Available Online at www.ijcsmc.com/journal

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 6, June 2022, pg.87 – 104

Design and Analysis of System to Detect Anomaly from Network Traffic to Improve the Security and Improve Performance

[1]Niranjan Singh Patel; [2]Dr. Tryambak Hiwarkar

[1]Ph.D Research Scholar; [2] Professor CSE

[1,2]School of Engineering and Technology

[1,2]Sardar Patel University Balaghat

DOI: https://doi.org/10.47760/ijcsmc.2022.v11i06.007

Abstract: As network infrastructure maintenance becomes more complex, network traffic monitoring becomes an increasingly vital part of the process. Many specialized solutions for online network traffic monitoring are available to fight against common (and well-known) assaults by quickly limiting portions of the traffic. It is possible, however, that there are unknown hazards in network traffic with slow-in-time changing characteristics. On–line tools are unlikely to pick up on non–rapidly changing variable values. Using data mining, it is still possible to discover these shifts. An approach for anomaly identification in network traffic monitoring, the Red-Blue State Merging Algorithm and the RTI Algorithm with process, are discussed in the study. The results of this study are based on genuine data from github, which is used in the paper.

Keywords: Red-Blue State Merging Algorithm, RTI Algorithm

I. Introduction:

The importance of safety cannot be overstated across a wide range of daily activities, from construction projects (in both the development and operation phases) to travel (by plane, train, and other modes of transportation). As our reliance on the Internet and related technologies grows, so does the need to ensure the safety and security of network traffic.

© 2022, IJCSMC All Rights Reserved

Monitoring network traffic can be done with a variety of real-time technologies. Tracking changes in traffic at a lower frequency is also significant. Analyzing network traffic monitoring data, the research provides an anomaly detection method that focuses on tiny dynamics. It is vital to keep in mind that not every abnormality that is discovered is necessarily a threat. A network security officer will need to look into these findings further before making a final judgment on whether to proceed.

It is important to note that irregularities in network traffic can have a substantial impact. Abnormalities encompass both lawful and illicit actions like as temporary shifts in client demand, flash crowds, etc. [1]. Our societal and economic infrastructures rely heavily on networks these days. This makes network security very critical, and detecting network traffic anomalies is an essential element of this. Despite the obvious relevance of network anomaly detection, there are no well-written surveys or instructional papers on the issue. This study aims to remedy this shortcoming by giving a lesson on anomaly identification in network traffic. It is our belief that network anomaly detection will become increasingly important and useful in the future.

Network anomaly detection may be divided into two broad categories: signature-based and non-signature-based techniques. In the signature-based technique, anomaly detection is done by looking for patterns that match established anomalous signatures. The regularity of IP addresses, for example, can be used to detect denial of service (DoS) activity [2]. This method's drawback is that it can't discover fresh abnormalities because the signature must already be known. Network traffic is analyzed using statistical methods in the second approach.

If you don't already know what you're looking for, this method can help you uncover new abnormalities. Non-signature-based anomaly detection is the emphasis of this research. Non-signature-based anomaly detection is described in this study in three primary ways: principal-component.

A combination of PCA (sketching) and signal analysis (PCA). Because of these three factors, it is difficult to compare different methodologies. Anomalies come in various forms. It is common for evaluation studies to focus on only a small number of abnormalities, and the data sets used in these research vary widely. In the face of a wide range of abnormalities, no one strategy consistently outperforms the others [3]. For each approach, new improvements are continuously being discovered.

© 2022, IJCSMC All Rights Reserved

Although the sketch-based technique is often less computationally and storage intensive than the other two ways, the other two approaches can occasionally produce higher detection performance. On the last page of the paper, we present a framework that integrates all three methods.

Automated analysis techniques including as PCA, sketching, time series analysis, wavelets, and the Kalman filter are now often used to detect network anomalies. In order to make this lesson accessible to a wide range of readers, we've divided the content into two sections. Our initial presentation is in Section II, where we cover the fundamental ideas of PCA, sketch, and wavelet-based techniques to detecting network anomalies. Non-specialists might benefit from the first level since it explains the fundamental concepts that underlie the various techniques to network anomaly detection.

The second level of the article, which is meant for the expert, is contained in the following sections. There should be a solid foundation in signal processing for the reader, and they should be prepared to learn about some of the most advanced signal processing techniques, which will likely entail a lot of mathematical formulae. In order to help the advanced reader develop their own network anomaly detection strategies, we give a very thorough explanation of the methodologies addressed in the study. "Anomaly detection strategies for each of the three network anomaly detection approaches are provided in a sequential order, except for the third approach, where the schemes described there are independent of each other. Table I gives an overview of the network outlier detection strategies proposed in this study.

Approaches	Progression of schemes
PCA-based	The basic scheme Using traffic feature distributions Using sketch subspaces Detecting small-volume, correlated anomalies Distributed PCA Problems and solutions with the PCA-based approach
Sketch-based	The basic scheme Identifying hierarchical heavy hitters Using sketches and non-Gaussian multi-resolution statistics
Signal-analysis-based	Statistics-based Wavelet-based Kalman-filter-based
Synthesis of different approaches	Network anomography
Anomaly extraction	Using associate rule mining

TABLE 1 Various Approaches to Network Anomaly Detection

© 2022, IJCSMC All Rights Reserved

II. BASIC CONCEPTS:

A. Network Traffic Anomaly:

It is important to note that irregularities in network traffic can have a substantial impact.

These abnormalities can include variations in connection traffic volume, IP source and destination address and port number distribution patterns, and so forth. Anomalies can be caused by both legal and illegal activity [4]. Transient variations in client demand, flash crowds, routing table modifications, and so on are all legitimate activities. DDoS, port scans, virus and worms, etc. are examples of illegal acts. Traffic volume time series anomalies are depicted in Figure 1 at the OD (source) and link traffic (destination) levels. At the connection traffic level, anomalies that are not immediately quantifiable are caused by anomalies that are directly detectable. The abnormalities at the OD flow level, which occurred at the period indicated by the red dots, are clearly visible. In contrast, irregularities at the connection traffic level are significantly less noticeable to the naked eye. Links c-d and b-c in Example 1 include irregularities that are difficult to detect, for example. As a result, identifying network anomalies is a difficult task.

