. In this system simply flow filtering provides the meta-data for the identifying the suspicious flow in the network which is less costly than the URCA. For the multiple random projections of traffic trace use sketches then Gamma laws is used for model the marginal's of the sub traces then identify deviations in the parameters of the models as anomalies [2]. System use IP addresses for detecting the anomalous flows. The scalable system is used for worm detection in backbone networks. An anomaly detector provides different detectors leverage suspicious attribute then system automatically constructs a flow-filter mask [3]. System calls as well as dump logs provides interesting intrusion patterns and they show how association rule can be used [4]. It describes heuristics for finding frequent item-sets that show large set of flow [5]. Summarization is key of data mining which is best suitable for finding a data set. Using the LogHound tool get the optimized implementation of Apriori and demonstrates how it is used to summarize traffic flow record [6]. To represents anomalies in packet payload data association rule mining is used to find the rare events [7]. To identify interesting event in trace they use frequent item-set mining from the MAWI traffic archive [8]. In the transaction of the large database difficulty for discovering [9]. They present new algorithm for solution of above problem. AprioriHybrid is the algorithm which is merging characteristic two or more algorithm and also expresses how the most characteristics are used. System uses association rule mining and an anomaly detector for the detecting the anomalous flow.

III. SYSTEM OVERVIEW

In this section give an overview of the system approach to anomaly extraction. In this section also discuss the details of each functional block are as follows histogram detection and

cloning , voting parameter, prefiltering, and association rule mining and apriori algorithm.

Flow Prefiltering:

Flow pre filtering is process in which we have set of Meta data .In that metadata we union of all Meta data for pre filtering the given time interval *'i'* In flow pre filtering we reduce the large set of Meta data flow

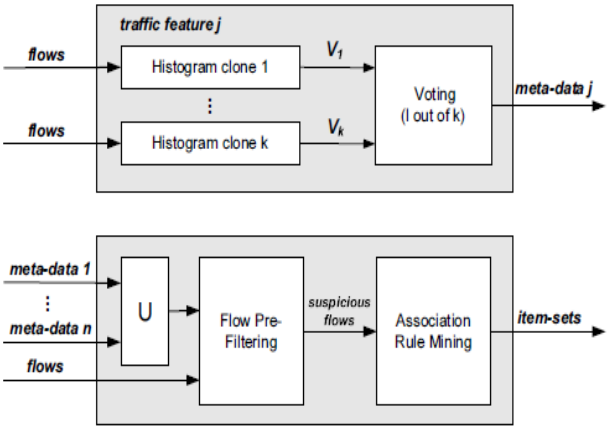


Figure 3: System Overview Diagram

Histogram Based detector:

Histogram based detector is shows the behavior of anomaly. Histogram Based detector also shows the traffic distribution in network flow.

In histogram based detector we are used the kullback- leibler Method for detecting the Anomaly. In Histogram Base detector we are generated the histogram bins. KL distance measured the similarity between the Source and Destination When we are sending the packet in a network we have some Important issue will be consider this issue are like SrcIP, DestIP, SrcPort, DestPort, No of Packet flow .

In figure 4 shows the normal behavior of system. In the first difference system positive and negative intervals are the same phase. We build the histogram detector table using KL distance. During time t intervals each histogram detector has been successfully evaluate using KL distance.

KL distance shows the roughly IP addresses of upper figures. The first difference of KL distance is normally deputed approximately normal deviations.

Histogram Cloning & voting:

Histogram Base detector generated histogram bins. Using that multiple histogram bins we are generated multiple histogram clone at a time interval *'i'*. Each Histogram clone used independent hash function. Each clone compile *vk* of traffic features with histogram bins. Histogram cloning reduces the collision & false positives. System analysis the final impact of different parameter setting of *l* and *k* for finding accuracy of approach.

