ANALYSIS AND IDENTIFICATION OF VOIP TRAFFIC USING STATISTICAL TECHNIQUE

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

ANJALI

11609860

Supervisor

MAX BHATIA

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Month…… Year ……

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0310

Supervisor Name : Max Bhatia **UID :** 16870 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Anjali	11609860	2016	K1637	8059288600

SPECIALIZATION AREA : Networking and Security **Supervisor Signature:** _____

PROPOSED TOPIC : Analysis and identification of VoIP traffic using statistical technique

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.80
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.40
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.60
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.40
5	Social Applicability: Project work intends to solve a practical problem.	7.20
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.20

PAC Committee Members		
PAC Member 1 Name: Prateek Agrawal	UID: 13714	Recommended (Y/N): Yes
PAC Member 2 Name: Deepak Prashar	UID: 13897	Recommended (Y/N): Yes
PAC Member 3 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 4 Name: Pushpendra Kumar Pateriya	UID: 14623	Recommended (Y/N): Yes
PAC Member 5 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 6 Name: Aditya Khamparia	UID: 17862	Recommended (Y/N): Yes
PAC Member 7 Name: Anupinder Singh	UID: 19385	Recommended (Y/N): NA
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Analysis and identification of VoIP traffic using statistical technique**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

ABSTRACT

Arranged information contain interconnected elements for which derivations are to be made. For instance, website pages are interconnected by hyperlinks, inquire about papers are related by references, telephone accounts are connected by calls, possible fear mongers are connected by correspondences. Networks have turned out to be ubiquitous. Correspondence systems, monetary exchange systems, systems depicting physical frameworks, and Informal communities are on the whole winding up observably dynamically imperative in our regular daily existence. Frequently, we are keen on models of how hubs in the framework impact each other (for instance, who pollutes whom in an epidemiological framework), models for anticipating a quality of interest in light of watched characteristics of articles in the framework. The technique of SVM is applied which will classify the data into malicious and non-malicious.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled

"ANALYSIS AND IDENTIFICATION OF VOIP TRAFFIC USING STATISTICAL TECHNIQUE" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Max Bhatia. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

ANJALI

11609860

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled "**ANALYSIS AND IDENTIFICATION OF VOIP TRAFFIC USING STATISTICAL TECHNIQUE"** submitted by **Anjali** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Max Bhatia

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOLOGEMENT

I would like to convey my most heartfelt and sincere gratitude to my mentor Mr. Max Bhatia of Lovely Professional University, for his valuable guidance and advice. His willingness to motivate me contributed tremendously to achieve the goal successfully.

I would also like to thanks dear God and my parents who has always inspired and encouraged me to achieve my goal successfully.

ANJALI

TABLE OF CONTENTS

TABLE OF CONTENTS

Chapter-1
Introduction

1.1 Network

The connection of more than one computer systems that provide benefits to each other is considered as a network. The computers connected to communicate and provide exchange of information to each other. The collections of computer devices that facilitate communication amongst each other are gathered here within this setup. The scenario in which numerous computers are gathered and connected with each other to exchange information and provide facilities to other resources is called a network. The information such as data communication is provided with the help of networking technology. There are software and hardware types of resources present within the sharing devices.

The numbers of protocols are utilized for the purpose of organizing the traffic of the network with respect to its size, topology and the content of organization for communication.

