International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 5
May 2017

# A Comprehensive Model for Handles the Traffic of Computer Networks and Uncovers Anomaly in Real Time

**Miss. Prachi N. Deshmukh,Dr. V. M. Thakare, Prof. R. N. Khobragade**
Department of Computer Science,
SGBAU, Amravati, India.

*Abstract-The large computer and communication networks lead to the generation of massive data flows. The difficulty of analyzing and managing these data in network security degrades the online detection of intrusion and suspicious connections. In this paper, to overcome this problem, proposeda comprehensive model that handles the traffic of computer networks and uncovers intrusions in real time. The model consists of dataset generator and intrusion detector. The dataset generator captures, analyzes and manages the live traffic using a dynamic queuing concept. It continuously constructs connection vectors from the live traffic and exports them either as datasets or sequentially into a pipe for further processing. The intrusion detector is based on an enhanced growing hierarchical self organizing map which classifies exported vectors to normal, anomaly or unknown connections. The model has been evaluated using synthetic and realistic data sources. It is able to process data flows within significant time and classifies the connections in the online mode effectively.*

***Keywords: intrusion detection, data aggregation, real time systems, performance monitoring***

## I. INTRODUCTION

With the emergence of High Speed Network (HSN), the manual intrusion alert detection become an extremely laborious and time-consuming task since it requires an experienced skilled staff in security fields and need a deep analysis [1]. The online operation mode of Intrusion Detection Systems has become a real challenge since the amount of generated heterogeneous and non-stationary data and the interconnection between communications networks are increasing rapidly [2]. Adult websites attract a large number of visitors and account for a substantial fraction of the global Internet traffic. However, little is known about the makeup and characteristics of online adult traffic [3]. Per-flow traffic measurement, which is to count the number of packets for each active flow during a certain measurement period, has many applications in traffic engineering, classification of routing distribution or network usage pattern, service provision, anomaly detection, and network forensics [4].

The problems of traffic matrix (TM) estimation and anomaly detection utilizing link-load traffic measurements. Models including structure regularized traffic monitoring (SRTM) and dynamic SRTM (DSRTM) are presented to realize traffic monitoring under static and dynamic routing configurations, respectively. Considering that real traffic data are usually approximately low-rank but exhibit strong spatial and temporal dependencies [5].

In this proposed methodology, propose a comprehensive model that handles the traffic of computer networks and uncovers intrusions in real time. The model consists of dataset generator and intrusion detector. The dataset generator captures, analyzes and manages the live traffic using a dynamic queuing concept. It continuously constructs connection vectors from the live traffic and exports them either as datasets or sequentially into a pipe for further processing. The intrusion detector is based on an enhanced growing hierarchical self organizing map which classifies exported vectors to normal, anomaly or unknown connections. The model has been evaluated using synthetic and realistic data sources. It is able to process data flows within significant time and classifies the connections in the online mode effectively.

## II. BACKGROUND

In addition, the batch model of alert management is no longer adequate given that labeling is a continuous time process since incoming intrusion alerts are often collected continuously in time. Furthermore, the static model is no longer appropriate due to the fluctuation nature of the number of alerts incurred by Internet traffic fluctuation nature [1]. The voluminous traffic amount increases system vulnerabilities and leads to emerge new and complex types of attacks. In order to effectively detect these attacks, the traffic must be professionally processed and analyzed. Accordingly, IDS models could operate effectively and achieve a high-performance over computer networks only in offline mode. Usually, IDS models are trained using an offline andoutdated dataset such as KDDCup99, GureKDD, and Koyot2006+. However, KDDCup dataset is still the most used benchmark in evaluating network security applications although there are many criticisms about the data collection method and the characteristics of the data as well [2]. The first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network. The data set contains approximately 323 terabytes worth of traffic from 80 million users, and includes traffic from several dozen major adult websites and their users in four different continents. The author analyze several characteristics of online adult traffic including content and traffic composition, device type composition, temporal dynamics, content popularity, content injection, and user engagement [3].

