An Efficient Network Traffic Classification Based on Unknown and Anomaly Flow Detection Mechanism

G.Suganya.M.sc.,B.Ed[1]

[1] *Mphil.Scholar, Department of Computer Science, KG College of Arts and Science,Coimbatore. Tamil nadu, India.*

Abstract— **Traffic classification technique is an important tool for network and system security in the environments such as cloud computing based environment. Modern traffic classification methods plans to take the gain of flow statistical features and machine learning methods, but the classification performance is affected by reduced supervised information, and unfamiliar applications. In addition detection of anomalies in the flow level is not considered in earlier approaches. Current work proposes Flow-level anomaly detection with the framework of Unknown Flow Detection approaches. Flow-level anomaly can be detected by using Synthetic flow-level traffic trace generation approach(SG –FLT). The two major challenges with such an approach are to characterize normal and anomalous network behavior, and to discover realistic models defining normal and anomalous traffic at the flow level. Unknown flow detection approach has been performed by Flow level propagation and finding the correlated flows to boost the classification accuracy. Performance evaluation is conducted on real-world network traffic datasets which demonstrates that the proposed scheme provides efficient performance than existing methods in the complex network environment.**

Keywords— **Traffic classification, unknown flow detection, anomaly flow detection, compound classification**

I. INTRODUCTION

Traffic classification system plays asignificantpart in and management architectures [1] and modern network security. For example, traffic classification is normally an important component in the products forintrusion detection [2] and QoS control [3]. With the advent of cloud computing the aggregate of applications organized on the Internet is rapidly increasing and several applications implement the encryption techniques. This condition makes it firmer to classify traffic flows consistent with their generation applications. Conventional traffic classification systemsdepend on on checking the exact port numbers used by dissimilar applications, or examining the applications' signature strings in the payload of IP packets. These methods encounter several problems in the current network such as user privacy protection, dynamic port numbers and data encryption. Presently, the methods incline to conduct classification by examining flow level statistical properties [4]. Considerabledevotion has been paid on the application of machine learning systems to flow statistical features based traffic classification. Still, the performance of the existing flow statistical feature based traffic classification isunconvinced in real world settings.

Severalunsupervised clustering algorithms and supervised classification algorithms have been applied to network traffic classification. From the labelled training samples of each predefined traffic class,the flow classification model is learned in supervised traffic classification [4]. This method classifieseveryflow into predefined traffic classes;consequently they cannot agree with indefinite flows produced by unknown applications. Furthermore, to attain high classification accuracy, the supervised methods wantenough labelled training data. In contrast, the clustering-based methods [5] can repeatedly group a set of unlabelled training samples and apply the clustering results to build a traffic classifier. In these methods, yet, the number of clusters has to be set huge enough to attain high-purity traffic clusters. It is a hard problem of plotting from a huge number of traffic clusters to a small number of real applications without supervised information.

The existing traffic classification methods suffer from poor performance in the critical situation where supervised information is inadequate and substantial unknown flows with anomalies are present.

Recently benign and malicious anomaliessuch asnetwork outages, worms,denial of-service attacks and flash crowds have the prospective to disturb critical servicesand infrastructures. Inspired by the remark thatdetection at the network edge is not compatible for comprisingsuch large-scale attacks, numerous anomaly detection systemsfor backbone networks have been developed. These systemswork on data collected at the flow level, meanwhile inspecting single packets is not possibleon high-speed backbone links.

Since annotation, anonymization, and modification of realtraffic traces fail in producingstandard evaluation traces,there is a strong need for another approach. We trustthat synthetic generation of standard traces has the possibleto report the three recognized problems. We visualize a synthetictrace generation scheme that yields normal flow trafficrendering a *baseline model*, and anomalous flow trafficconstructed on a diversity of *anomaly models*. Still, the challengewith generating synthetic flow-level traces is twofold: Initially,we need to describe what is considered normal and anomalousnetwork behavior, and then we need to find the suitablenormal and anomalous traffic models. We propose an eventdrivenmethod for defining normal and anomalous networkbehavior, and require the outline for anoriginal flow-basednetwork traffic model directedat anomaly detection.

II. RELATED WORK

Several supervised classification algorithms have been applied to traffic classification by taking into account various network applications and situations.

