TRACING OF VOIP TRAFFIC IN THE RAPID FLOW INTERNET BACKBONE

A.Jenefa[1], Blessy Selvam[2]

[1]*Teaching Fellow, Computer Science and Engineering, Anna University (BIT campus), TamilNadu, India*
[2]*Teaching Fellow, Computer Science and Engineering, Anna University (BIT campus), TamilNadu, India*

Abstract

VoIP traffic application gaining a terrific admiration in the recent couple of years. VoIP Traffic Classification has concerned for network management and it comes to be more complicating because of modern applications behaviors and it has attracted the research community to develop and propose various classification techniques which don't depend on 'well known UDP or TCP port numbers. To overcome the problem of unknown flow classification and achieve effective network classification, a new innovative novel work called Multi Stage Fine-Grained classifier is proposed in this research for classifying the VoIP traffic flow with high accurate classification. The datasets of VoIP network traffic measurements taken from our campus WI-FI and the experimental results shows that the proposed work outstrips the existing approaches in the Rapid flow Internet Backbone. Without investigate the packet payloads, our proposed Fine-Grained classifier effectively classifies the Peer-to-Peer encrypted traffic in the real time network. Our experimental results shows high accuracy and small error rate in classifying the Peer-to-Peer network traffic.

Keywords: *Multi Stage Fine-Grained Classifier, Rapid VoIP traffic Flow (SKYPE, VoIP, GAMING, Other) classification, Machine Learning*

----------***----------

1. INTRODUCTION

Popular Network VoIP traffic classification becomes more puzzling because recent applications periodically changing its network behaviours. VoIP Traffic classification has enlarged in significance this decade, as it is now transferring not only file, text, image and mail but also helpful to watch TV shows, streaming, play online games and to communicate using VoIP and Skype. Network traffic classification is helpful for service disparity, network security, quality of service and research purpose. Real time traffic classification has the ability to solve challengeable network management problems for Internet Service Providers (ISP's). The Payload-Based Traffic classification, Port Based Traffic classification and Machine learning are the three prevailing methods in the arena of network traffic classification. In the early framework, Port-based method was broadly used and it's nominal for old-fashioned applications which often use pre-defined port assigned by IANA [1] .The Increase of new applications that has no IANA registered ports, but instead the new applications uses port numbers already registered to masquerade their traffic and avoid filtering or firewalls. As new application design and user behaviour declared port-based traffic flow classification untrustworthy and payload based approaches which inspect packet content to identify byte strings associated with an application, or performs more complicated signature matching methodology. The Payload based method identifies network flow traffic by penetrating the packet payload for signatures of well-known applications. Still, Payload packet inspection methods have numerous limitations. First of all it can only identify traffic for which signatures are known and are unable to classify any other traffic. Maintaining an up-to-date list of signatures is hazardous. Finally, the Payload Based traffic classification fails if the payloads are encrypted. Generally Peer-to-Peer network traffic payloads are encrypted (*i.e. Skype streams* [2] [4] [10]). Classification of network traffic using port-based or payload-based identification approaches has been completely weakened in recent days.

So it's upright to move towards Machine Learning based packet classification in which statistical characteristics of IP Packet flows are involved. The Machine Learning (ML) based approach [11] [19] uses the statistical attributes of traffic flow so that it can avoid the limitations of Payload and Port Based methods.

In this paper, we proposed a innovative Machine Learning approach called *Multi stage Fine-Grained Classifier* to classify the Rapid VoIP traffic, Skype traffic and other kind of traffic generated from the University Internet Backbone.

VoIP allows voice communication over IP phones. Normally, an audio codec is used to convert the analog voice signals to digital signals, so that the voice payload size depends on the audio codec.Skype allows everyone to make both video and audio calls over the Internet. Skype uses Sinusoidal Voice over Packet Coder (SVOPC) [14] [15] [21] [22] [23] for the conversion of Digital signals.

Our attention in this section is inspired chiefly by lawful interception obligations [5][6][7][8][9].Governments around the world call for telecommunication companies and obliged to offer facilities and services to ensure that they can intercept their systems and services[7] of network traffic.

The VoIP data set comprises encrypted P2P connections or other types of encrypted traffic. So traffic identification boons great challenge to lawful interceptions. Lawful Interception practice generally comprises of a law enforcement agency delivering a warrant instructing the Internet Service Provider to provide information about the clients connected in calls.

As mentioned, the Supervised Machine Learning approach [16] is suitable for classifying the VoIP traffic which analyses only the commonly used statistical flow features such as Flow-Duration, Packet Size, Inter-Arrival Gap, No of bytes/packets transferred ,3-Tuple Information and etc. Our goal is to build an efficient and accurate classification for VoIP traffic using Machine Learning techniques

This research paper is structured as follows: Section 2 gives information about VoIP characterization and other kind of encrypted traffic. Section 3, 4, 5 describes the System model of VoIP identification used in this research. Section 4 presents the experimental results produced and finally conclusion is given in session 5.

2. VOIP CHARACTERIZATION

In spite of secrecy of VoIP inter-nals, investigations have permitted a number of its statistical flow characteristics to be identified [3] [4] [12] [20]. The following figure 1 shows that the different application of VoIP uses the same port to forward the flow packets. So to characterize the VoIP traffic among different application in the network, the statistical features of IP packets are concerned [13] in this research work to identify the application flow.