In order to keep up with the high throughput of modern routers or switches, the online module for per-flow traffic measurement should use high-bandwidth SRAM that allows fast memory accesses. Due to limited SRAM space, exact counting, which requires to keep a counter for each flow, does not scale to large networks consisting of numerous flows. Some recent work takes a different approach to estimate the flow sizes using counter architectures that can fit into tight SRAM [4].

Define spatial and temporal regularization matrices based on Moore–Penrose pseudoinverse and Laplacian matrix to structurally regularize the TM variables. [5].

This paper introduces Section I introduction. Section II discusses Background. Section III discussesprevious work. Section IV discusses existing methodologies. Section V discusses the attributes and parameters and how these are affected on images. Section VI proposed method and outcome result possible. Finally Section VIII concludes this review paper.

## III. PREVIOUS WORK DONE

HassenSallayet al. (2013)[1], proposes an efficient real time adaptive intrusion detection alert classifier dedicated for high speed network. The classifier is based an online self-trained SVM algorithm with several learning strategies and execution modes.

Maher Salem et al.(2013)[2], proposed a novel online method that constructs connections from the massive data flow for evaluating IDS models. The proposed method overcomes this challenge by using a queuing concept of dynamic window size. It captures network traffic and hosts events constantly and handles them synchronously within time slot windows inside the queue in order to construct connection vectors based on certain features.

Faraz Ahmed et al.(2016)[3], proposedthe first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network. The week-long HTTP logs include traffic from several dozen major adult websites and their users in four different continents. Reduce network traffic by pushing copies of popular adult objects to locations closer to their end-users.

Min Chenet al. (2016)[4], proposed a scalable counter architecture called Counter Tree. And propose a two-dimensional sharing scheme, where each counter can be shared by multiple virtual counters and each virtual counter can be shared by multiple flows.

Q. Zhang et al.(2015)[5], proposed a structure regularized traffic monitoring model for traffic matrix estimation and anomaly detection. Define spatial and temporal regularization matrices based on Moore–Penrose pseudoinverse and Laplacian matrix to structurally regularize the TM variables. Besides, in view of the fact that anomalies in traffic usually happen rarely and last briefly, sparsity-regularization is further implemented on traffic volume anomalies. This enables models to jointly deal with the traffic monitoring issues of TM estimation and anomaly detection. The SRTM model is designed for static routing configurations and the online traffic monitoring model DSRTM is presented for dynamic settings.

## IV.EXISTING METHODOLOGIES

### 1]An online self trained alert classifier

An efficient real time adaptive intrusion detection alert classifier dedicated for high speed network. The classifier is based an online self-trained SVM algorithm with several learning strategies and execution modes. Evaluate the classifier against three different data-sets and the performance study shows an excellent result in term of accuracy and efficiency. The predictive local learning strategy presents a good tradeoff between accuracy and time processing. In addition, it does not involve a human intervention which makes it an excellent solution that satisfies high speed network alert management challenges.

An introduce in this section the measures used in this experiments. The method defines the true positive (TP), false positive (FP), true negative (TN), false negative (FN) as follows:

- TP: a true positive alert classified as true positive.
- FP: a false positive alert classified as true positive.
- TN: a false positive alert classified as false positive.
- FN: a true positive alert classified as false positive.

In order to study the performance of different alert classification algorithms, and method compute the following measures: Sensitivity ([TP/(TP +FN)]), Fall-out ([FP/(FP +TN)]), Accuracy ([(TP +TN)/(TP +FN+FP +TN)]), Specificity ([TN/(FP +TN)]), Precision ([TP/(TP +FP)]), Negative predictive value ([TN/(TN + FN)]), False discovery rate ([FP/(FP +TP)]), Matthews correlation coefficient, F measure, online learning processing time.