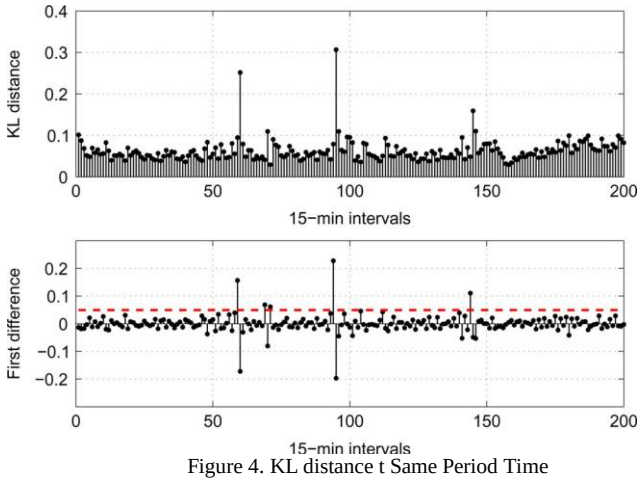


Figure 4. KL distance t Same Period Time

Histogram Cloning & voting:

Histogram Base detector generated histogram bins. Using that multiple histogram bins we are generated multiple histogram clone at a time interval *'i'*. Each Histogram clone used independent hash function. Each clone compile *vk* of traffic features with histogram bins. Histogram cloning reduces the collision & false positives. System analysis the final impact of different parameter setting of *l* and *k* for finding accuracy of approach.

Apriori Algorithm

Apriori is algorithm proposed by R.Agraval and R.Srikant in 1994 for mining frequent item sets for Boolean association rules. This algorithm used prior knowledge of frequent itemset properties. Apriori employs and iterative approach knows as a level wise search, where k-item sets are used to expose (k-1) item set.

First, A necessary condition for finding and association rule of from A->B is sufficiently high support. therefore, for finding such rules system have first to find item sets within the transaction that occur sufficiently frequent. This are called frequent item sets. Second, system can observe that any subset of a frequent item set is necessarily also a frequent item set this is called the apriori property. Third, we can exploit this observation in order to reduce the no of item set that need to system considered in the search. Ones frequent item set of lower cardinality are found, only item sets of longer cardinality need to be considered that contain one of the frequent item sets already found. This allows reducing the search space drastically as will see.

Pseudo-code:

Ck: Candidate item set of size k.
Lk: frequent item set of size k.
L1 = {frequent items};
For (k= 1; *Lk* !=Ø ; k++)

Do begin

Ck+1 = candidates generated for *Lk*,
For each transaction *t* in database do

Increment the count of all candidate in *Ck* +1 that are contained in *t*

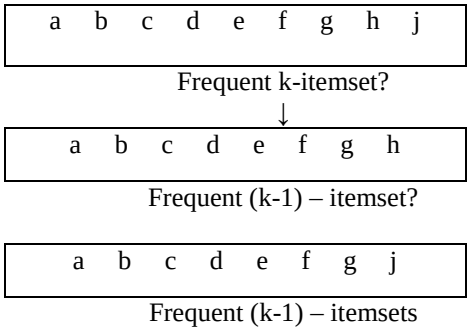
Lk+1= candidates in *Ck*+1 with minimum support

End

Return *Uk Lk*

Exploiting the Apriori Property

If we know the (k-1) item sets, which are candidates for being frequent if we know all frequent (k-1) then we can construct a candidate set CK for frequent k-item set by joining two frequent (k-1)- item set that differ by exactly 1 item join step. Only these item sets CAN BE frequent k-item sets.



Assume that we know frequent itemset of size k-1 considering a k itemsets; we can immediately conclude that by dropping two different items, we have two frequent (k-1) itemset. We take two (k-1) item sets which differ only by one item and take their union this step is called join step and used to construct potential frequent k-itemset.

Parameter Estimation

The various parameters and there ranges are used for evaluation of this system are listed in following table 2:

Table 2. Parameter estimation.

Parameter	Description	Range
N	Number of detectors	5
W	Interval length	[5,10,15]min
M	Hash function length	[5,12,1024,2048]
K	Number of clones	1-50
L	Voting parameter	*1-k*
S	Minimum support	1%-10%

The detail description of each parameter are describe in following and discus selection criteria for each parameter

Number of detectors (n)

In this system we use five detectors for increase the further information and identifying the anomaly's each detector monitors the one of the following feature

1. Source IP address
2. Destination IP address
3. Source port number
4. Destination port number
5. Number of packet per flow

And the other features that might be very useful for anomaly detection are the number of packet per flow, the average packet size, and the flow duration.

Interval length (w)

It determines the detectable anomaly scale; i.e. it is harder to detect the short disruption that contains only few flows with longer intervals. But it is not always desirable to detect such short disruption. Hence, the desired number of daily or weekly anomalous alarm can used to set the interval length w.