Fig 1: Different Applications of VoIP Identification

3. SYSTEM MODEL OF VOIP IDENTIFICATION

Supervised Machine Learning approach demands an aforementioned knowledge to classify the Peer-to-Peer traffic flows. The phases of Supervised Machine Learning approaches are

- *Training Phase:* The Training phase that builds a set of classification model or rules.
- *Identification Phase:* The model that has been constructed in the Training phase is used to classify real time VoIP traffic.

The Machine Learning involves mainly two steps. First, extensive features are defined based on statistical characteristics of application protocols such as flow duration, inter-arrival times, packet length etc. A Fine Grained Machine learning classifier is then taught to associate set of statistical features with known traffic classes, and apply the well-trained machine learning classifier to classify the VoIP traffic using previously trained rules. The Multi Stage Fine Grained Classifier is also suitable to identify encrypted protocols.

The Fine-Grained Classifier (FGC) algorithm is used to classify flows into application behaviours based on flow classification based on packet size cumulative and flow similarity grouping. The flow is classified into applications by packet size distribution and then the flows are grouped as sessions by 5-tuple information. The Fine-Grained Classifier (FGC) algorithm classifies the traffic without examining the packet payloads. This method works even if the packet payloads are encrypted. So this Fine-Grained Classifier (FGC) algorithm is sufficed to classify the encrypted Peer-to-Peer traffic.

4. SUPERVISED TRAINING PHASE

The Training phase is managed offline to classify the encrypted Peer-to-Peer traffic application in the Internetwork Backbone. So it's supportive to spot out a VoIP application from a mix of applications. The objective of the offline training phase is to find out the statistical flow information that should be unique to or different from other applications, to be the origin for comparison. For this reason, this Training phase first collects a set of VoIP traffic traces and stabs to abstract the statistical flow based information from the real time traffic traces.

A traffic filter is very much useful to collect the traffic traces from the network since the filter sieves out immaterial information from the traffic collection. Our experimental research uses a one day trace collected at the edge of our Anna university backbone network. Any Traffic monitoring analyser can be used to extract the flows. Initially, to group flows into clusters, the degree of similarity between flows is to be measured.

4.1 Method 1 (M1) Using Similarity Distance Calculation

By Euclidean distance, the resemblances between the flows are measured; a small distance between the two packets shows a strong similarity (whereas) a large distance shows a minimum similarity. In between two flows, the similarity distance can be calculated as

$$dist(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where n is the number of flows collected from the Internet Backbone. If the distance between two flows is very trivial then the two flows would be grouped together or otherwise they are grouped as dissimilar groups. The flow having slighter value distance will be grouped together named as $S_1, S_2 \ldots\ldots S_n$.The above process is repetitive till all flows get clustered.

4.2 Method 2 (M2) Average Packet Size Calculation

The average packet size calculation is determined by finding out the centroid point of the packet sizes distribution.

This technique permits to classify the flows once it has to be allotted to a cluster. The average packet sizes for each VoIP application is computed and gained knowledge for each application which will be helpful to identify the application in the real time phase.

$$avg\_pkt\_size_{S_i} = \sum_{j=1}^{n} \frac{pk_j}{n}$$

Whereas pk_j be the frequently used packet sizes and 'n' be the number of flows in S_i .The above progression is repeated for each application to calculate the average packet sizes for classification process. At extreme all the applications of VoIP traffic shows the unique behaviour concerning patterns of packet length. Thus, dissimilar packet sizes are useful to discriminate certain VoIP applications. Generally, the packets have the similar sizes across all flows but it's necessary to inspect that the average packet size per flow remains constant across all flows in the network traffic.

//Extracting the features from the VoIP-Traffic Flow
//**Input:** Set of flow Packets with different correlation $\{x_1, x_2 ..... x_n\}$ $\{y_1, y_2 ..... y_n\}$
//**Output:** Correlating the same group of flows by similarity calculation and Tuple information

For i=1 to n // executing till EOF

$$\text{SimilarityDistance}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Check and Group the correlated Packet Flows by Correlated three-tuple of F_a and F_b

For each ($x_i \in F_a$ or $y_i \in F_a$)

Grouping x_i and $y_i \in F_a$ and Label the Flow as L_a

End

For each ($x_i \in F_b$ or $y_i \in F_b$)

Grouping x_i and $y_i \in F_b$ and Label the Flow as L_b

End

For each ($x_i \notin F_a$ and $y_i \notin F_b$)

Grouping the Flow Label as $M_a = \{x_i\}$ and $M_b = \{y_i\}$

// A new Traffic Application determined

End

End

Algorithm 1: Similarity Flow Label Group in VoIP Traffic Application

5. SUPERVISED IDENTIFICATION PHASE

The supervised clustering of identification phase which contains dissimilar heuristics to improve the classification of encrypted Peer-to-Peer traffic. The Supervised classification technique is useful for accurate classification of VoIP traffic. A set of experimental examination has reviewed in our traffic traces to examine numerous VoIP applications in the network backbone. The following first three techniques are used to group the similar flows and the last method is used for flow classification of VoIP traffic. Fig 2 shows the system model of identification of VoIP traffic in the network.