In [4]Moore and Zuevpresented a Naive Bayes estimator to classify traffic by application. Individually, current work capitalizes on hand-classified network data, utilizing it as input to a supervised Naive Bayes estimator.Theachievable of highlevel accuracyis illustrated with the Naive Bayes estimator.

In[6] Kim et al.presented ports-based Corel Reef method which host seven common statistical feature based methods and behaviour-based BLINC method using supervised algorithms on seven dissimilar traffic traces.This studygenerated several insights: (a) Theefficiency of port-based classification in recognizing legacy applications is quiet impressive, moresupported by the use of packet size and TCP flag evidence. This detailclarifies why research consideration has shifted to noticing and classifying new applications that use port camouflaged and encryption, i.e., traffic intentionally trying to avoid traffic classification. Inappropriately, growing attention to classifying traffic for determinations not essentiallyaccepted by originator of the traffic is likely to increase this group of traffic, persuading an arms race between those demanding to classify traffic, and those demanding to avoid having their traffic classified.

In [7]Lorieret al. presented grouping of traffic flows into a lesser number of clusters using the expectation maximization (EM) algorithm and physically label each cluster to an claim. The clusters are functionaland the clustering and classification algorithms specify that a worthyfit has been attained to the data. Firstexaminationspecifies that the clusters are steady over a variety of different data with the similar overall features. The prevailing clusters offeranother way to disaggregate a packet header stream and assume it to demonstrate useful in traffic analysis those emphases on a specific traffic type. For instance, simulation of TCP optimisations for high performance bulk transfer. But, additional work is essential to entirely meet our originalobjective of clustering traffic into groups that a network manager would distinguish as related to the specific application types on their network.

In [8] Bernaille *et al.*presented the*k*-means algorithm to clustering of traffics and labelledthe clusters to applications by means of a payload analysis tool. The early results detected with the method on a small trace are promising. The method is capable as it letsprimary classification of applications and is fairlymodest. Still, the method has certainlimitations that have been discussed below. Most of these limitations are easy to overwhelmed, while others are more important and affect most classification methods to date. In Multi-homed networks, large networks frequently have numerous connections to the Internet. In this case, this approachcan be extended todisplay all access links and aggregate information on a machine where the classification will occur.

In [9] Erman*et al.*presented unidirectional statistical features to simplify traffic classification in the network core. In this the problem of traffic classification is considered in the network core. Classification at the core is provoking because only incompleteinformation about the flows and their contributors is available. This problem can be addressed by developing aoutline that can classify a flow using only unidirectional flow information. This approach could be evaluated using latest packet traces that collected and pre-classified to establish a base truth.

III. PROPOSED WORK

In proposed system, Flow-level anomaly detection method can be used with the framework of Unknown Flow Detection approaches. The method of synthetic generation of flow-level traffic traces (SG –FLT) has been discussed for detection of Flow –level anomaly in network. In this method, normal and anomaly models of flows can be examined. In order to find the normal and anomaly flow, an event driven approach has been defined.

A. SYSTEM ARCHITECTURE

Initially a system model in figure 1 has been developed for finding the unknown flow in a network based on flow correlation. The anomaly detection of the flows can be detected using synthetic generation of flow-level traffic traces which is based on event driven approach. Then the detected unknown and anomaly flows can be clustered by means of k-means algorithm. At the training phase, less number of labelled traffic flows and a large number of unlabelled traffic flows are united to comprise an unsupervised training data set for traffic clustering. The characterized traffic flows are utilized to train a traffic classifier namely nearest neighbour. In the testing stage, a compound classification on the correlated flows has been performed in preference to classifying individual traffic flows. Comparative analysis can be done for the proposed system by using real time datasets.