As far as we know, there is no official definition for outliers. According to one of the ideas, "an outlying observation is one that seems to diverge considerably from other members of the sample in which it occurs. Many outlier detection methods have been developed throughout the years. In general, statistical and density-based outlier identification approaches are the most common. Statistical techniques focus on a single variable or variable set. In the case of multidimensional data analysis, this method necessitates further post processing of the produced results. When an item becomes an outlier, it must be defined: if at least one variable value is pointed to as and outlier, an assumed proportion of variables behave in such a way, or values of all variables are pointed as outlying observations.

B. The PCA-Based Approach

N-dimensional data points are mapped onto new major axes vi, where vi = 1, 2,..n in the PCA coordinate transformation. The following is a list of the properties of the primary components. It is clear that the first main component is pointing in the

© 2022, IJCSMC All Rights Reserved

direction of the data's largest variance. When the first principle component is removed, the second principal component indicates the highest residual variation in the residue data, and so on.

Principal components are ranked in order of how much variance they account for. PCA is frequently used to reduce dimensions. If k n main components account for most of the variation in n-dimensional data, then the data dimension can be reduced to k. Using PCA to detect network anomalies is described in the following sections, first using traffic volumes and subsequently using traffic attributes such as IP addresses and port numbers to identify anomalies in the network.

C) Volume-Based Detection:

PCA may be used to detect traffic volume irregularities as follows: [5]. As an input to the PCA model, we need to know how much traffic is flowing through each of the network's n connections during each of the m time slots in the measurement window. We may calculate this by multiplying the n-by-m traffic matrix A by ai,j. You can see how this is done by normalizing A to get X, which has zero mean and unit variance. We use PCA to analyse X. As an example, let ui represent the projection of the data onto the primary axis. U1 has the largest volatility, followed by u2, u3, and so on. For example, we define an empirical threshold uT, which means that all the values in the normal set, and all the values in the abnormal set, fall within this range. There are two sets of axes, one normal and the other aberrant. Their major axes correspond. x1 and x2 are the link traffic vectors projected onto V1 and V2 from x (a given column of X). An anomaly is declared if kx2k 2 exceeds a certain threshold. Q-Statistics, a statistical test, is used to assess the needed degree of confidence. [31]

For PCA anomaly detection to operate, connection traffic must be low effective dimension in order for it to work. There is a lot of variability in connection traffic across different network conditions, as seen in Figure 2. There is a lot of variety in the first three or four major components. PCA's ability to detect network anomalies is based on the low effective dimensionality of traffic data.

© 2022, IJCSMC All Rights Reserved

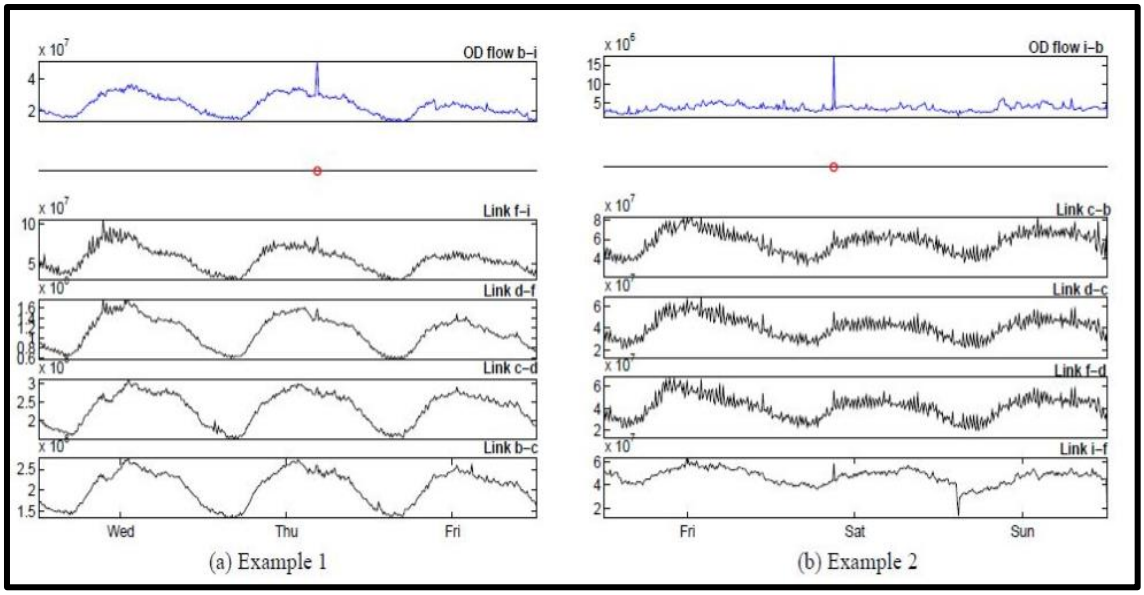


fig. 1. Illustration of anomalies at the OD flow level (top row) and at link traffic level (lower four rows)

Anomaly	Definition	Traffic feature distributions affected
Alpha flows	Large-volume point-to-point flows	SIP, DIP (possibly SP, DP)
DoS	Denial of service attack (distributed or single-source)	SIP, DIP
Flash crowd	Large volume of traffic to a single destination, typically from a large number of sources	DIP DP
Port scan	Probe to many destination ports on a small number of destination addresses	DIP, DP
Network scan	Probe to many destination addresses on a small number of destination ports	DIP, DP
Outage events	Traffic shifts because of equipment failures or maintenance	SIP, DIP
Point-to-multipoint	Traffic from a single source to many destinations, e.g., content distribution	SIP, DIP
Worms	Scanning by worms for vulnerable hosts, which is a special case of network scan	DIP, DP

Table ii traffic feature distributions affected by various anomalies

III. Algorithm Used and Result Outcomes:

Red-Blue State Merging Algorithm:

Flex fringe is developed on top of the Evidence-Driven blue fringe State-Merging algorithm (EDSM), which was initially described in [48]. For labelling and merging states that are congruent, the winning DFA inference technique in the Abbadingo One DFA Learning Competition utilized a red-blue frame work. An Augmented Prefix Tree Acceptor (APTA) is built using the input data and merges acceptable pairs of states until a compact hypothesis is formed. When the activities associated with each state provide the same result, merging their

© 2022, IJCSMC All Rights Reserved

states is judged appropriate (or consistent) (accepting or rejecting). The entering and outgoing transitions of the two states it was merging from are the transitions of a newly united states. Merging options, such as a succession of similarly labelled outbound transitions to distinct state types, are also merged into one. It is termed as purpose when all nondeterministic choices are explored. No matter what level of this resolution method it is at, the starting pair of variables cannot be united when two conflicting states are merged. Two things may be gleaned from such an approach. A DFA-based learning system, EDSM employs both positive and negative training data in order to construct an accurate model of the world. Additional restrictions are imposed on the future state mergers, which may prove erroneous if an incorrect merging was performed. Since an incorrect merging might have an avalanche impact on learning, the algorithm must stick to the well combining approach backed up to the utmost extent by the input data.