### 2]Queuing concept of dynamic window size

Processing massive data flow in intrusion detection systems (IDS) become a serious challenge. It is considered as a major deficiency while handling heterogeneous and non stationary data stream to uncover anomaly in the online operational mode. The proposed method overcomes this challenge by using a queuing concept of dynamic window size. It captures network traffic and hosts events constantly and handles them synchronously with in time slot windows inside the queue in order to construct connection vectors based on certain features.

The proposed method consists of data aggregators, queue as container for the predetermined time slot windows, and a dataset exporter. The queue in a method can contain n windows and each window w is a time slot of t seconds, e.g. 5 seconds.

### 3] A large scale and in-depth measurement and analysis of online adult traffic

Present the first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network. The data set contains approximately 323 terabytes worth of traffic from 80 million users, and includes traffic from several dozen major adult websites and their users in four different continents. The author analyze several characteristics of online adult traffic including content and traffic composition, device type composition, temporal dynamics, content popularity, content injection, and user engagement. An analysis reveals several unique characteristics of online adult traffic. The author also analyzes implications of findings on adult content delivery. The findings suggest several content delivery and cache performance optimizations for adult traffic, e.g., modifications to website design, content delivery, cache placement strategies, and cache storage configurations.

### 4]A two-dimensional sharing scheme, two offline method and Enhanced Counter Tree architecture.

The author designs a scalable counter architecture called Counter Tree. And propose a two-dimensional sharing scheme. And also propose two offline methods to estimate flow sizes. Moreover, also propose the Enhanced Counter Tree architecture.

In this method, employ three metrics to evaluate the performance of different per-flow traffic measurement schemes:

**Memory Requirement:**Due to the constraint of on-chip space, the author wants to use as small memory as possible to achieve per-flow traffic measurement. In the sequel refer to SRAM simply as memory. This focuses on the memory requirement for implementing the counter architectures. The collection of flow labels is beyond the scope of this paper. Some memory efficient schemes for flow label collection can be found in literature.

**Processing Overhead**: To keep up with the line speed, the processing overhead for recording a packet should be small, such that the implementation of the measurement module will not cause a performance bottleneck. In most counter architectures, the processing overhead for recording a packet mainly results from memory accesses and hash computations.

**Estimation Accuracy**: With a given memory space, author want the estimates of flow sizes to be as accurate as possible. Let the true size of a flow be *s* and the estimated size be ˆ*s*. And use the relative $Bias(\frac{\hat{S}}{S})$ and relative standard error $StdErr(\frac{\hat{S}}{S})$ to evaluate the estimation accuracy, which are defined as follows:

$$Bias(\frac{\hat{S}}{S}) = E(\frac{\hat{S}}{S}) - 1, \quad \ldots\ldots\ldots\ldots \quad (1)$$

$$StdErr(\frac{\hat{S}}{S}) = \sqrt{Var\left(\frac{\hat{S}}{S}\right)} = \sqrt{\frac{Var\,(\hat{S})}{S}} \quad \ldots\ldots\ldots\ldots\ldots(2)$$

5]**Structure regularized traffic monitoring and dynamic structured compressed sensing-based traffic monitoring**

Define spatial and temporal regularization matrices based on Moore–Penrose pseudoinverse and Laplacian matrix to structurally regularize the TM variables. Besides, in view of the fact that anomalies in traffic usually happen rarely and last briefly, sparsity-regularization is further implemented on traffic volume anomalies. This enables models to jointly deal with the traffic monitoring issues of TM estimation and anomaly detection. The SRTM model is designed for static routing configurations, and the online traffic monitoring model DSRTM is presented for dynamic settings.