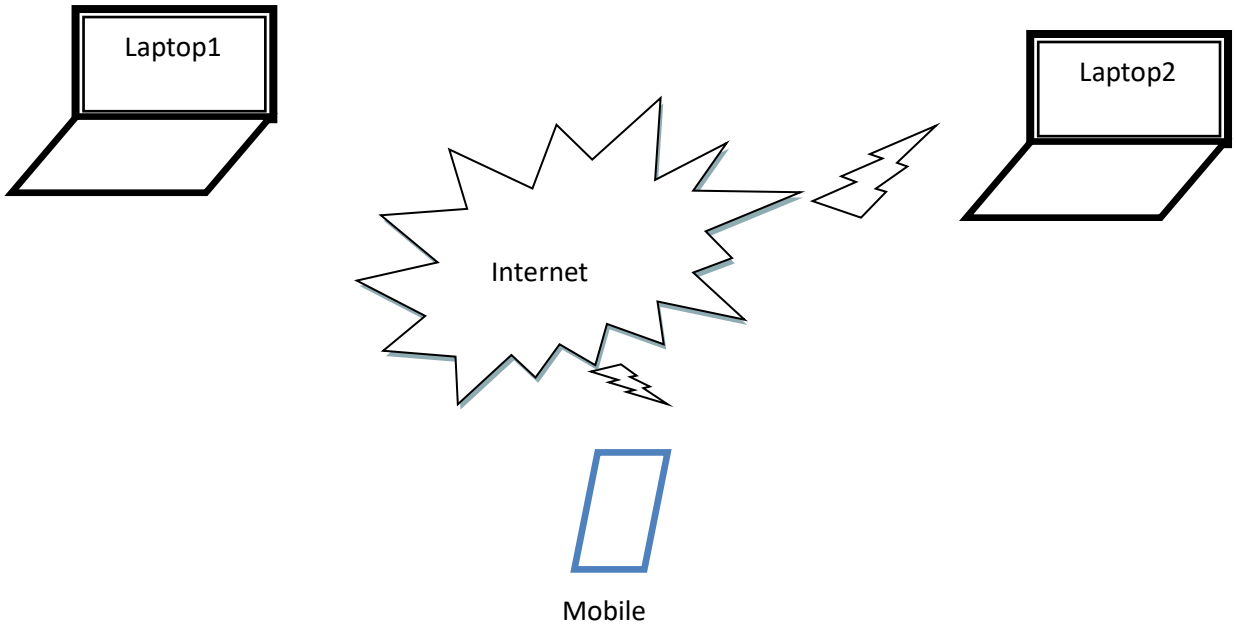
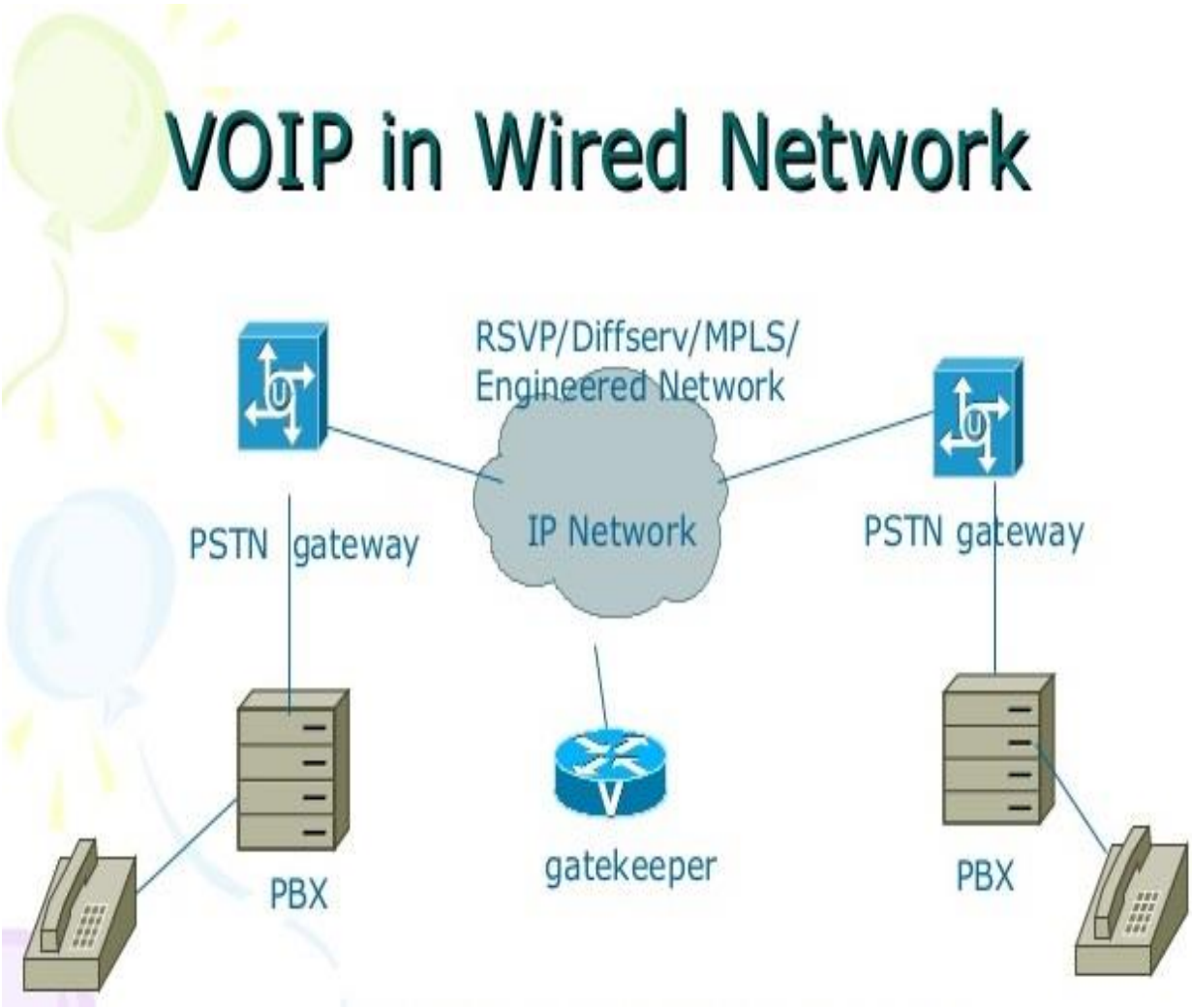


Fig.1: Representation of Computer Network

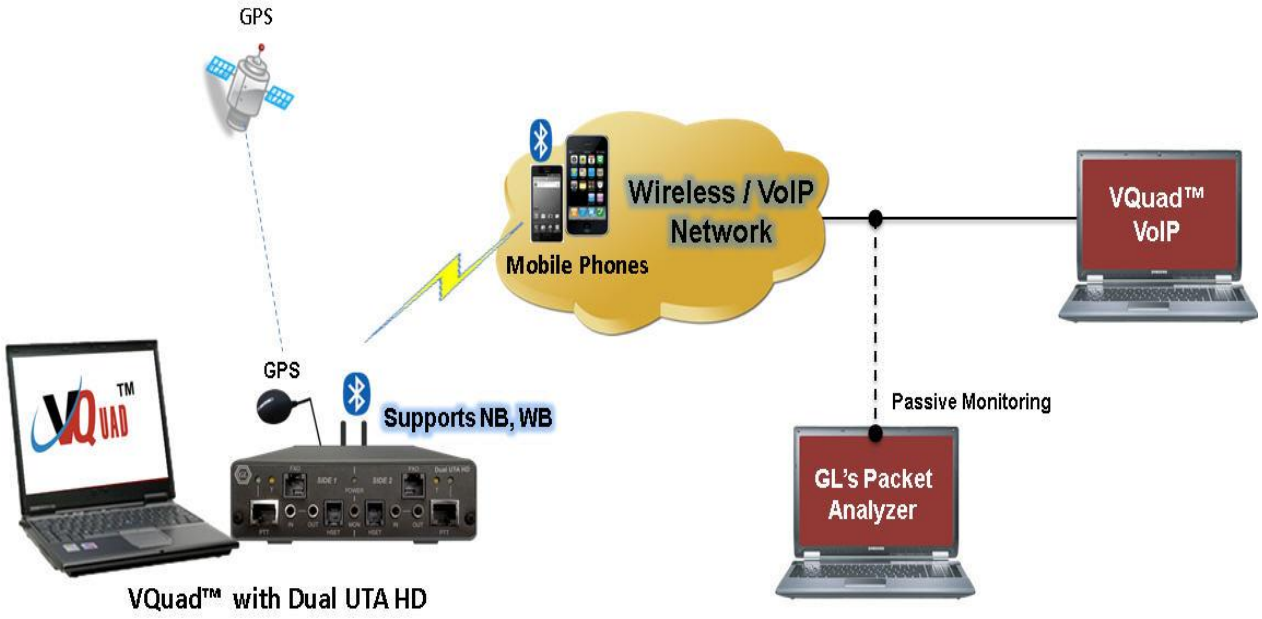
The above Fig.1 shows the representation of computer network. In diagram there are two laptops and one mobile system that are communicating with each other through the internet.