Red-blue combines may be performed quickly by retaining a core of red knobs and a periphery of blue knobs, which reduces the number of feasible merge pairings. The merging procedure and the calculation of the evidence score are based on this technique. Red nodes indicate the components of an automaton, whereas blue nodes represent the offspring of red nodes that operate as merger possibilities.

For example, if a merger of red and blue cannot be accomplished through a slightly red node, the learner might do one of three things: 1) compute the node for merging the pair, 2) paint the bulge of the blue the colour of the red, or 3). Distinct methods and sequences for generating evidence scores, as well as the red-blue state portrayal, may result in different algorithms with varying performance. The original study reported in [6] used the difference between accepting and rejecting states as the evidential score associated with each prospective merger. However, as far as the sequence.

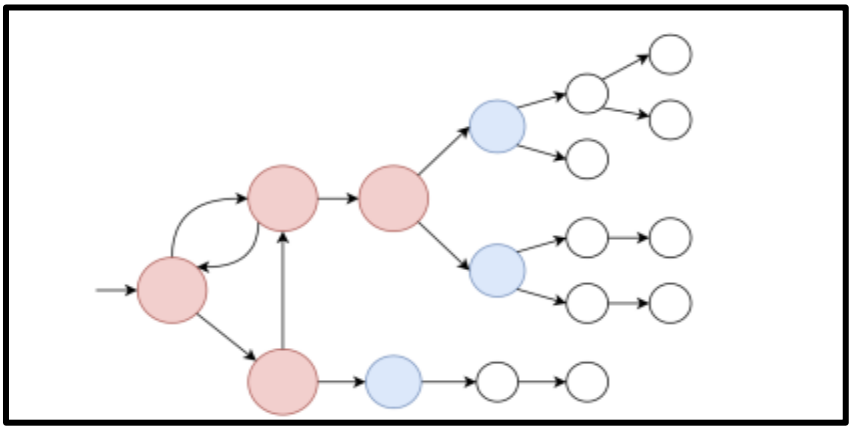


Figure 2: A picture of an example DFA through the purpose procedure in the red-blue outline

© 2022, IJCSMC All Rights Reserved

The RTI Algorithm:

Flexfringe was originally built on the base of an extension of the standard red-blue fringe EDSM approach for DFA credentials. Deterministic real-time mechanisms were the focus of this strategy (DRTA). DRTAs, as discussed in this proposal, have the ability to capture the behavior of real-world systems more effectively. The timing of the occurrences that are identified by these systems has a substantial influence on the performance of such systems. In order to cope with real-time systems, the Real Time Identify (RTI) approach was created in, however, there are two major differences between this strategy and the one that was previously discussed: Each change is associated with a suspension protector that represents the timing requirement linked to that change, which makes the outline of a scheduled indication worth essential. (2) A new SPLIT operation is introduced in the consolidation procedure along with the MERGE and COLOR procedures, which is essential. The delay gates in the APTA were a source of inconsistencies, and the method's authors intended to help fix them. They were able to do this by including the split function in the procedure, which is discussed in further detail below.

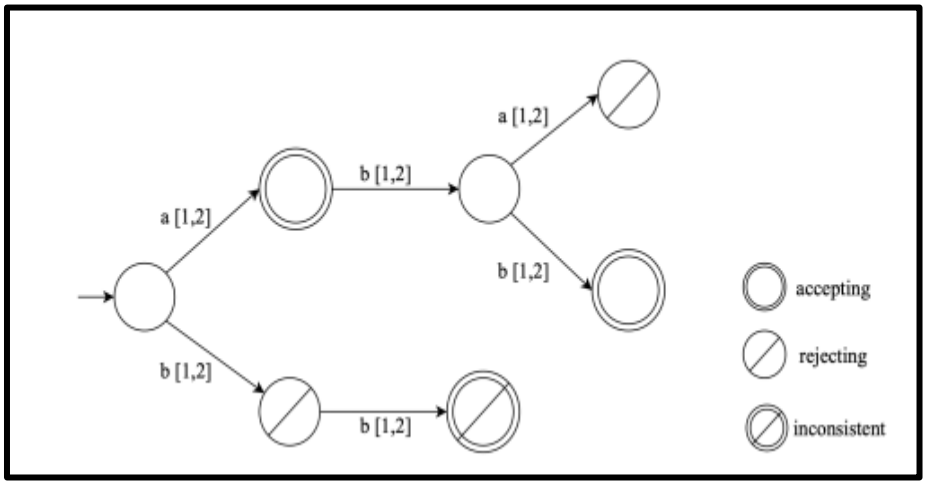


Figure 3: a real-time APTA generated

IV. Experimental Structure:

A conventional assessment strategy for machine learning tasks is used in this thesis, for each database separated hooked on an exercise dataset, a validated model and a trial set, with the development process separated hooked on the exercise stage, the tuning/validation phase, and lastly, the difficult phase. A detailed description of the generation of these sets, but a small example is presented in this line as well. The preparation set was only recycled throughout the

© 2022, IJCSMC All Rights Reserved

system's preparation stage, whereas the other 2 sets were only recycled throughout the analysis stage.

The training data was made up entirely of benign network traffic, and it was used to create a "database" of benign behavioral profiles by extracting multivariate FSMs from NetFlow data and fitting detection models to each FSM. There were two kinds of procedures for evaluating employed throughout the validation process. The first kind included benign traffic as well, and it was used as a analysis effort to the built scheme so that the knowledge process' limits could be fine-tuned for a low false alert rate. The 2nd kind included mixed traffic and was used in standard contrast studies to determine the optimum performance setup for the current plan.