### A. SRTM Formulation:

Utilizing the link measurements *Y* and routing matrix *R*, establish the SRTM model in the least-square error sense as follows:

$$\min\frac{1}{2}||\mathrm{Y}-\mathrm{RX}-\mathrm{RA}||_{\mathrm{F}}^{2} \ +\lambda 1\|A\,||_{1}\,1+ \ \Lambda s\|CS \ \ X\,||_{F}^{2}+ \lambda_{T}\|XCT\,||_{F}^{2}$$

$$X,A\epsilon R^{F*T}$$

where *CS* and *CT* are the introduced spatial and temporal constraint matrices indicating the spatial and temporal correlations of TMs, respectively, $\|*\|F$ is the Frobenius norm, $\lambda 1$, $\lambda_S$ , and $\lambda_T \geq 0$ are sparsity-, spatial-, and temporal-regularizingparameters, respectively. The *l*1-norm item $\|A\,||_1$ is adoptedhere as the closest convex surrogate of $\|A||_0$since *l*0-normminimization is NP-hard and cannot be optimally solved.

### B. DSRTM Formulation:

To describe dynamic network settings, consider therouting configuration to be time-varying and denote $R_t$asthe routing matrix during the $r^{th}$ time interval. And define thedynamic traffic monitoring model by

$$y_{t=}R_t(x_t+a_t)+D_t \ ,$$

Where$y_t$,$x_t$, $a_t$, and $y_t$are link-load measurements, OD traffic flows, anomalous flows, and measurement noises in the $r^{th}$ time interval, respectively. According to, although the routing configurations of some real networks change slowly, the assumption that X= $[x_1 x_2. . .x_t]$ lies in an approximately low-rank subspace is still reasonable because of the routing-independent spatiotemporal properties of TMs. With the recursive observations $y_t$ , $R_t$ , and the previously obtained estimations $x_t$ −1, . . . , *x*1,*att*−1, . . . ,*a*1, the author aim to recover the flow-level anomaly free traffic flows $x_t$and anomalous flows *at* in real time by the DSRTM model

$$\min\frac{1}{2}||y_t - R_t x_t - R_t a_t||_2^2 \ +\lambda 1\| \ a_t||_1 \ 1+ \ \lambda_S \ \|CS \ x_t||_2^2 + \lambda_T \lambda_T \textstyle\sum_{t=0}^{t-1} \beta^r \|x_t -, x_{t-r}||_2^2$$

$$X, A\epsilon R^F$$

where$\lambda 1 \geq 0$,$\lambda_S \geq 0$, and $\lambda_T \geq 0$ are sparsity-, spatial-, and temporal-regularizing parameters, respectively, *CS* is the spatial constraint matrix formulated in the same way as that in the static-routing scenario, and $0 < \beta < 1$ is an introduced forgetting parameter that determines the exponentially reduced weights imposed on past data. To see the relation between the DSRTM and SRTM models, we rewrite the optimization problem (10) as

$$\min\frac{1}{2}||y_t - R_t x_t - R_t a_t||_2^2 \ +\lambda 1\| \ a_t||_1 \ 1+ \ \lambda_S \ \|CS \ x_t||_2^2 + \lambda_T \lambda_T \|X_t C_T^{'}||_F^2$$

$$X,A\epsilon R^F$$

where$X_t$= $[x_{1,}x_2, \ . \ . \ . \ ,x_t]\ \epsilon R^{F*T}$denotes the flow-level TM at time instant *t* whose first $t-1$ columns are fixed by the previous estimations $x_1$ , . . . $x_t$ $-1$and

$$C'_T = \left| \begin{array}{ccc} 1 & 0 \ldots & 0 \\ 0 & 1 \ldots & 0 \\ \vdots \ \vdots & & * \\ -1 & -1\ldots & 0 \end{array} \right| \begin{array}{c} \sqrt{\beta^{t-1}} \\ \\ \sqrt{\beta^{1}} \\ \\ \sqrt{\beta^{0}} \end{array} \left| \ \vdots \ \right| \quad \vdots \quad t*t$$

## V. ANALYSIS AND DISCUSSION

The main concluded statements are the following:

• The performance results for both cases are not significantly different.

• The processing time for with-limit case is lowest that without limit. This gain becomes more significant when the data size becomes large.

• By comparing the small difference in performance and the important gain in processing time we can conclude that the limit case is better for us in our further investigation [1].