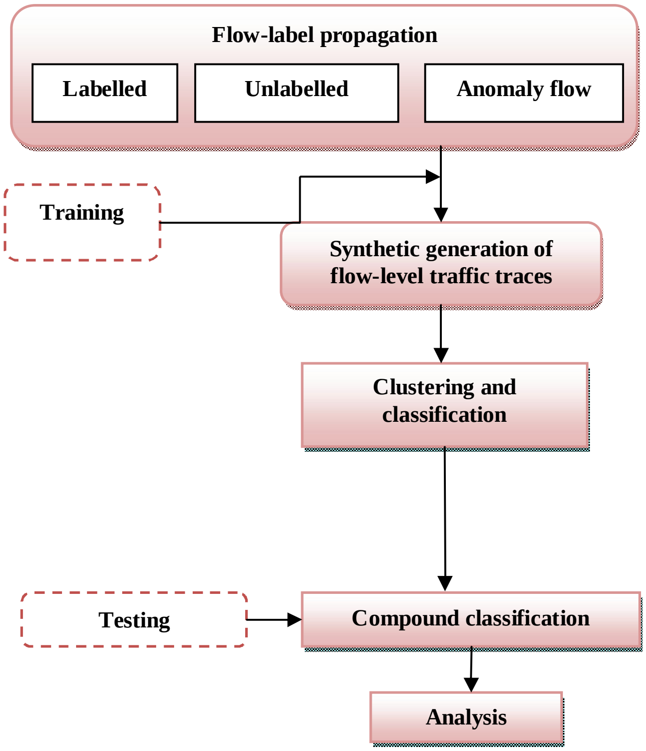


Figure 1: Proposed system achitecture

IV. METHODOLOGY

A. FLOW LABEL PROPAGATION

The proposed method aims to classify traffic flows based on the flow level statistical properties. A flow consists of successive IP packets having the same 5-tuple: {source ip, source port, destination ip, destination port, transport protocol}. Traffic flows are constructed by inspecting the headers of IP packets captured by the system on a computer network.

We start with a small set of pre-labelled traffic flows to create a supervised data set for cluster-application mapping and a training set for traffic clustering.

Suppose the labelled flow set is A = {xa1, xa2, . . .} with the labels La = {ya1, ya2, . . .}, where each flow is a real vector in the statistical feature space and the dimension of the vector is determined by the number of flow statistical properties a large set of unlabelled traffic flows = {xb1, xb2, . . .}, in the target network. Then anomaly flow can be defined as C ={ xc1,xc2,……} in the target network.. By merging the labelled, unlabelled flow sets and anomaly flows the training set T for traffic clustering can be obtained as follows

$T= A \cup B \cup C$

Moreover, an automatic process is applied to extend the labelled flow set by searching the correlated flows between A ,B and C

For each flow **x**a in A, the automatic process searches for its correlated flows in B and C with the same **3-tuple: {destination ip, destination port, transport protocol**. The following algorithm provides the propagation of flows based on 3-tuple based heuristic as follows

Input: Small flow set A and its corresponding label set La; large unlabelled flow set B; anomaly flow set C

Output: extended set of labeled flows E and its corresponding label set Le

```
1. Create output flow set E← A
2. Create output label set $L_e \leftarrow L_a$
3. for i←1 to |A| do
4 for j ←1 to |B|&& |C| do
5. Check and compare 3-tuple of $x_{ai}$,$x_{bj}$ and $x_{cj}$
6. if $x_{ai}$, $x_{bj}$ and $x_{cj}$share same 3-tuple then
7. Put $x_{bj}$ and $x_{cj}$into E;
8. Put $y_{ai}$ into $L_e$ // $y_{ai}$ is determined as the label of $x_{bj}$
9. End
10. End
11. End
```

B. SYNTHETIC GENERATION OF FLOW-LEVEL TRAFFIC TRACES(SG –FLT)

A normal and anomalous network behavior is defined; the subsequently step is to locate appropriate models that give a realistic report of normal and anomalous traffic. General parameter in traffic modeling defines that each model should be planned for a specific idea; in present case of work this is generating standard traces for flow-level anomaly detection.

Traffic Model Timescale: Anomaly detection systems of Timescale works at ranges from more than a few days to minutes. Therefore, the present model must be accomplished of relating the long-timescale behavior in addition to the short-timescale behavior of network traffic.

Modelling Characteristics of Flows: Flow traces generation involves modeling of diverse flow parameters which are regularly used for detection of anomaly: Parameters used by volume metrics such as packets and bytes, and Parameters used by spatial metrics namely source and destination addresses and ports.

Versatile and Realistic anomaly models: In order to generate benchmark traffic traces, the model must be able to generating anomalies of changing intensities, and it needs to consider the collision of an anomaly on network traffic in normal condition.