The choice of the degree of analysis of the NetFlow data is a crucial part of the tests. As previously stated, the current proposal may work at a host and connection levels of analysis; however, the preferable level is determined by the nature of the incoming information and the desired degree of assessment.

Detection fineness. As an instance, if a joining equal of examination is chosen, an additional fine grained detection process can be accepted, meanwhile the kind social replicas inferred from the circulation documented inside apiece innocuous link will implement abundant fewer sound compared to designs derived from the traffic of each host in the network. In addition, by detecting at the correlation level, the results can be best separated and construed in contrast to the results acquired for apiece host, particularly when a host is related by both kind and hateful performances, with the malicious circulation including a low underground of its recognized movements.. A significant number of flows must be provided for both the training and testing stages of the network in order to implement a connection-level detection technique. Because of the large number of flows involved, it is impossible to extract an accurate behavioral profile from such a connection, and even if it is found during testing it is extremely questionable if identification findings acquired from this linkage are accurate.

It is necessary to talk about the types of tests and included in the assessment process after determining the analytical level to be employed in the detection technique. Results were compared in the research in accordance with one of 4 groups. Evaluation of FSM detection techniques integrated in their structure will begin first. Consequently, the effectiveness of the protection system and detection algorithm utilized to create the benign communication profiles

© 2022, IJCSMC All Rights Reserved

may be judged. There are several ways to make meaningful conclusions about the influence of apiece finding method to the learning and detection processes. A second consideration is the relative value of the various NetFlow feature sets in achieving the detection objective. The mixed training technique uses just fundamental NetFlow characteristics, as previously described in this thesis, in an effort to make the created system as widely applicable as feasible. Though NetFlow characteristics were present in two of the datasets, just a subset of the most typical elements were used in the analyses.

Dataset	Training set	Test set
CTU-13	Benign flows from Scenario 3	Remaining Scenarios + Malicious flows from Scenario 3
UNSW-NB15	Benign flows from Scenario 1	Remaining Scenarios + Malicious flows from Scenario 1
CICIDS2017	Monday	Rest Days

Table iii: Preparation and Examination sets divided for respectively database

Outcomes on CTU-13 Dataset:

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	3	2	0	0	5	0	1	0	5	0	1
2	0	1	1	1	1	2	0	0	1	2	0	0
3	1	4	1	0	1	4	1	0	1	4	1	0
4	0	3	1	1	1	4	0	0	1	4	0	0
5	1	3	0	0	1	3	0	0	1	3	0	0
6	0	2	1	1	1	2	1	0	0	2	1	1
7	1	3	0	0	1	3	0	0	1	3	0	0
8	0	3	1	1	1	4	0	0	0	4	0	1
9	1	3	2	9	1	5	0	9	3	5	0	7
10	10	3	2	0	10	5	0	0	10	3	2	0
11	2	1	1	0	2	1	1	0	2	1	1	0
12	3	3	0	0	3	3	0	0	0	2	1	3
13	1	3	1	0	1	4	0	0	1	4	0	0
Total	21	35	13	13	24	45	3	10	21	42	6	13

Results for feature set 1

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	3	2	0	0	4	1	1	0	4	1	1
2	1	1	1	0	1	2	0	0	0	2	0	1
3	1	3	2	0	1	4	1	0	1	4	1	0
4	1	4	0	0	1	4	0	0	0	4	0	1
5	1	3	0	0	1	3	0	0	1	3	0	0
6	1	2	1	0	1	2	1	0	0	3	0	1
7	1	3	0	0	1	3	0	0	0	3	0	1
8	1	4	0	0	1	4	0	0	0	4	0	1
9	1	4	1	9	0	4	1	10	0	4	1	10
10	10	4	1	0	10	4	1	0	10	4	1	0
11	2	1	1	0	2	1	1	0	2	1	1	0
12	3	3	0	0	3	3	0	0	3	3	0	0
13	1	4	0	0	1	4	0	0	0	4	0	1
Total	25	39	9	9	23	42	6	11	17	43	5	17

Results for feature set 2

© 2022, IJCSMC All Rights Reserved

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	4	1	0	1	4	1	0	1	4	1	0
2	1	2	0	0	1	2	0	0	1	2	0	0
3	1	5	0	0	1	4	1	0	1	4	1	0
4	1	4	0	0	1	4	0	0	1	4	0	0
5	1	3	0	0	1	3	0	0	1	3	0	0
6	1	3	0	0	1	2	1	0	1	3	0	0
7	1	3	0	0	1	3	0	0	1	3	0	0
8	1	4	0	0	1	4	0	0	1	4	0	0
9	10	4	1	0	10	4	1	0	10	4	1	0
10	10	4	1	0	10	4	1	0	10	4	1	0
11	2	1	1	0	2	1	1	0	2	1	1	0
12	3	3	0	0	3	3	0	0	3	3	0	0
13	1	4	0	0	1	4	0	0	1	4	0	0
Total	34	44	4	0	34	42	6	0	34	43	5	0

Results for feature set 3

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	3	2	0	0	4	1	1	0	4	1	1
2	1	1	1	0	1	2	0	0	1	2	0	0
3	1	3	2	0	1	4	1	0	1	4	1	0
4	1	3	1	0	0	4	0	1	1	4	0	0
5	1	3	0	0	1	3	0	0	1	3	0	0
6	1	2	1	0	1	2	1	0	1	2	1	0
7	1	3	0	0	1	3	0	0	1	3	0	0
8	1	3	1	0	1	4	0	0	1	4	0	0
9	1	3	2	9	0	4	1	10	1	5	0	9
10	10	3	2	0	10	4	1	0	10	4	1	0
11	2	1	1	0	0	1	1	2	2	1	1	0
12	3	3	0	0	3	3	0	0	3	3	0	0
13	1	3	1	0	1	4	0	0	1	4	0	0
Total	25	34	14	9	20	42	6	14	24	43	5	10

Results for feature set 4

Table iv: Finding consequences per piece conventional and uncovering procedure used on the CTU-13 database with only the main hosts comprised

The LOF based technique has the lowest frequency of false alarms, but all three setups are able to detect all hostile hosts. While the Isolation Forest approach, which uses the first set of features (protocol, source bytes, and destination bytes), and the Gaussian KDE approach, which uses the fourth established of structures (destination port, protocol, duration, source bytes and destination bytes), provide low false alarm rates, it should be noted that these locations are associated with a lower discovery rate because they misclassify data.