Hash function length (m)

The hash function length are also used in detection sensitivity versus aggregation trade-off as discussed for parameter w. the smaller m the more flows are aggregated per hash function bin. Smaller bins are preferable for anomaly extraction.

Voting parameter (l & k)

The parameter k determines the total no of histogram clones that is used. The parameter k has an impact on the portability that a feature value remains in the meta-data after voting, and thus on accuracy. The parameter l determines the lower bound for the number of clones that need to select feature value that included in the final metadata. The parameter setting for l and k can obtained using equation 1 and 3 by simulation. Finally the parameter l and k serve to balance the number of false and true positives produced by pre-filtering.

Minimum support (s)

To make the problem tractable, we introduced the concept of minimum support. The user has to specify this parameter – let us call it minsupport. The parameter s determines the frequency threshold above which an item set is extracted by apriori as a possible set of anomalous flows. Our problem now becomes- find all rules that have a given minimum confidence and involves item set whose support is more than minsupport.clearly, ones we know the supports of all the itemset, we can easily determine the rules and their confidence. Hence we need to concentrate on the problem of finding all item sets which have minimum support. It is call such item sets as frequent item sets. The minimum support can be used as a user variable for zooming in and out of the most significant item-sets. The administrator can progressively decreases s until sufficient anomalous item-set have been investigated.

IV. MATHEMATICAL MODAL

1] U is main set of users (in network e.g. 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 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

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.

V. CONCLUSION

System takes as input a large set of flows and aim at finding the flow associated with the one or more than one event. It is essentials for detecting the core origin of detected anomalies that provide details of mitigation with network forensic. System first introduces a histogram-based detector that provides the fine-grained meta-data for filtering suspect flows. Furthermore, system introduced a method for extractions and applies frequent item-set mining to find large sets of flows with identical values in one or more features. System present anomaly extraction approach is generic and can be used with different anomaly detector that provides meta-data about identified anomalies. System reduced the work-hours needed for the manual verification of anomaly alarms.

REFERENCES

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

[2] G. Dewaele, K. Fukunda, P. Borgant, P Abry, and K. Cho, "*Extracting hidden anomalies using sketch and non Gaussian multiresolution stastical detection procedures*" in Proc LSAD, 2007, pp. 145-152.

[3] S. Ranjan, S. Shah, A. Nucci, M. M. Munafo, R. L. Cruz, and S. M. Muthukrishnan, "*Dowitcher: Effective worm detection and containment in the Internet core*," in Proc. IEEE INFOCOM, 2007, pp.2541-2545.

[4] W. Lee and S. J. Stolfo, "*Data mining approaches for instrusion detection,*" in Proc. 7th USENIX Security Symp., 1998, vol. 7 p.6.

[5] V. Chandola and V. Kumar, " *Summariazation-Compressing data ino an informative representation,* " Knowl. Inf. Syst., vol. 12, pp. 335-378, 2007.

[6] R. Vaarandi, "*Mining event logs eith SLCT and LogHound,*" in Proc.IEEE NOMS, Apr.2008, pp.1071-1074.

[7] M. V.Mahoney and P. K. Chan, "*Learning rules for anomaly detection of hostile network traffic,*" In Proc. 3rd IEEE ICDM, 2003, pp. 601-604

[8] K. Yoshida, Y. Shomura, and Y. Watanabe "*Visualizing network status,*" in Proc. Int. conf. Mach. Learning Cybern., Aug. 2007, vol.4 , pp. 2094-2099.

[9] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules in large database,"in Proc. 20 th VLDB, Santiago de Chile, Chile, Sep. 12-15, 1994,pp. 487-499.

Mayur Devidas Khairnar he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of development.

Sandesh Ramesh Shinde he is student of Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of security.

Girish Sunil Patil he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of pune. His interest in the field of security.

Swapnil Bajirao Suryawanshi he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Instituted, Nashik Under University of Pune. His interest in the field of database administrator.

K. S. Kumavat, ME, BE Computer Engg. Was educated at Pune University. Presently she is working as Head Information Technology Department of Brahma Valley College of Engineering and Research Institute, Nasik, Maharashtra, India. She has presented papers at National and International conferences and also published papers in National and International Journals on various aspects of Computer Engineering and Networks. Her areas of interest include Computer Networks Security and Advance Database.