C. NEAREST CLUSTER BASED CLASSIFIER

Another In this k-means clustering has been used to construct nearest cluster based classifier. The aim of K-means clustering is to partition the traffic flows into k-clusters in order to reduce the cluster- sum as follows:

$$\underset{c}{argmin} \sum_{i=1}^{k} \sum_{X_j \in C_i} \left\| x_j - m_i \right\| \quad \rightarrow(1)$$

where **m**$_i$ denotes the centroid of C_i and it is the mean of flows in C_i.

Consider a initial set of k centroids which are selected randomly $\{m_1^0, m_2^0, \ldots \ldots \ldots . m_k^0\}$.The clustering algorithm can be done by alternating the assigning and updating stages.

During assigning phase, each cluster is assigned to the closest mean of cluster.

$$C_i^t = \{x_j : \left\| x_j - m_i^t \right\| \leq \left\| x_j - m_i^t \right\| \text{ for all } l=1,\ldots k\} \quad \rightarrow(2)$$

While in update stage, the centroid of the flows in the cluster can be chosen by calculating the new means as follows:

$$m_i^{t+1} = \frac{1}{C_i^t} \sum_{x_j \in C_i^t} x_j \quad \rightarrow(3)$$

The output of the cluster-class mapping is utilized to build a traffic classifier for each testing flows. Consider the traffic classes formed by the cluster-class mapping are denoted by $\psi = \{\omega_1. . . \omega_q\}$. The traffic classes are presented by using the results of *k*-means clustering and the flow statistical features. For class ω_i, it can be illustrated by a set of cluster centroids,

$M_i = \{m_j : C_j \epsilon \omega_i\}$

The classification rule for individual flow is as follows:

$$y= \underset{j}{argmin} \min_{m \in M_j} ||x - m|| \quad \rightarrow(4)$$

D. COMPOUND CLASSIFICATION

In the classification phase, the compound classification is applied on the correlated flows modelled instead of classifying individual traffic flows. Consider the flows are $X=\{x_1, \ldots \ldots . x_g\}$.In order to improve the classification accuracy ,the correlated information can be used. The compound classification has been developed by aggregation of flow predictions of the classifier with the weighted measures.

For a given flows $X=\{x_1, \ldots \ldots . x_g\}$, g flow predictions are$y_{x1}$,……$y_{xg}$

The flow predictions can be straightforwardly transformed into weighted measures

$$v_{ij} = \begin{cases} 1 & if\ y_{xj}\ indicates\ the\ i-th\ class \\ 0, & Otherwise \end{cases} \qquad \rightarrow(5)$$

The compound decision rule using the weighted measures is as follows:

Assign X → ω_l if $\sum_{j=1}^{g} v_{lj}$=max$_{i=1...q}$ $\sum_{j=1}^{g} v_{ij}$ →(6)

It proves that all flows in X are classified into ω_l.

V. EXPERIMENTAL RESULTS

In this section, a number of experiments have been carried out to study the impact of unknown applications and anomaly flows to the supervised classification methods. Then, the proposed method is compared with two methods such as when unknown applications are measured in the traffic classification experiments and a traffic classification using synthetic generation of flow-level traffic traces (SG -FLT) .`

Two common metrics are used to measure the classification performance in anomaly with un known application scenario and unknown application framework.

Overall accuracy is defined as the ratio of the amount of all correctly classified flows to the sum of all testing flows

Accuracy =$\frac{number\ of\ correctly\ classified\ flows}{number\ of\ testing\ flows}$

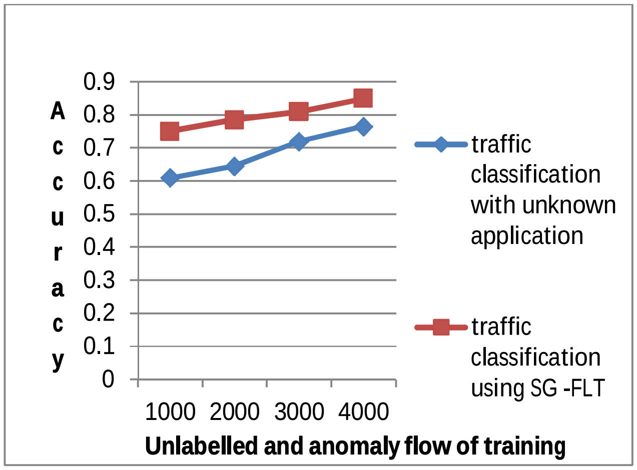
F-measure is calculated by

F-measure =$\frac{2 \times precision \times recall}{precision + recall}$

Where precision is defined as the ratio of correctly classified flows over all predicted flows in a class (unknown +anomaly flows) and recall is defined as the ratio of correctly classified flows over all ground truth flows in a specified class.