Results on UNSW-15 Dataset

Similar to CTU-13, initial studies on UNSW-NB15 focused on testing the suggested technique on different feature sets and algorithms that fit in various stages of the model. Table 5 shows the findings at each of the dataset's four divisions, including traffic exclusively from significant hosts.

In order to properly test the detection success of the proposed technique, only the most important hosts of the dataset were originally considered. CTU-13 data shows that the bulk of the dataset's total flows are related to a few big hosts, much as in the CTU-13 dataset.

© 2022, IJCSMC All Rights Reserved

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	10	0	3	1	10	0	3	1	10	0	3
2	1	10	0	3	2	10	0	2	2	10	0	2
3	1	10	0	3	1	10	0	3	1	10	0	3
4	1	10	0	3	2	10	0	2	2	10	0	2
Total	4	40	0	12	6	40	0	10	6	40	0	10

Results for feature set 1

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	10	0	3	0	10	0	4	0	10	0	4
2	2	10	0	2	0	10	0	4	2	10	0	2
3	1	10	0	3	0	10	0	4	1	10	0	3
4	2	10	0	2	1	10	0	3	2	10	0	2
Total	6	40	0	10	1	40	0	15	5	40	0	11

Results for feature set 2

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	3	10	0	1	1	10	0	3	1	10	0	3
2	3	10	0	1	2	10	0	2	2	10	0	2
3	4	10	0	0	1	10	0	3	1	10	0	3
4	3	10	0	1	0	10	0	4	2	10	0	2
Total	13	40	0	3	4	40	0	12	6	40	0	10

Results for feature set 3

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	1	10	0	3	0	10	0	4	0	10	0	4
2	1	10	0	3	0	10	0	4	4	10	0	0
3	1	10	0	3	0	10	0	4	1	10	0	3
4	0	10	0	4	1	10	0	3	2	10	0	2
Total	3	40	0	13	1	40	0	15	7	40	0	9

Results for feature set 4

Table v: Finding consequences per piece set and uncovering procedure recycled on the UNSW-NB15 database with individual the main hosts comprised

For the highest-performing setup, it's clear that the LOF-based technique acting on the third piece set delivers the greatest metrics compared to the others. It is clear from looking at the findings that various detection techniques outperform one another for different feature sets. There's little doubt that LOF-based models do best with sets 2 and 3, but models based on the Gaussian KDE framework do better in sets 4 and 5. Both the Isolation Forest and the Gaussian KDE techniques appear to be appropriate for the initial feature set. It is wealth noting that all settings accurately classify all kind hosts, with the highest execution shape identifying around 80 percent of dangerous hosts. In Figure 4, the aggregated findings from the full dataset are shown for each confer, making these results easier to comprehend.

© 2022, IJCSMC All Rights Reserved

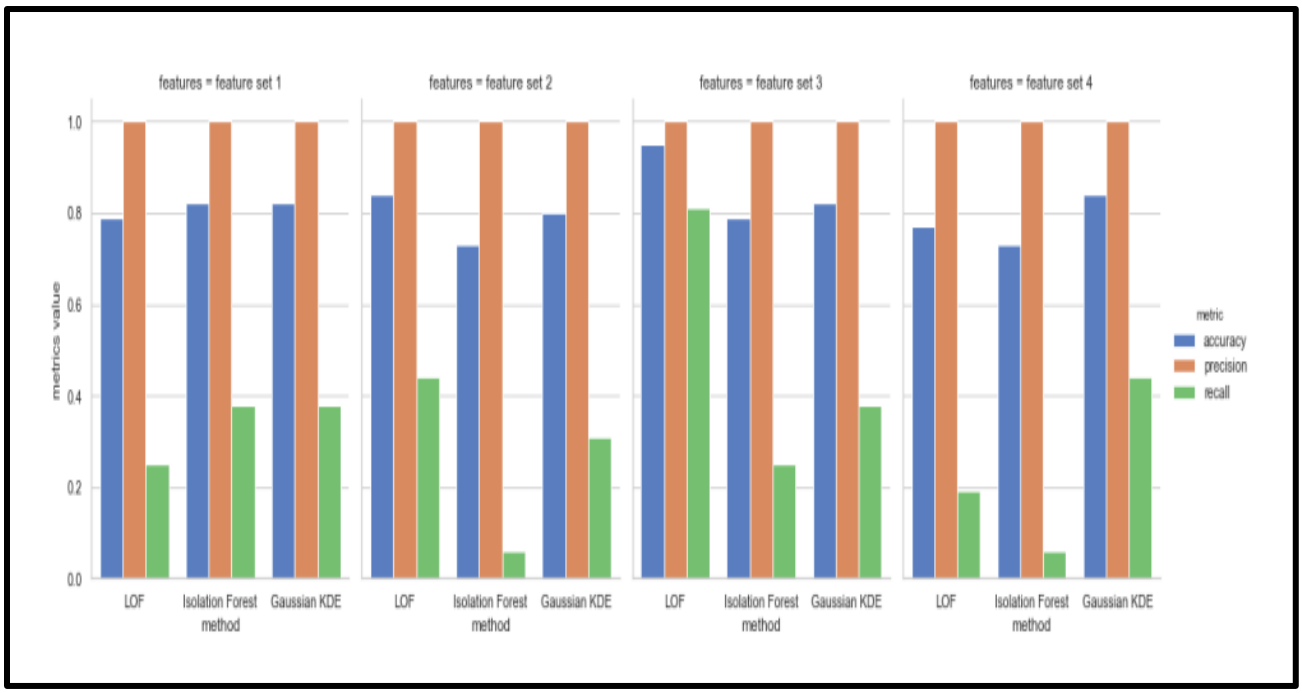
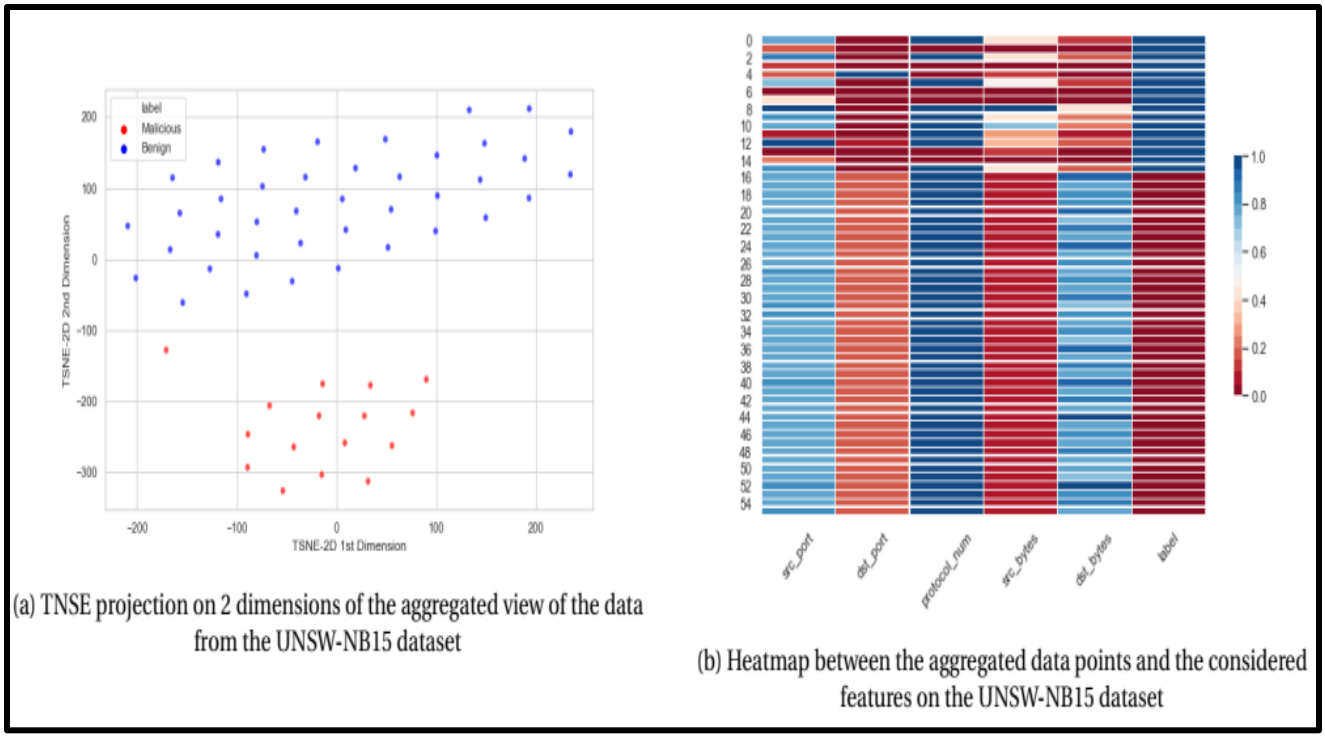