5.1 Method 1 (H1): The Transport Level Grouping

To differentiate the VoIP applications in the network backbone, the transport level protocol grouping is helpful to label the applications into different clusters: i) Transmission Control Protocol (TCP) is used by web, chat, mail, and FTP. ii) User Datagram Protocol (UDP) is generally used by Network Management Traffic and games. iii) Encrypted Peer-to Peer traffic and streaming traffic uses either TCP or UDP. Thus, the technique H1 supports to accumulate the similar applications for as a minimum to certain extent. Table 1 shows the grouping of Applications using Transport Layer Protocol.

Table-1: Classification of Applications using the Transport Level Grouping

Traffic Identification	**Application Identification**	**Transport Layer Protocol Used**
Chat	MSN Messenger, Yahoo Messenger,AIM,IRC	TCP
ftp(Data)	ftp,databases	TCP
Web	http, https	TCP
Mail	smtp,pop,nntp,imap,identd	TCP
p2p	BitTorrent,eDonkey,Gnutella, Pee Enabler, WinMX, OpenNap, MP2P, FastTrack, Direct Connect	TCP/UDP
Attack	Port scans,IP address scans	-----
Streaming	mms(wmp),real,quicktime,shoutcast, Vbrick streaming.	TCP/UDP

5.2 Method 2 (H2) The Correlated Bag of Flow sets (5-Tuple Information)

The Correlated flow sets such as by port number and IPs, the flows are grouped together. Generally, the operating system assigns successive port numbers for related flows. Figure 4 shows how the operating system allocates the successive port numbers for each individual flow of packets. Here in this research work we used (Source_IP, Destination_IP, Inter-arrival Time, Port number and Flow ID) to cluster similar behaviours. If the Source_IP address and the Destination_IP address of two dissimilar flows are same with the similar port number, then the flows will be grouped together. But all flows having same port number does not belong to the same flow. So the Inter-arrival time (difference of each separate arrival time of different flows) is considered to group the similar flows in the traffic. If the arrival time of any two flows is within a threshold value, then it will be grouped together or consider as different flow. Table 2 shows the grouping of similar flows using 5-Tuple information.

5.3 Method 3 (H3) The Similarity Distance Calculation

The similarity distance of IP traffic between the flows is calculated by Euclidean distance. Euclidean distance is calculated to identify the similarity relation between two different set of flows in the traffic. It's helpful to find out the identical flows in the traffic. If the distance between the flows is within a specified threshold value, then it will be clustered together or else it will be grouped as different cluster.

5.4 Method 4 (H4) Flow Identification

After grouping identical flows, the average packet size of each similar flow is calculated and then the mean value obtained is compared with the knowledge gained for each VoIP traffic to choose which application it should be. To identify the VoIP applications, the unknown flows are compared one after another. Our proposed system works well even if the packet payloads are encrypted. So, this technique is helpful to identify the encrypted Peer-to-Peer traffic. These techniques work very well, if the average packet size remains constant across the flows in the network backbone. Figure 4 show that the different VoIP applications have different packet sizes. So it's helpful for our research to effectively identify the application based on its unique behaviour of VoIP applications. The following algorithm describes the Fine-Grained Classification of packet size distribution for each Peer-to-Peer encrypted VoIP application traffic network.