The evaluated the method in online operational mode at the university campus and in offline operational mode using the dump data from KDDCup99 program. In addition, configure the method to consider these protocols: ARP, ICMP, rcp, UDP and the following services: jtp, ssh, telnet, smtp, smb, njs, xmpp, http, ntp, dhcp, syslog, snmp, rdp. The evaluation of the proposed method has been divided into two parts: the first one demonstrates the results of queuing concept in offline and online operational mode. The second part uses constructed connections from proposed method to evaluate their enhanced GHSOM (EGHSOM) [2].

The authors discuss potential implications of their measurement and analysis of online adult traffic. They are particularly interested in understanding the impact of different content access patterns on CDN caching. To this end, they analyze the caching performance for adult websites by looking at server-side HTTP response codes and cache hit ratios [3].

The author has implemented the two estimators based on the Counter Tree, i.e., CTE and CTM respectively. CTE and CTM share the same module for online packet recording, which performs the operations. Hence, when evaluate the online operations of the Counter Tree, use CT as the abbreviation. The compare those with the most related counter architectures: (1) randomized counter sharing (MLM) and (2) Counter Braids (CB) [4].

If the author impose low-rank constraints instead of spatiotemporal constraints on the considered anomaly free traffic flows, i.e., set $\lambda_s$and $\lambda_T$as 0’s and replace the structural regularization items by nuclear-norm minimization $\lambda_* ||RX||_*$, the SRTM model transforms into the LPS model recently proposed. Despite the theoretical guarantee of the LPS model in uniquely recovering low-rank and sparse matrices, real traffic data might not always satisfy the required conditions. To be specific, theoretical conditions of the LPS model assume the TMs to be low-rank, but most real traffic data is approximately low-rank. On the other hand, the LPS model supposes that the routing matrix $R$ is orthogonal by rows and has strong column incoherence, which might be violated by many real traffic models, as they usually exhibit structural dependencies instead of strict stochasticity [5].

## VI. PROPOSED METHODOLOGIES

The proposed method interconnects the above major steps of an IDS model and consists of two parts: The Dataset Generator and Intrusion Detector as illustrated in figure 1.

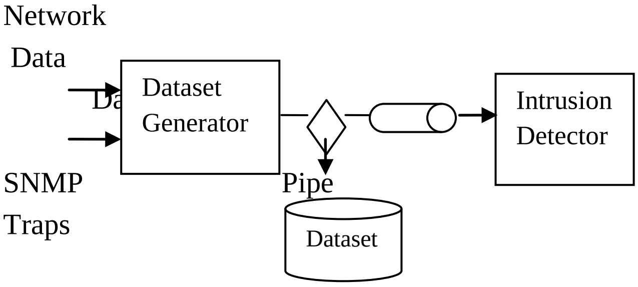

Figure 1: General overview of the comprehensive model

The constructed connection vectors from the Dataset Generator represent the link between these two parts. The output of the Dataset Generator is used as input for the Intrusion Detector. Finally, the classification of the constructed vectors delivers information about the security state of the observed computer and communication network (i.e. normal, anomaly or unknown).

### 1.Dataset Generator

The model addresses network data flows and hosts' activities effectively by using the Dataset Generator that analyzes, processes, and correlates them. The method uses a benefit of the well-known tool TCPDUMP to aggregate the online massive traffic by mirroring it on a specific port. Moreover, the method exploit the SNMP mechanism to send traps from network hosts when a certain activity is matched, e.g. by a failure login of a user. To ease the illustration, assume that the aggregator handles the dump data and traps into an appropriate time slot, e.g. every 5 seconds as figure shows. The generator analyzes and processes aggregated dump traffic and constructs corresponding network packets. Similarly, it constructs hosts' events from the received SNMP traps at the server side. It correlates the packets and traps upon ID match to construct connection vectors, based on certain features such as protocol_type, service, count, inside the queue and export them as datasets.