Figure 4: Combined assessment metrics per recognition technique and piece set for the UNSW-NB15 database

As shown in Figure 4 the LOF-based strategy acting on the 3rd piece usual delivers the finest provide a from all assessment measures. Although, as previously stated, all configurations have a precision value of 100 percent since all innocuous hosts are properly recognized, the recall values obtained for the majority of the evaluated settings are extremely low. Only the greatest setup (with a recall value of 81 percent) manages to discover more than half of the malicious hosts, while the other settings fall short. A closer look at the dataset revealed that the hosts that were incorrectly detected in a piece of the dividers existed also linked to an important amount of kind movements, thus complicating the identification process using our suggested technique. It is possible to mitigate some of the effects of this phenomena by combining a LOF-based method with the third feature set. A more fine-grained detection strategy could be used to address this problem, suggesting that a link level study of this information would be advantageous.

© 2022, IJCSMC All Rights Reserved



(a) TNSE projection on 2 dimensions of the aggregated view of the data from the UNSW-NB15 dataset

(b) Heatmap between the aggregated data points and the considered features on the UNSW-NB15 dataset

Figure 5: Conception of the combined opinions in the LOF gathering setting for the UNSW-NB15 database

Results on CICIDS2017 Database

The tests done on the CICIDS2017 database followed the very same trend as those performed on the other data set in the assessment of the suggested classifier. In the beginning, the studies only employed traffic connected to the dataset's key hosts, with 80 percent of the overall training flows and 91.6 percent of the total test flows comprising the flows used in the testing phase. Table vi shows the findings of the main hosts for each day of this dataset on the various assessments that were reviewed. The training set in this example (Monday flows) had no hostile sites, unlike the prior reported datasets. However, for the sake of transparency, the data from that day has been included in this chart.

© 2022, IJCSMC All Rights Reserved

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
Monday	0	13	0	0	0	13	0	0	0	13	0	0
Tuesday	1	12	1	0	1	12	1	0	1	12	1	0
Wednesday	1	13	0	0	1	13	0	0	1	13	0	0
Thursday (morn.)	1	11	2	0	1	11	2	0	1	13	0	0
Thursday (aft.)	1	11	2	0	1	11	2	0	0	13	0	1
Friday (morn.)	1	10	1	2	1	10	1	2	1	11	0	2
Friday (aft. • DDoS)	1	11	2	0	1	12	1	0	1	13	0	0
Friday (aft. • port scan)	1	12	1	0	1	11	2	0	1	13	0	0
Total	7	93	9	2	7	93	9	2	6	101	1	3

Results for feature set 1

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
Monday	0	13	0	0	0	13	0	0	0	13	0	0
Tuesday	1	12	1	0	1	12	1	0	1	12	1	0
Wednesday	1	13	0	0	1	13	0	0	1	13	0	0
Thursday (morn.)	1	11	2	0	1	11	2	0	1	13	0	0
Thursday (aft.)	1	11	2	0	1	11	2	0	1	11	2	0
Friday (morn.)	1	9	2	2	1	10	1	2	1	11	0	2
Friday (aft. • DDoS)	1	12	1	0	1	11	2	0	1	11	2	0
Friday (aft. • port scan)	1	12	1	0	1	12	1	0	1	12	1	0
Total	7	93	9	2	7	93	9	2	7	96	6	2