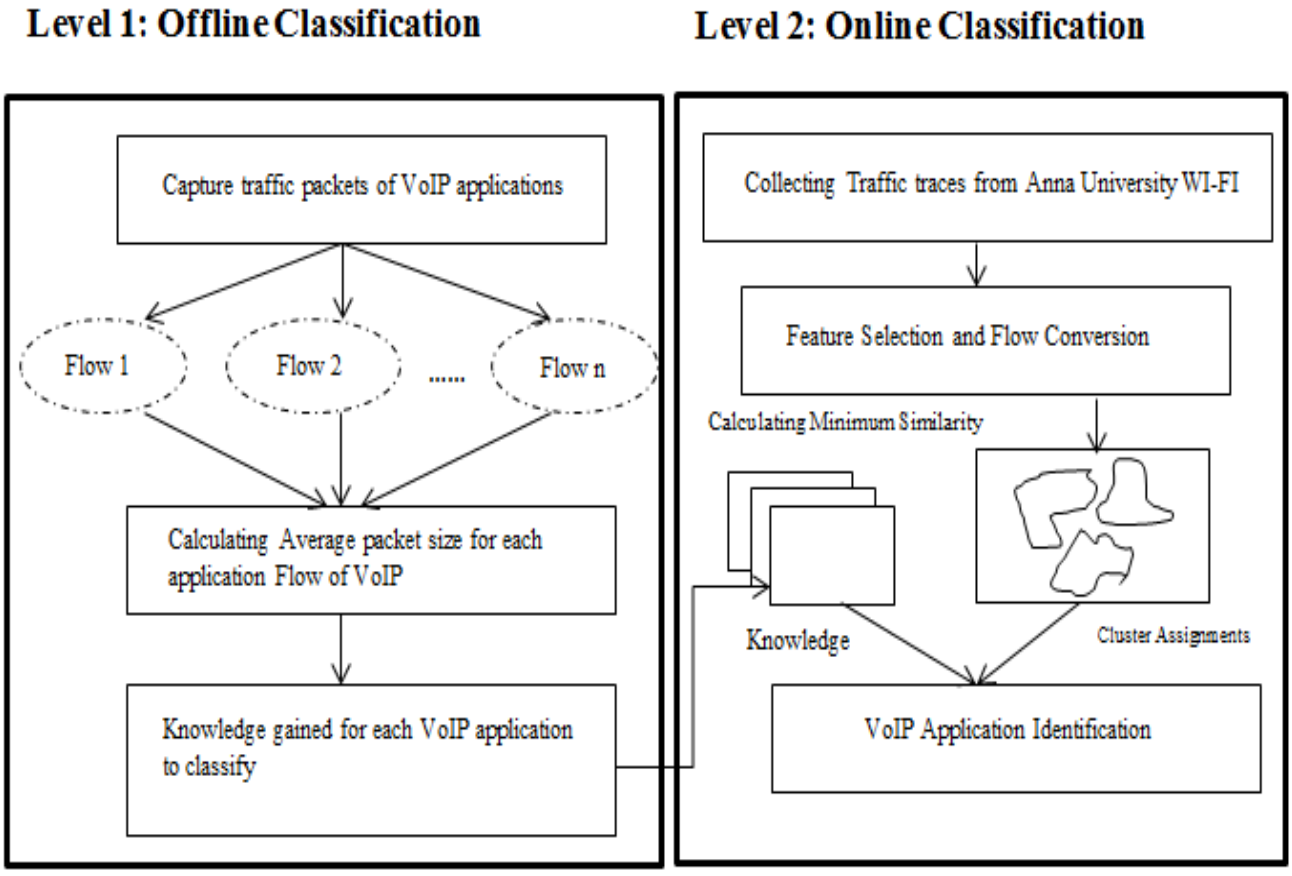


Fig 2: System Model of VoIP Identification

//Packet Size Calculation for the VoIP-Traffic Flow

//**Input:** Set of flow Packets with Packet Sizes $\{P_1, P_2 .....P_n\} \in F_a$ and $\{Q_1, Q_2 .....Q_n\} \in F_b$

//**Output:** Classification of VoIP traffic flow based on different Packet Sizes.

For each i=1 to n

Grouping each Similarity Flow as $X_i \leftarrow P_i \parallel Y_i \leftarrow Q_i$

End

For each i=1 to n // calculating the mean for each VoIP Application

// Average Packet Size Calculation for each VoIP Traffic Application

$$\text{Average\_Packet\_size} = \sum_{i=1}^{n} \left(\frac{X_i}{n}\right) \parallel \sum_{i=1}^{n} \left(\frac{Y_i}{n}\right)$$