### 2. Intrusion Detector

Constructed connections by the generator can be exported via a pipe concept to the anomaly detector and so they can be classified as normal, suspicious, or unknown in real time. The anomaly detector is based on the enhanced growing hierarchical self organizing map (EGHSOM). Basically, the GHSOM construes high dimensional data on several layers with several maps to explore supplementary details. In the training process, the input vectors are presented to fixed number of neurons on the map using certain number of iterations. The final best matching units on the maps are called the GHSOM model. However, there are several shortcomings on the GHSOM model, which have been addressed on the work, major enhancements are:

- Meaningful map initialization
- Splitting threshold technique to boost the final topology
- Merging best matching units to robust the final GHSOM model
- Dynamic classification-threshold confidence to detect unknown attacks

The main key aspect of the EGHSOM is using a dynamic classification-confidence distance threshold to classify the incoming connection vectors via the pipe as normal, suspicious or unknown. The EGHSOM is able to update its detection model in real time and hence it is able to uncover new attacks.

The detection model classifies the connections and sends the detection result to the controller, which manages and distributes the connections.

The entire processes of EGHSOM perform the detection and update during the real time, which avoid degrading the network performance and emerging an overhead throughout the detection process. Both parts of the model operate in the online operational mode in real time. The proposed model aggregates massive data flows of large scale network, analyzes and processes them to construct connection vectors, and sends them via a pipe for classifying the security state of the network system.

## OUTCOME AND POSSIBLE RESULT

The method configured the window time slot in the dataset generator tobe 5 seconds. For the gathered packets and events, the method monitored two relevant performance metrics, which are the number of processed packets inside each window and the required time to process these packets and export them as a dataset based on the exporting phase.

In this section, the method will briefly show that the detector can effectively classify the exported traffic from the dataset generator. The detector as shown in figure1 is based on the proposed EGHSOM model which is able to classify the constructed connections from the dataset generator to three labels, normal, anomaly and unknown.

## VII. CONCLUSION

This paper presents a comprehensive model that consists of two main parts, the dataset generator and the intrusion detector. The former aggregates network packets and hosts' events, processes them and performs a correlation process to construct connection vectors using a dynamic queuing concept, and then it exports these connections in a dataset form or directly into a pipe for the detector. The detector is based on an intelligent EGHSOM model which is able to classify these connections to normal, anomaly or unknown based on a threshold margin. The model has been evaluated on a test network and on a real network firm. It could handle a large amount of packets for each window and effectively reveal anomaly connections. However, in the live test it could not reach a bit rate of 1 Gbps or even 10 Gbps to examine the plausibility, capability and scalability of the model in the high speed networks. This shortage refers to the limited number of users and small number of running services as well.

## FUTURE SCOPE

In the future, the model will be enhanced to be able to classify the unknown connections. Moreover, it will be further evaluated on a large scale computer networks up to 10 Gbps.

## REFERENCES

1] HassenSallay, Adel Ammar, Majdi Ben Saad, and Sami Bourouis," A Real Time Adaptive Intrusion Detection Alert Classifier for High Speed Networks", in proc.12th international Symposium on Network Computing and Applications IEEE 2013. PP. 73-80.

2] Maher Salem and Ulrich Buehler," Reinforcing Network Security by Converting Massive Data Flow to Continuous Connections for IDS", in proc. ICITST IEEE 2013.pp 570-575.

3] Faraz Ahmed, Faraz Ahmed, M. ZubairShafiq, and Alex X. Liu,"The Internet is For Porn: Measurement and Analysis of Online Adult Traffic", in proc. 36th ICDCS IEEE 2016.PP. 88-97.

4] Min Chen, Shigang Chen, and ZhipingCai," Counter Tree: A Scalable Counter Architecture for Per-Flow Traffic Measurement", in proc. IEEE TNET 2016.PP. 1-14.

5] Qi Zhang and Tianguang Chu," Structure Regularized Traffic Monitoring for Traffic Matrix Estimation and Anomaly Detection by Link-Load Measurements", in proc. IEEE vol.65. NO. 12. DECEMBER 2016.PP. 2797-2807.