The network can be of two types such as:

Wired networks: They are those which utilize wires for providing information amongst each other. For the wellspring of correspondence wired system can utilize Optical fiber link, coaxial link, bent match link, and so on.



Wireless types of networks: The network which doesn't require any wire or those uses wireless sources such as microwaves or communication satellites as a medium for communication is known as wireless network.



1.2 VoIP Traffic

VoIP remains for Voice over Internet Protocol that utilizations web or other information organize instead of utilizing regular Public Switched Telephone Network (PSTN). A rapid growth has been seen in use of internet for voice communications that results in reduce cost of equipment, operation and maintenance. The VoIP is a strong innovation that enables individuals to convey through voice utilizing IP convention rather than phone lines. The property gauges, high sticker price, restricted combination with existing communication situations are a portion of the elements that have appointed this

innovation in a specialty advertise. Presently a day's circumstance has been changed because of approach of reference bullet and additionally minimal effort VoIP phone connectors open source apparatuses.

This has turned out to be simple and basic for web providers to give their customers VoIP calls expecting practically zero exertion, If any notwithstanding standard xDSL network. Headway in VoIP additionally coordinates the advancement of merged systems that help both video and voice administrations not exhibited by customary PSTN

Despite the fact that VoIP is minimal effort or free innovation still different telecom administrators endeavour to disguise VoIP activity deliberately to stay away from location what's more, escape from charges i.e. Access Promotion Charge (APC) by changing particular parameters in VoIP bundles.

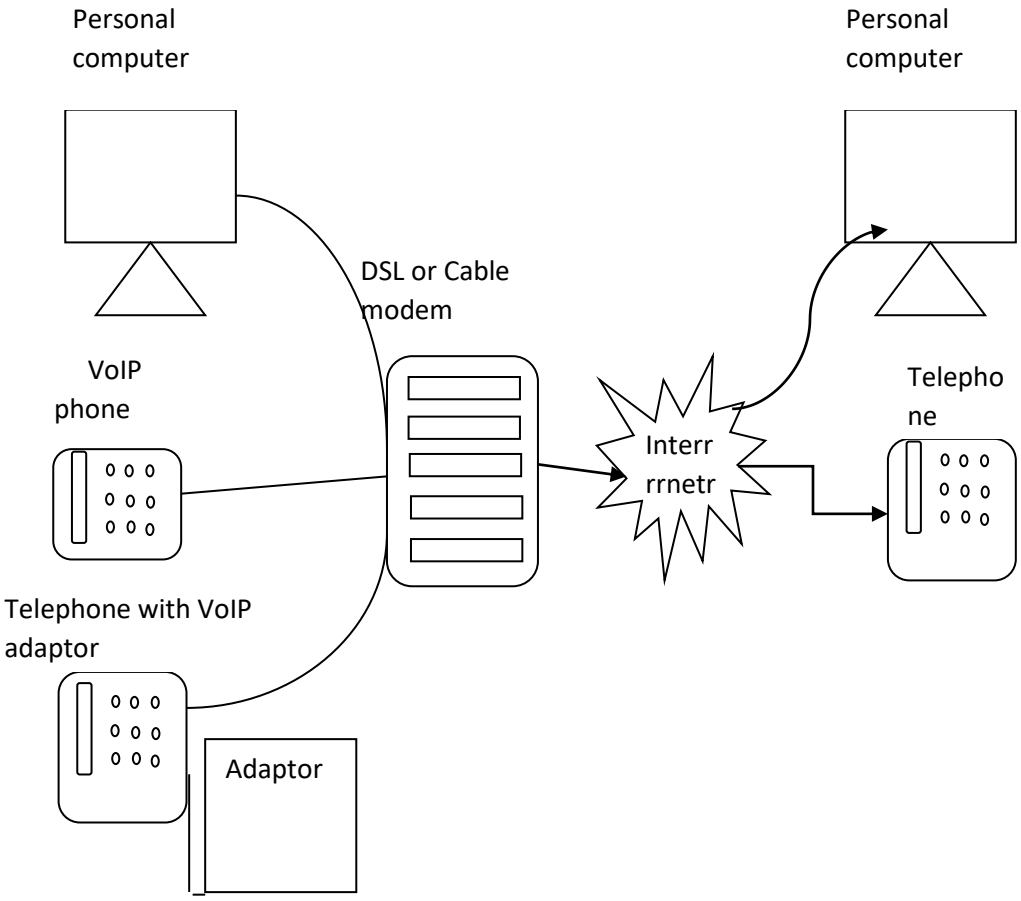


Fig. 2: Working of VoIP

The explanation for concealing the correspondence for illicit use is to dispose of assessments connected by government that confine the call rates to their clients. Illegal telecom development is seen as a hazard to the national security and it brings about a tremendous misfortune to the national exchequer and to the current administrators also. The character has been covered up in illicit telecom dark movement in which calls are brought outside the nation and regarded as nearby calls. The network traffic is need to be monitored and analyzed by law enforcement agencies to avoid frauds. Suppliers need to group the sort of movement transported through their system especially VoIP calls. The biggest piece of conventional pay wellspring of suppliers is VoIP that is the reason primary concentrate is on it. That is the reason they get less advantage from their major and most business undertaking customers as an essential bit of the action goes undetected and uncharged. To confine diminish development there is a need to make techniques to dismember IP action, recognize and distinguish diminish VoIP calls and after that basically piece them or charge them. The above shown Fig. 2 is the diagram of VoIP this is showing the working of VoIP. Standard VoIP tradition, for instance, SIP and H.323 are extraordinarily unmistakable in the conveyor condition and in various distinctive fields not obliged to VoIP, for instance, conveyance individual and talk. Notwithstanding these gauges based applications, there are different applications, for example, Skype or voipstunt that rather depend on exclusive correspondence conventions, codecs and other cross breed applications incompletely in view of open measures, for example, google talk and thingamabob. The outcome is that VoIP is getting to be in some courses like P2P (shared), as:

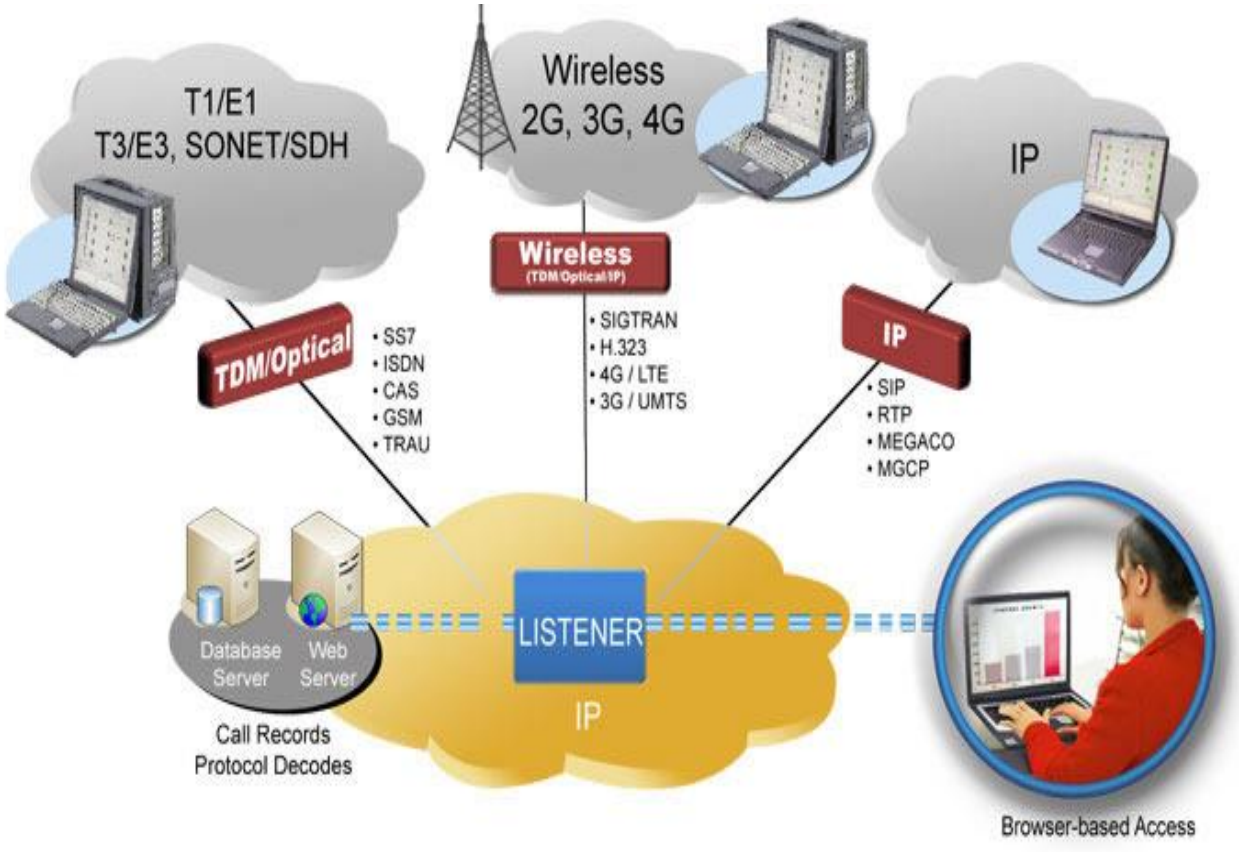
New applications show up, develop and vanish regularly.
Some VoIP applications (e.g. Skype) are utilizing P2P as correspondence transport for building the correspondence foundation and intersection firewalls.
A commonplace situation where numerous standard-based VoIP application neglect to work.
More or less, VoIP arrangements are frequently utilized at corporate level as a practical answer for phone correspondences, though exclusive VoIP applications are utilized for giving individuals a chance to talk either PC to-PC or PC to-phone utilizing a PC furnished with an extraordinary application and a headset.

1.3 Issues of VoIP

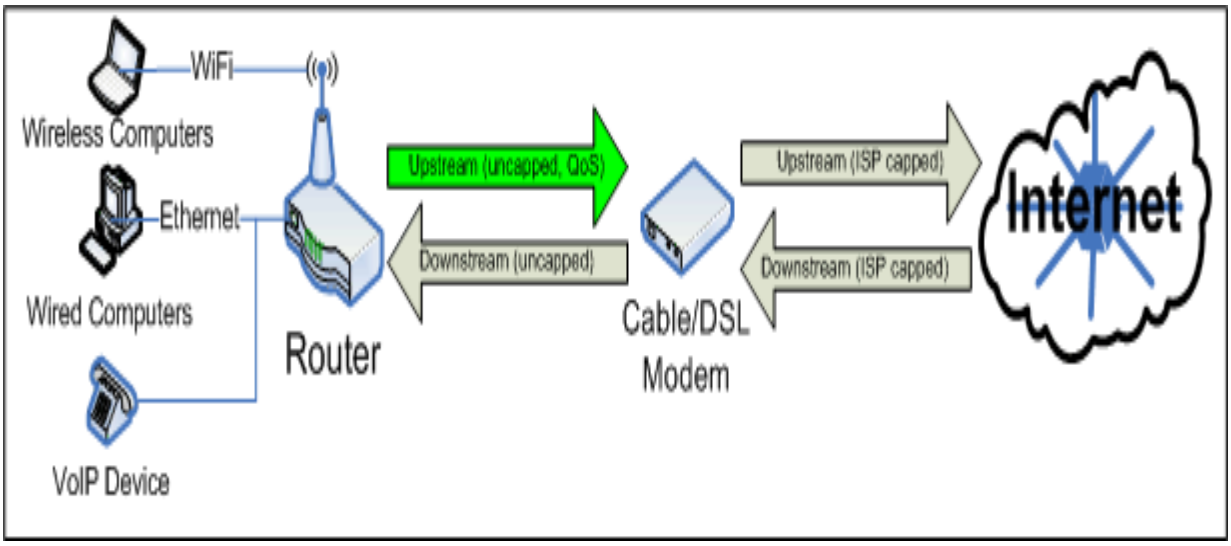
There are different benefits provided by VoIP technology along with that there are various issues with which users have to deal while using it are given below:

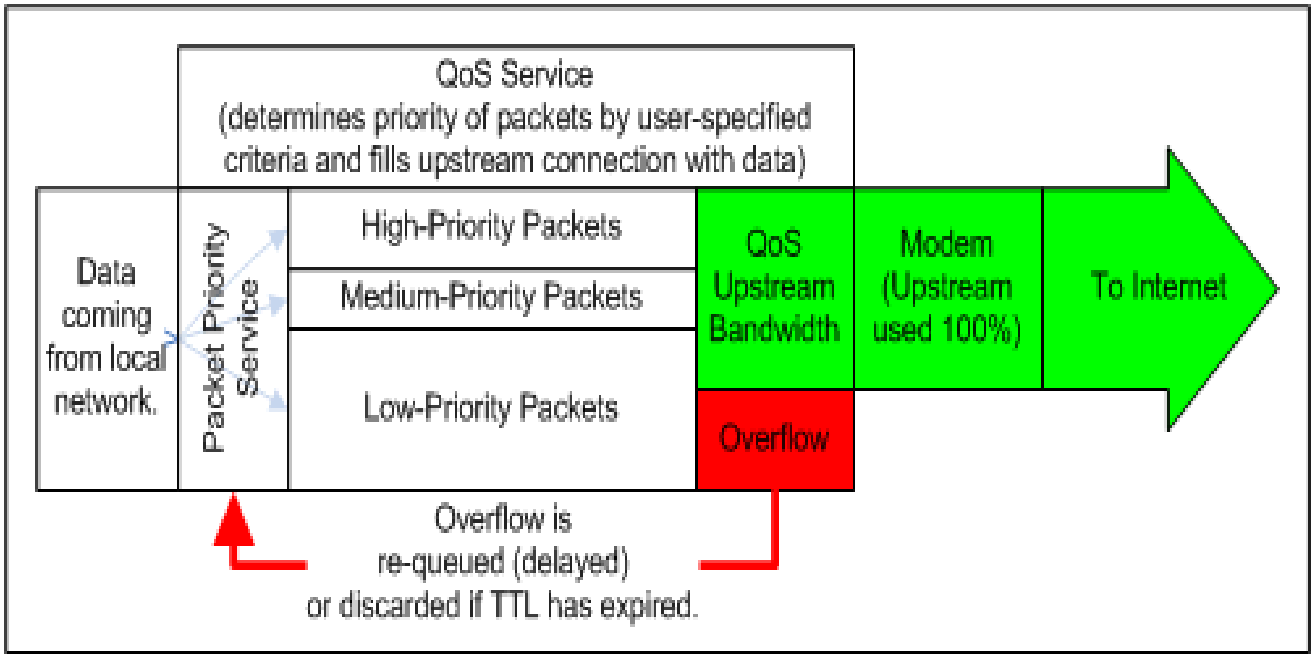
- **Complicated administration and system design:** The joining of various administrations like voice, video, information, et cetera into a similar system makes planning of system design troublesome. As various conventions and gadgets are included for each administration and different qualities are considered for every medium. It likewise causes different blunders what's more, makes it harder to explore and limit them.

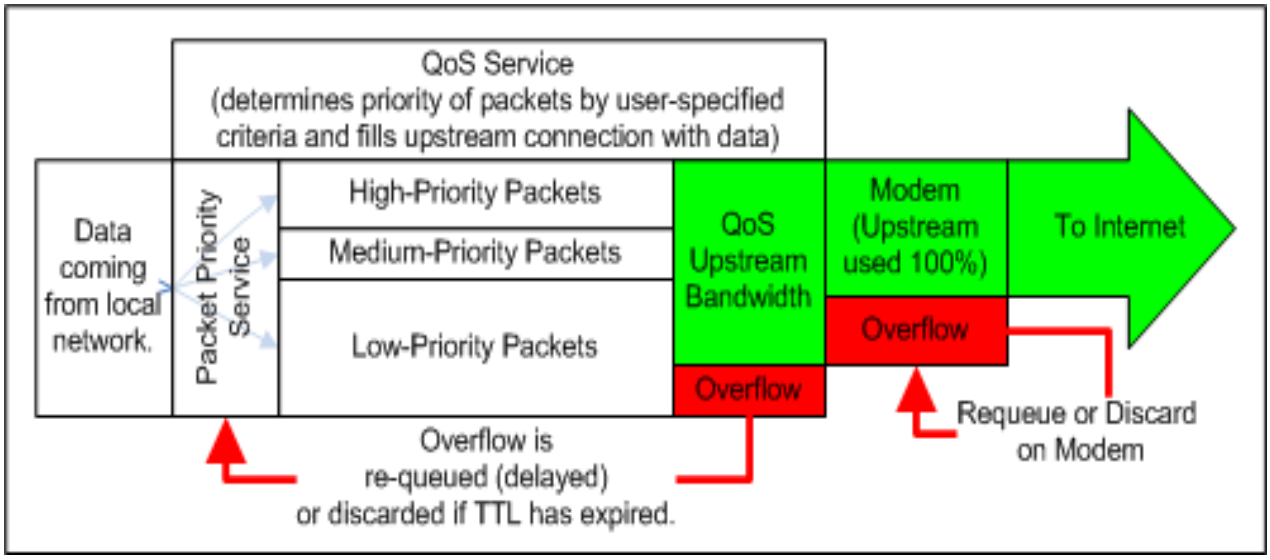
- **Interoperability issues between different applications, or things:** The H323, SIP, IAX, and MGCP are diverse conventions that have been proposed for sending of VoIP frameworks. This creates interoperability issues between the created VoIP gadgets in view of various conventions. Interoperability issues still come up between items utilizing a similar convention because of the huge number of convention renditions, and the methods for usage.