The comparative graph for the classification of unknown flows and classification of anomaly with unknown class is illustrated below:

Figure 2: Accuracy comparision graph



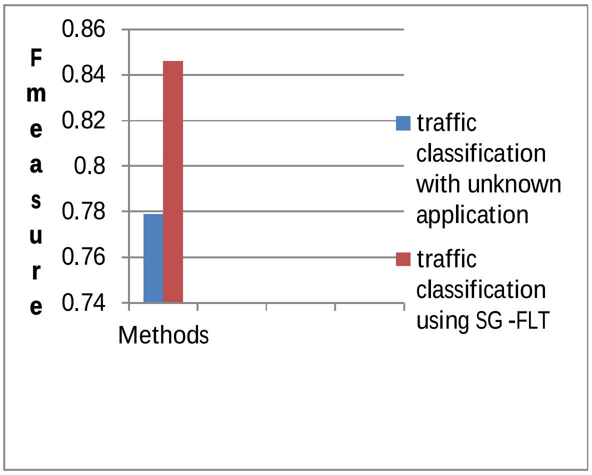


Figure 3: F-measure comparision graph

Thus the above graph in figure 2 and 3 shows that proposed system of traffic classification using synthetic generation of flow-level traffic traces (SG -FLT) provides higher accuracy and F-measure when compared with existing method of private traffic classification with unknown application.

VI. CONCLUSION

In the present work, traffic classification in network has been done by based on detection of flow-level anomaly detection by using the synthetic trace generation approach. Synthetic trace generation approach has been proposed in which it has the potential of making standard traffic traces available to a broad community. A reference model is created which allows defining the normal and anomalous network behavior. The present method introduced techniques to adequately utilize flow correlation information. Firstly flow label propagation is utilized which can automatically accurately label more unlabelled flows with anomaly flows to enhance the ability of nearest cluster based classifier (NCC). Another method is compound classification which can unite a number of flow predictions to make more accurate classification of weighted flows. Experimental result in terms of accuracy and F-measure provides better result when compare with the existing system.

REFERENCES

[1]. T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multileveltraffic classification in the dark," SIGCOMM Comput. Commun. Rev.,vol. 35, pp. 229–240, Aug. 2005.

[2]. Y. Xiang, W. Zhou, and M. Guo, "Flexible deterministic packet marking:an IP traceback system to find the real source of attacks," IEEE Trans.Parallel Distrib. Syst., vol. 20, no. 4, pp. 567–580, Apr. 2009.

[3]. M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-servicemapping for QoS: a statistical signature-based approach to IP trafficclassification," in Proc. 2004 ACM SIGCOMM Conference on InternetMeasurement, pp. 135–148.

[4]. A. W. Moore and D. Zuev, "Internet traffic classification using Bayesiananalysis techniques," SIGMETRICS Perform. Eval. Rev., vol. 33, pp. 50–60, June 2005

[5]. A. McGregor, M. Hall, P. Lorier, and J. Brunskill, “Flow clusteringusing machine learning techniques,” in Proc. 2004 Passive and ActiveMeasurement Workshop, pp. 205–214.

[6]. H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee,“Internet traffic classification demystified: myths, caveats, and the bestpractices,” in Proc. 2008 ACM CoNEXT Conference, pp. 1–12.

[7]. Lorier, McGregor, M. Hall, P., and J. Brunskill, “Flow clusteringusing machine learning techniques,” in Proc. 2004 Passive and ActiveMeasurement Workshop, pp. 205–214.

[8]. L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian,“Traffic classification on the fly,” SIGCOMM Comput. Commun. Rev.,vol. 36, pp. 23–26, Apr. 2006.

[9]. J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, “Identifying anddiscriminating between web and peer-to-peer traffic in the network core,”in Proc. 2007 International Conference on World Wide Web, pp. 883–892.

.

.