End

Algorithm 2: Fine-Grained Classifier

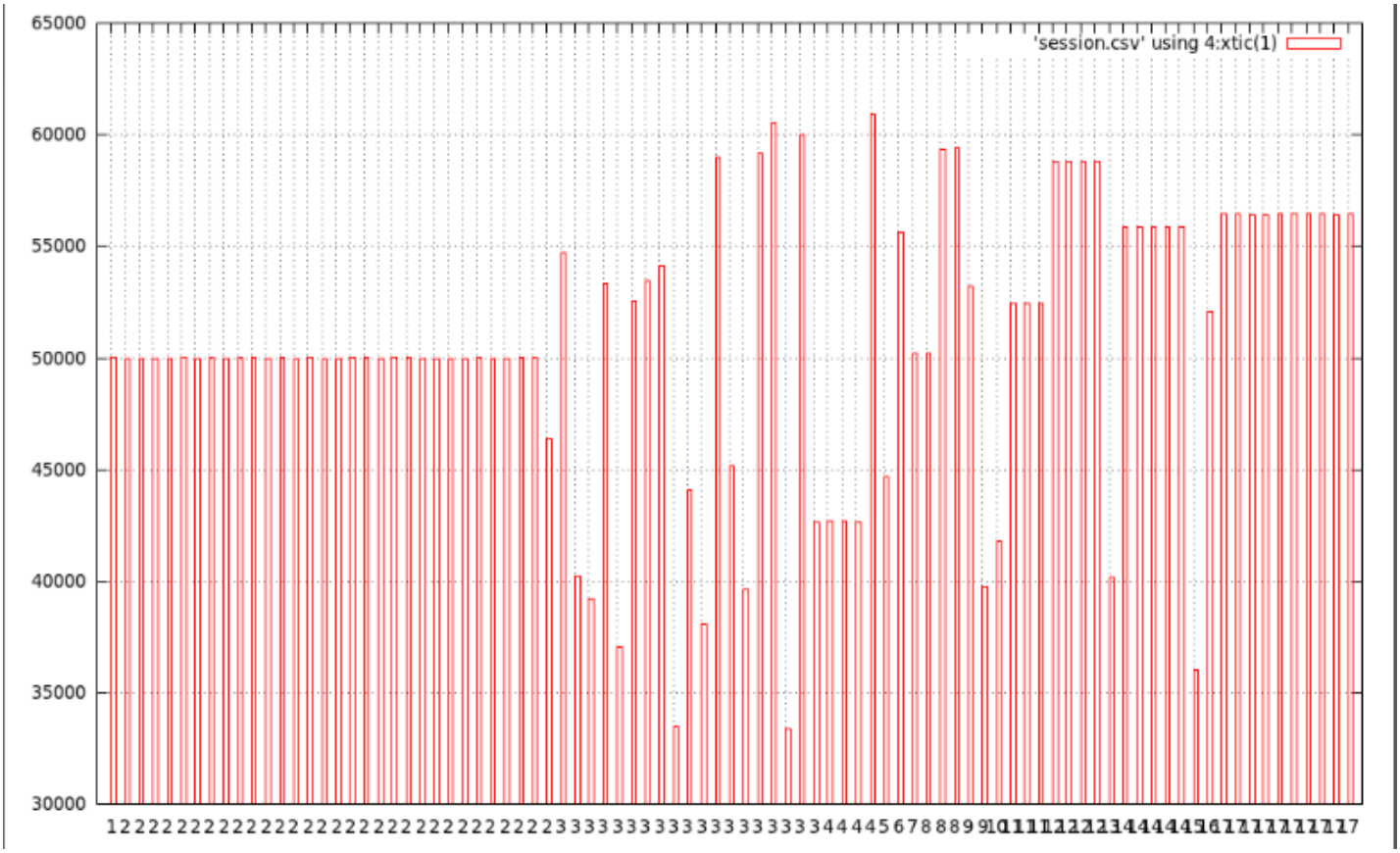
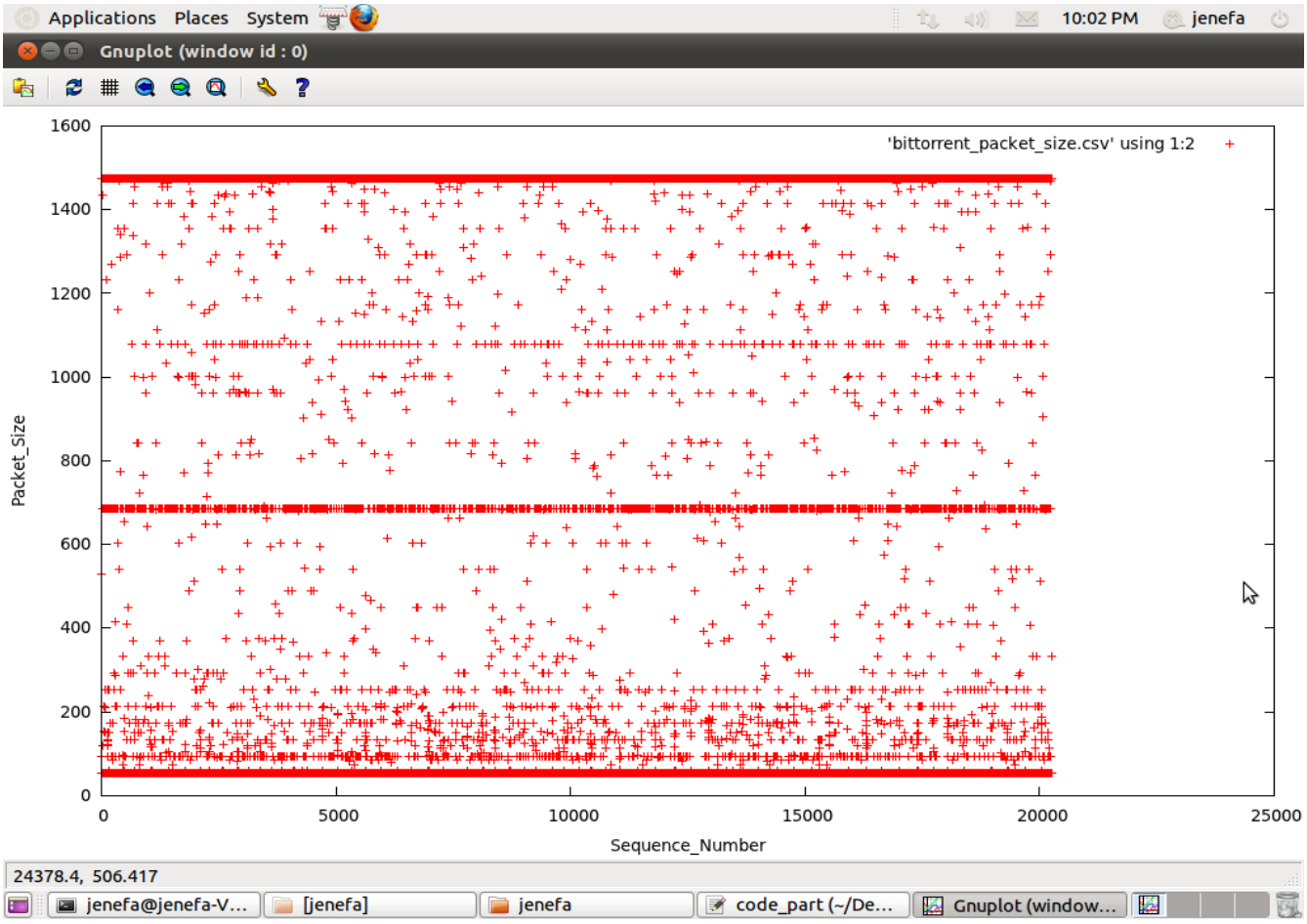