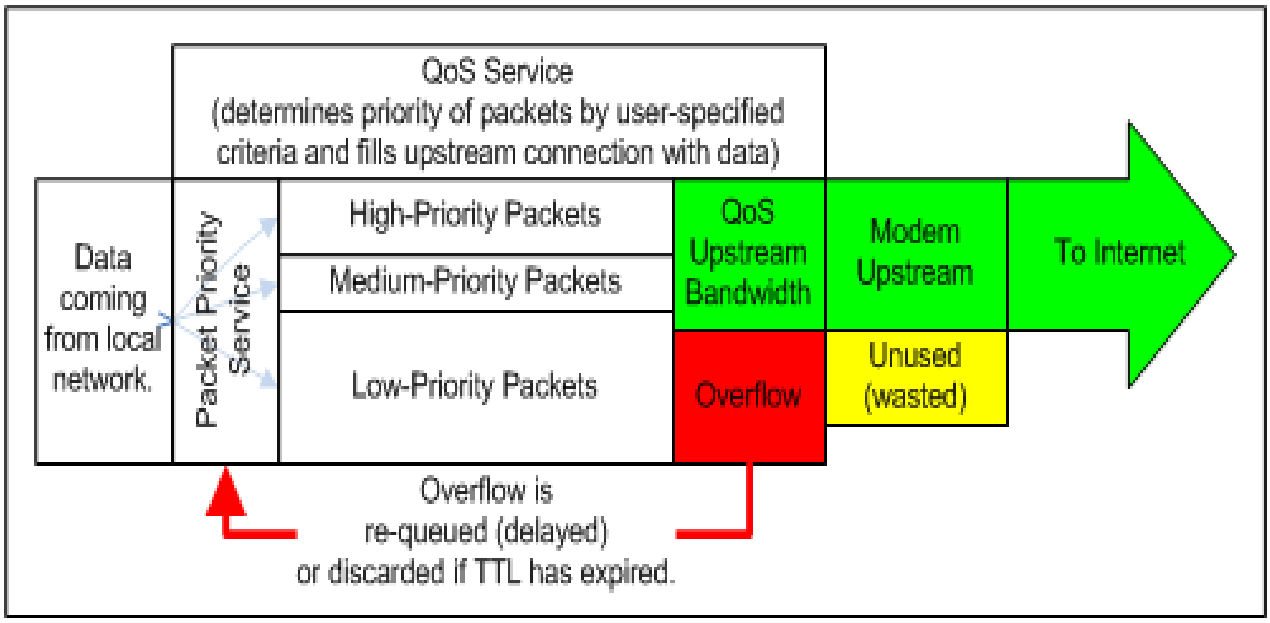


- **Quality of administration (QoS) issues**: At the point when IP innovation was outlined the QoS perspectives was not been considered in it. This makes it IP innovation wasteful to help movement with various QoS imperatives regardless of the advancement of various methodologies (Differentiated services, Integrated services) for the difference in the QoS gave by an IP deal with.

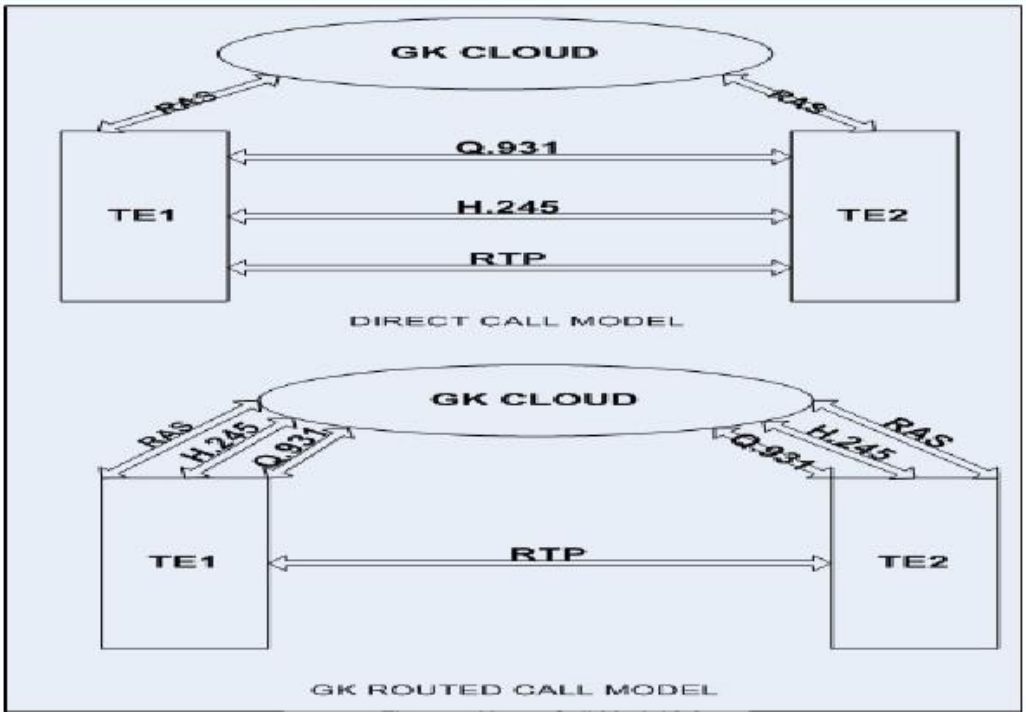








- **Security issues:** In the legacy phone structure Public Switched Telephone Network (PSTN), the real security issue is the square undertaking of discoursed that require physical access to telephone lines. In VoIP security issues are altogether more than that. In reality, in VoIP frameworks numerous components like IP telephones, get to gadgets, media doors, intermediary servers, and conventions are engaged with setting up a call and exchanging media between two endpoints. Every component has powerless elements that might be misused by assailants to complete security assaults.