Results for feature set 2

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
Monday	0	13	0	0	0	13	0	0	0	13	0	0
Tuesday	0	12	1	1	1	11	2	0	1	12	1	0
Wednesday	0	12	1	1	1	12	1	0	1	13	0	0
Thursday (morn.)	0	13	0	1	1	12	1	0	1	12	1	0
Thursday (aft.)	0	12	1	1	0	9	4	1	0	10	3	1
Friday (morn.)	1	11	0	2	1	10	1	2	1	11	0	2
Friday (aft. • DDoS)	1	13	0	0	1	12	1	0	1	12	1	0
Friday (aft. • port scan)	1	12	1	0	1	12	1	0	1	12	1	0
Total	3	98	4	6	6	91	11	3	6	95	7	3

Results for feature set 3

Scenario	LOF				Isolation Forest				Gaussian KDE			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
Monday	0	13	0	0	0	13	0	0	0	13	0	0
Tuesday	0	12	1	1	0	12	1	1	1	12	1	0
Wednesday	0	13	0	1	1	12	1	0	1	11	2	0
Thursday (morn.)	0	12	1	1	1	11	2	0	1	11	2	0
Thursday (aft.)	1	12	1	0	1	11	2	0	1	11	2	0
Friday (morn.)	1	10	1	2	1	9	2	2	1	9	2	2
Friday (aft. • DDoS)	1	12	1	0	1	11	2	0	1	11	2	0
Friday (aft. • port scan)	0	13	0	1	1	11	2	0	1	11	2	0
Total	3	97	5	6	6	90	12	3	7	89	13	2

Results for feature set 4

Table VI: Finding consequences per piece set and finding procedure recycled on the CICIDS2017 database with merely the main hosts comprised

It's clear that the Gaussian KDE-based strategy, using the initial selection of characteristics and logging just one false alert, yields the greatest performance across all measures. It should be noted at this point that there are other settings that show a somewhat advanced amount of noticed hosts but the Gaussian KDE systematic model functioning on the first set of features is measured the best accomplishment one, meanwhile it knowingly outdoes all of these outlines in positions of untrue alarm amounts. That there isn't as much difference in the finest detection methods across diverse piece circles as in the additional two database is another intriguing discovery. The Gaussian KDE-based models, in instance, prefer to It's clear that the Gaussian KDE-based strategy, which uses the initial selection of characteristics and logs only one false alert, has the best performance across all measures. At this point, it's important to note that there are other settings that show a somewhat advanced amount of noticed hosts (in particular, one more host), but the Gaussian KDE systematic model that works on the first set of features is thought to be the best because it has a much lower rate of false alarms than all of these other settings. Another interesting finding is that the best detection methods don't differ as much between different

© 2022, IJCSMC All Rights Reserved

features sets as they do between the other two datasets. Gaussian KDE-based models, for example, are more likely to.

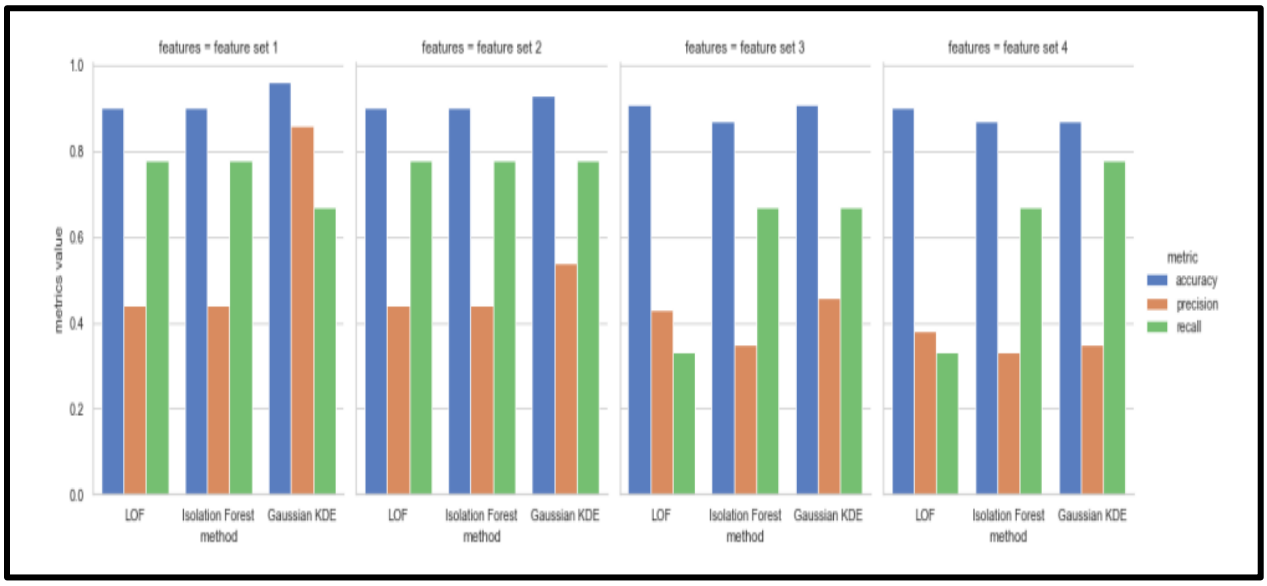


Figure 6: Combined assessment metrics per discovery technique and piece set for the CICIDS2017 database with only the main hosts involved

Figure 6 shows the values of correctness, exactness, and memory for all of the setups that were tested across the whole dataset. It's easy to see that the Gaussian KDE-based approach that uses the first piece set does better than all the additional shapes in two of the three metrics that were looked at.