Fig 3: Similarity Flow Grouping of IP Packets

Table- 2: Classification of VoIP Applications using 5-Tuple information

Source IP_ address	Destination IP_ address	Port Number used	Grouping of Flows (Flow_id)
10.0.0.1	192.168.2.51	43321	X
10.0.0.1	192.168.2.51	43321	X
10.0.0.1	192.168.2.51	43321	X
10.0.0.1	192.168.2.51	43321	X+1
10.0.0.1	192.168.2.51	43321	X+1
10.0.0.1	192.168.2.51	43321	X+1
10.0.0.1	74.125.236.181	54467	Y



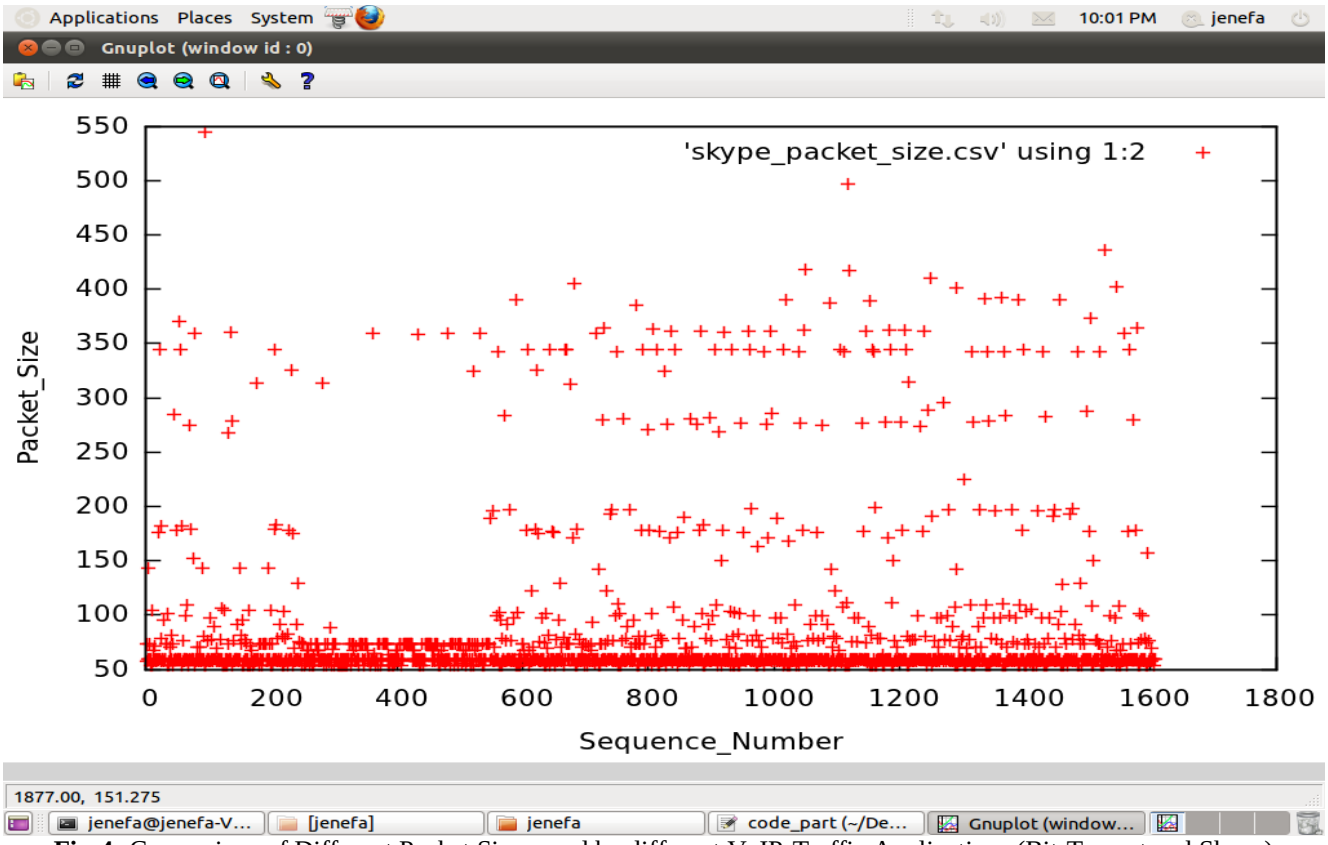


Fig-4: Comparison of Different Packet Sizes used by different VoIP-Traffic Applications (Bit-Torrent and Skype)

6. IMPLEMENTATION DETAILS

Our purpose of this research is to produce an efficient classification algorithm for VoIP traffic identification. Generally, the Supervised and the unsupervised algorithms of Machine Learning are used to resolve the difficulties in network traffic classification. Here an innovative supervised algorithm named as Multi Stage Fine-Grained classifier is used to classify the network traffic. Initially the Traffic filters; TCPDump [17] or Wireshark or Net Mate [18] is used for traffic filtration to prevent irrelevant traffic. By implementing our own proposed supervised learning classification algorithm, effectively classify the application based on its unique behaviour of encrypted Peer-to-Peer traffic. Table 3 and Figure 5 shows the experimental tests taken at our Anna University campus.