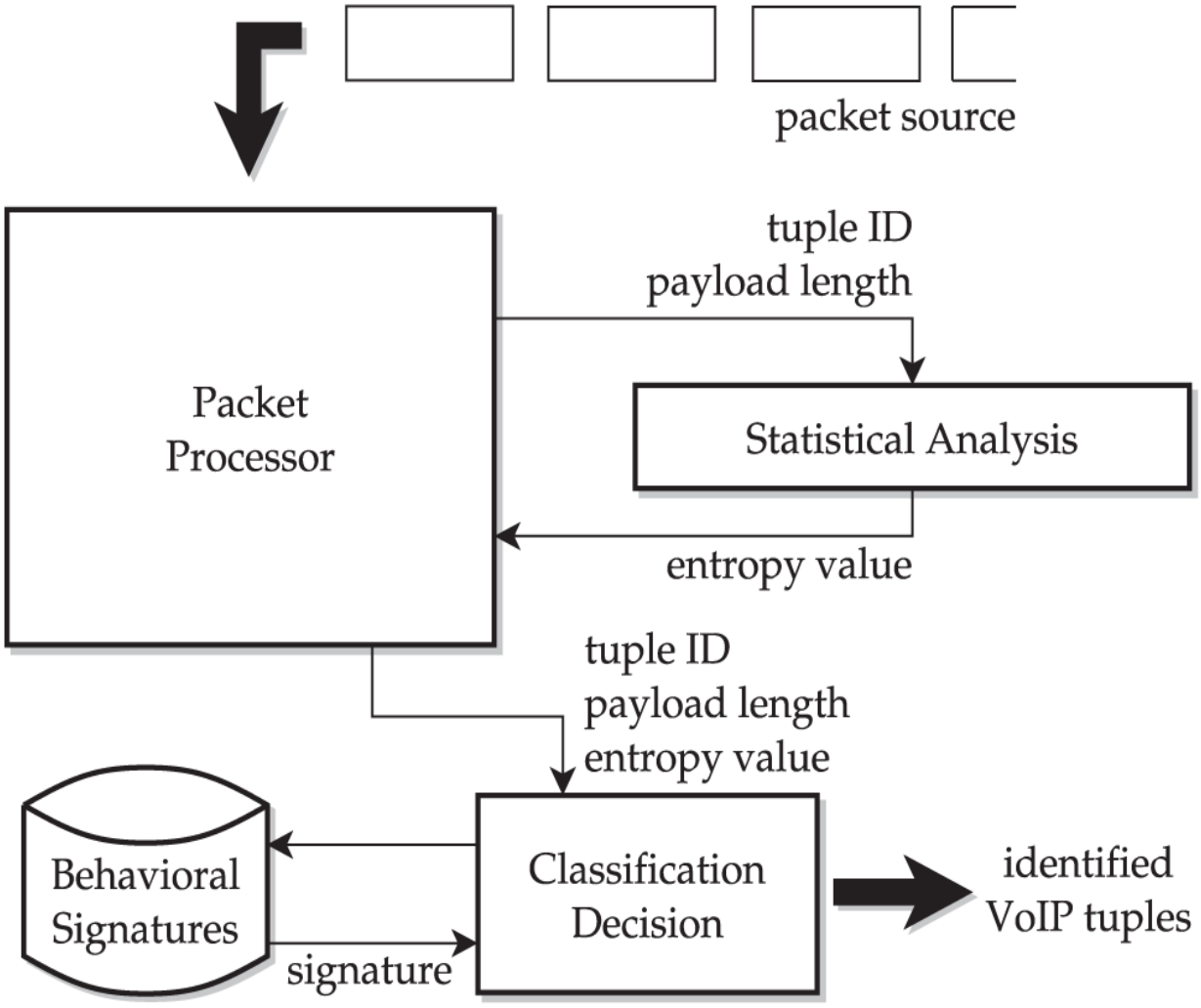


Among the above exhibited issues, VoIP security issues are winding up more veritable in light of the fact that customary security contraptions, traditions and outlines can't palatably shield VoIP systems from late keen ambushes.

1.4 Statistical techniques

Data Analysis can be portrayed as the path toward reviewing and surveying the data that is amassed from different sources. Data cleaning is basic as this will help in getting rid of the overabundance information and arriving at the correct conclusions. Data examination is the think method of cleaning, exploring and changing data with the help of various mechanical assemblies and frameworks. The objective of data examination is to recognize the important information which will support the fundamental administration process.

There are diverse systems for data examination which consolidates data mining, data observation and Business Intelligence. Examination of data will help in compacting the results through examination and illustration of the important information. Data examination helps in choosing the idea of data and working up the reactions to the request which are valuable to the master.



Keeping in mind the end goal to find the arrangement of the issue and to reach to the particular and quality outcomes, different factual procedures can be connected. These systems will help the analyst to get exact outcomes by drawing connections between various factors. The factual procedures can primarily be partitioned into two.

1.4.1 Parametric test: Parametric bits of knowledge consider that the example data relies upon certain settled parameters. It mulls over the property of the populace. It accept that the specimen information is gathered from the populace and populace is ordinarily appropriated. There are equivalent odds of event of the considerable number of information display in the populace. The parametric test depends on different suppositions which are should have been holding great. Unmistakable parametric tests are Analysis of Variance (ANOVA), Z test, T test, Chi Square test, Pearson's coefficient of relationship, Regression examination.

1.4.1.1 T-Test: T-test can be portrayed as the test which helps in perceiving the tremendous level of capability for a situation mean or between the methods for two cases. It is in like way called as a T-Distribution. The t-test is driven when the case size of the general population is near nothing and instability of the majority isn't known. The t-test is utilized when the majority (n) isn't more prominent than 30. There are two sorts of T-Test:

- Dependent mean T Test-It is used when same variables or get-togethers are tried.