© 2022, IJCSMC All Rights Reserved

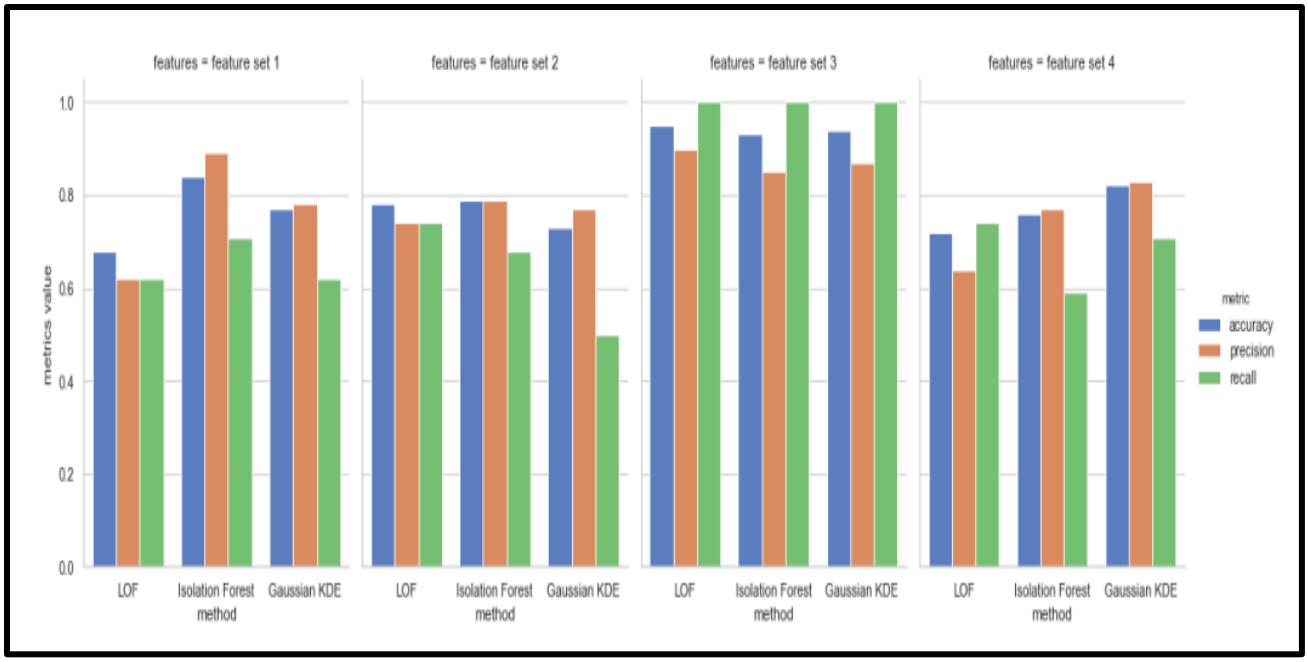


Figure 7: Combined assessment metrics per discovery technique and piece usual for the CTU-13 database with only the main hosts involved

Going on after the highest execution parameters, it must be piercing out that the finest execution finding procedure changes across the dissimilar piece circles utilized for assessment. When the initially piece established is employed, the Isolation Forest replicas easily beat the remainder algorithms. For the instant article agreed (destination port, protocol, source bytes, destination bytes) both the LOF and the Isolation Forest methods offer equivalent consequences, with the first finding significantly extra dangerous hosts, and the instant linked with somewhat fewer false alarms.

V. Conclusion:

In the paper the outlier analysis approach for network traffic data anomaly detection was presented. Based on real data an off–line analysis was performed, while an on–line analysis was modelled with the proposed methodology. Discussed algorithms of anomaly detection were used and two variants of reporting frequency were checked. We believe network anomaly detection will become more important in the future because of the increasing importance of network security.

© 2022, IJCSMC All Rights Reserved

References

[1]. Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund "Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications," in *Proc. of the 4th ACM SIGCOMM conference on Internet measurement (IMC)*, 2004, pp.101-114.

[2]. D. Moore, G. Voelker, and S. Savage, "Inferring internet denial of service activity," in Proc. of the USENIX Security Symposium, August 2001, pp. 9-22.

[3]. K. Nyalkalkar, S. Sinha, M. Bailey, and F. Jahanian, "A comparative study of two network-based anomaly detection methods." in Proc. Of IEEE INFOCOM, 2011, pp. 176-180.

[4]. Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund "Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications," in Proc. of the 4th ACM SIGCOMM conference on Internet measurement (IMC), 2004, pp.101-114.

[5]. A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network wide traffic anomalies," SIGCOMM Computer Communications Review, vol. 34, no. 4, pp. 219-230, 2004.

[6]. Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in Proc. of 27th Annual Asilomar Conference on Signals, Systems, and Computers, 1993

[7]. Anup Bhange "DDoS Attacks Impact on Network Traffic and its Detection Approach" International Journal of Computer Applications (0975 – 8887) Volume 40– No.11, February 2012

[8]. Anup Bhange "Anomaly detection in network traffic: A statistical approach" International Journal of IT, Engineering and Applied Sciences Research (IJIEASR) Volume 1Issue3Pages16-20

[9]. Anup Bhange "Comparative analysis of several cryptography algorithm with its effectiveness towards the security and its performance" Journal of the Gujarat Research SocietyVolume21Issue6Pages42-46

[10]. https://www.flowmon.com/en/blog/science-of-network-anomalies

[11]. Anup Bhange "Design and Evaluation to Calculate the Performance of Hybrid Cryptography to make Secure Transaction over Network" Turkish Journal of Computer and Mathematics Education Vol.11 No.3 (2020), 1208-1218

[12]. M. Li. An approach to reliably identifying signs of DDOS flood attacks based on LRD traffic pattern Recognition. Computers & Security, 23(7): 549-558, 2004**.**

[13]. C.S. Sastry, S. Rawat and A.K. Pujari. Network traffic analysis using singular value decomposition and multiscale transforms. Information Sciences, 177(23): 5275-5291, 2007**.**

[14]. H. Hajji. Statistical analysis of network traffic for adaptive faults detection. IEEE Transactions on Neural Networks, 16(5):1053–1063, September 2005

[15]. O. Salem, S. Vaton, and A. Gravey. An efficient online anomalies detection mechanism for high-speed networks. In IEEE Workshop on Monitoring, Attack Detection and Migitation (MonAM 2007), November 2007.

[16]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. of 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile, pp. 487-499. Morgan Kaufmann, September 12-15, 1994

[17]. H. Arsham, Time series analysis and forecasting techniques, http://obelia.jde.aca.mmu.ac.uk/resdesgn/arsham/opre330Forecast.htm

[18]. P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies", in Proc. of Internet Measurement Workshop, 2002, pp. 71-82.

© 2022, IJCSMC All Rights Reserved