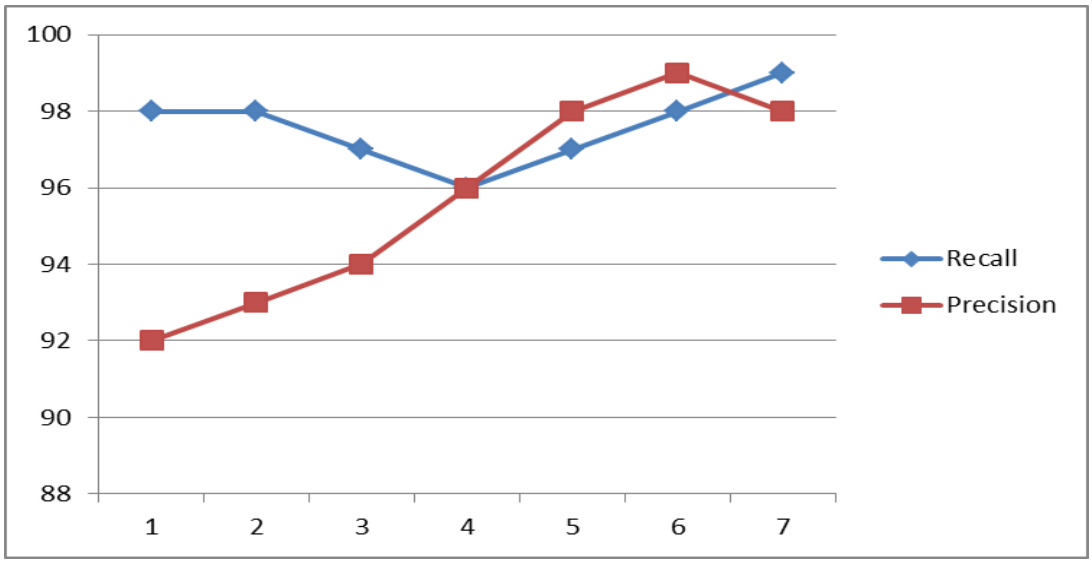


Fig 5: Recall and Precision of VoIP traffic

Table-3: Peer-to-Peer Classification Results

Packet Counts	Gaming	VoIP	Skype	Recall and Precision
<10	83.34	73.45	76.23	78%
<20	89.72	82.32	78.78	82%
<30	95.43	85.12	85.92	86%
<40	96.29	86.31	88.14	89%
<50	97.89	94.23	92.12	96%
<60	99.99	97.65	98.76	97%
Final	100%	98%	99%	99%

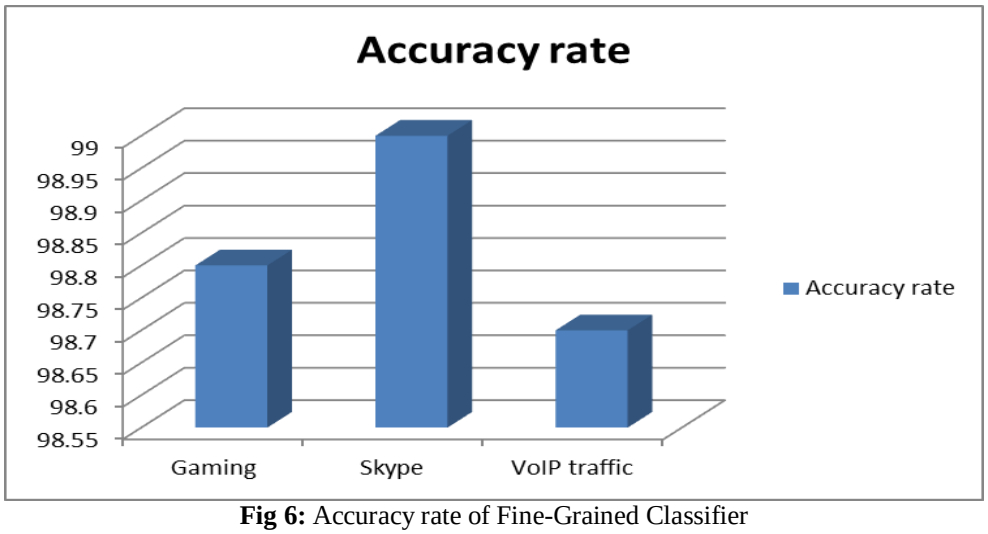


Fig 6: Accuracy rate of Fine-Grained Classifier

7. CONCLUSION

This research work emerges on classification of encrypted Peer-to-Peer traffic in two phases: an offline Training phase and an online Application Identification phase. The offline Training phase finds out the similar characteristics of VoIP application in the rapid real time network. In our research, the Fine-Grained classifier algorithm is proposed to classify the real-time encrypted traffic with the help of gained-knowledge from the offline training phase without investigate the packets payloads. With the gained knowledge database for each VoIP application, the Multi Stage Fine Grained Classifier can effectively identify the VoIP application that session of flows fit into. The Proposed research work effectively classifies the encrypted Peer-to-Peer traffic with high accuracy rates and small error ratio. The accuracy rate of proposed online classifier, an average of 98.79% is achieved in the internet backbone.

ACKNOWLEDGEMENTS

This research paper is made possible through the help and support from everyone. First and foremost, I would like to acknowledge and extend my heartfelt gratitude to my PhD guide Mr. BalaSingh Moses who has helped me to complete this research possible. Second, I would like to thank my department colic's, for their vital encouragement and support. Most especially to god, who made all things possible. At last, the authors would like to thank our Anna University Campus admin for allowing us to monitor the VoIP traffic

REFERENCES

[1] IANA, "Internet Assigned Numbers Authority", http://www.iana.org/assignment/port-numbers.

[2] Skype web site, http://www.skype.com

[3] D.Bonfiglio, M.Mellia, M.Meo, D.Rossi, P.Tofanelli," Revealing Skype Traffic: when randomness plays with you", ACM Sigcomm'07, Kyoto, Japan, Aug.2006.

[4] "International carriers, traffic grows despite Skype popularity", Telegeography Report and Database, available on line http://www.telegeography.com/,Dec 2006.

[5] CALEA Online, http://www.calea.org, accessed 13 February 2009.

[6] Upson, S.2007 Wiretapping Woes, IEEE Spectrum, May 2007.

[7] Maloku, N., Aljaz, T., Dolenc, F. 2003 Legal Call Interception in Next Generation Networks, Proceedings of the 7th International Conference on Telecommunications.

[8] Baker, F., Foster, B., Sharp, C. 2004 Internet Engineering Task Force, CISCO Architecture for lawful Intercept in IP Networks, http://www.ietf .org/rfc/rfc3924.txt, Accessed 13 February 2009.

[9] Bellovin, S., Blaze, M., Bricell, E., Brooks, C., Cerf, V., Diffie, W., Landu, S., Peterson, J., Treicher, J.2006 Security Implications of Applying Communications Assistance to Law Enforcement Act to Voice over IP, Information Technology Association of America.

[10] P.Biondi, F. Desclaux, "Silver Needle in the Skype." Black Hat Europe '06, Amsterdam, the Netherlands, Mar.2006.

[11] Riyad Alshammari, A.Nur Zincir-Heywood, "Identification of VoIP encrypted traffic using a machine learning approach",Journal of King Saud University-Computer and Information Sciences'15.

[12] Bongfiglio, D., Mellia, M., Meo, M., Rossi, D. 2009 Detailed Analysis of Skype traffic, IEEE Transactions on Multimedia, vol 11, no1, pp. 117-127.

[13] Lam Hoang Do, Philip Branch, "Real Time VoIP Traffic Classification", CAIA Technical Report 090914A, pp 1-3, July 2009.

[14] Dario Rossi, Marco Mellia, Michela Meo, " A Detailed Measurement of Skype Network Traffic"- A Review Report, 2009

[15] Suh, k., Figuiredo, D., Kurose, J., Towsley, D., 2006 Characterizing and Detecting Skype Relayed Traffic, IEEE INFOCOM '06.

[16] Nguyen, T. and Armitage, G. 2008 A Survey of Techniques for Internet Traffic Classification using Machine Learning, IEEE Communications Surveys & Tutorials, vol.10 no.4.

[17] Tcpdump, http://www.tcpdump.org, Accessed 13 February 2009.

[18] Netmate,http://www.ip-measurement.org/tools/netmate, Accessed 13 February 2009.

[19] R, The R Project for Statistical Computing, http://www.r-project.org, Accessed 13 February 2009.

[20] S.A.Baset and H. Schulzrinne, "An analysis of the Skype peer to peer internet telephony protocol," in IEEE Infocom '06, Barcelona, Spain Apr.2006.

[21] A.Guha, N.Daswani, and R.Jain, "Anexperimental study of the Skype peer-to-peer VoIP system," in 5th Int. Workshop on Peer-to-Peer Systems, Santa Barbara, CA, Feb 2006.

[22] L.DDe Cicco, S.Mascolo and V.Pailmisano, "Skype video responsiveness to bandwidth variations," in ACM NOSSDAV'08, Braunschweig, Germany, May 2008.

[23] J.Lindblom, " A Sinusoidal voice over packet coder tailored for the frame-erasure channel," IEEE Trans. Speech and Audio Processing, vol. 13, no.5, pt.2, pp.787-798, Sep.2005

BIOGRAPHIES

A.Jenefa is a Teaching fellow in Anna university, Trichy and she received her M.Tech in Computer Science and Engineering from Karunya University. Her research interest includes networking field with focus on Internet traffic Monitoring and Analysis, Network Management and Network Security.

Blessy Selvam is a Teaching fellow in Anna university, Trichy and she received her M.E in Computer Science and Engineering from Anna University. Her research interest includes networking field with focus on Internet traffic Monitoring and Analysis, Mobile Mining and Cognitive science