- Independent mean T Test-It is used when two unmistakable social events tried. The two unmistakable social affairs have stood up to differing conditions.

1.4.1.2 Z Test: This test is used when the people is regularly scattered. The example size of the masses is colossal or little, however the difference in the people is known. It is used for differentiating the techniques for the people or for recognizing the criticalness level of refinement between the strategies for two self-governing illustrations. Z test relies upon the single essential regard which makes the test more favorable.

1.4.1.3 Examination of Variance (ANOVA): At the point when there are no less than two straight out data, by then Examination of Variance is utilized. Examination of qualification can be for the most part of two sorts a) one-way ANOVA, b) Two-way ANOVA. One way ANOVA is utilized when the mean of no under three than three get-togethers are mulled over. The factors in every get-together are same. Two-way ANOVA is utilized to find if there is any relationship between two independent factors and ward factors. Examination of Variance depends upon different questions. ANOVA expect that there is a needy variable which can be measured at consistent interims. There

are independent components which are straight out, and there should be no under two arrangements. It moreover expect that the masses is regularly scattered and there is no unprecedented part is accessible.

1.4.1.4 Chi Square test: This test is generally called Pearson's chi-square test. This test is used to find an association between no less than two free straight out components. The two components should be allotted at the all level and should involve no less than two self-ruling get-togethers.

1.4.1.5 Coefficient of Correlation: Pearson's coefficient of relationship is used to draw a connection between two components. It is demonstrated by 'r'. The estimation of r goes between +1 to - 1. The coefficient of relationship is used to recognize whether there is a positive connection, negative association or no connection between two components. At the point when the esteem is 0, it shows that there is no relationship between two factors. When it is under 0, it demonstrates a negative affiliation, and when the esteem is more than 0, at that point it shows a positive affiliation.

1.4.1.6 Relapse Analysis: This is utilized to gauge the estimation of one variable which depends upon the estimation of another variable. The variable whose respect is normal is the down and out factor, and the variable which is used to predict the estimation of another variable is called free factor. The assumptions of backslide examination are that the variables should be measured at the steady level and there should be a straight association between two components.

1.4.2 Non-parametric test: Non-Parametric Statistics does not consider any assumption relating to the parameters of the masses. It clears up that data is ordinal and isn't vital to be usually scattered. The non-parametric test is generally called a movement free test. These tests are moderately less troublesome than the parametric test. Unmistakable non-parametric tests join Fisher-Irwin Test, Wilcoxon Matched – Pairs Test (Signed rank test), Wilcoxon rank-add up to test, Kruskal-Wallis Test, Spearman's Rank Correlation test.

1.5 Types of classifiers:

- **k-Nearest Neighbor:** In this kind of classifier, a patter x is grouped by allotting class name to it that is most much of the time spoke to among it's k closest examples. The class with least normal separation is utilized to dole out a test design that demonstrates that this technique is delicate to remove work. The Euclidean separation metric is utilized for getting least normal separation. All highlights are standardized into same range this is the fundamental necessity of this metric approach. The k-closest neighbor classifier is a regular nonparametric classifier that is said to yield great execution for ideal estimations ofk.

- **Bayesian Classifier:** In supervised parametric classifiers theory, most general approach used is quadratic discrimination. When dealing with d-dimensions the obtained decision boundaries by these classifiers can become very complicated. Most of the discriminant function generation computation has been done off-line. This approach can be more influenced by revile of dimensionality as in this quadratic discriminant an expansive number of parameters should be considered. If there should be an occurrence of little preparing tests its execution is influenced definitely. In case of small training samples its performance is affected drastically.

- **Multi-layer Perceptron (MLP):** The multi-layer perceptron classifier is a fundamental encourage forward counterfeit neural system. They have utilized a solitary concealed layer at first for straightforwardness (streamlines picking the quantity of neurons) and after that went for two covered layers for better request execution. The covered units were picked unmistakably for each datum set. The quantity of shrouded neurons was discovered tentatively finished various trials. A dependable guideline is to pick the quantity of concealed neurons with the true objective that the total number of weights in the net is by and large n/10, n being the total number of getting ready core interests. The neural framework was readied using the back-inciting computation, According to the multi-layer perceptron arranged using the back-multiplication learning count approximates the perfect discriminant work portrayed by Bayesian theory.

- **SVM Classification:** SVM is a characterization calculation in light of improvement hypothesis and at first created by. Here, a protest is seen as a n-

dimensional vector and it isolates such questions with a n-1 dimensional hyperplane. This is known as an immediate classifier. There are various hyperplanes that are utilized to group information.

Chapter-2
Literature Survey

Mazhar Rathore, et.al, (2016), Web get to providers (ISPs) are enthusiastic about recognizing VoIP calls either to square unlawful business VoIP or sort out the paid customers VoIP calls. Stamp based, port-based, and configuration based VoIP ID techniques are not more correct and not beneficial as a result of complex security and tunneling instruments used by VoIP. In this paper, creators have proposed another plan in view of non specific control, vigorous and effective factual examination that aides in recognize scrambled, non-encoded, burrowed VoIP media streams utilizing conventional approach. They have chipped away at effective process for rapid ongoing system movement. Precision correlation of the proposed framework with existing strategy security component, can distinguish scrambled burrowed VoIP and implementable at perhaps one-way or two-way organize interface. It tends to the issue of any relationship to recognize VoIP streams to either arrange or square. They have attempted their answer on many insights of more than 10 VoIP applications. The connections and results exhibit that their proposed technique is the best among all the present methods. This technique has 97.54% TP and .00015% FP. It is the better choice for media transmission specialists and ISPs to recognize VoIP acquires quick immense Data condition.

Muhammad Shafiq, et.al, (2016), have recommended network traffic classification as a central topic for researchers in the field of computer science. It is vital errand for web access suppliers (ISPs) to know which sorts of system applications stream in a system. Network Traffic Classification is the first step to analyze and identify different types of applications flowing in a network. With the help of this technique internet service providers or network operators are able to manage the overall performance of a network. There are many methods traditional technique to classify internet traffic like Port Based, Pay Load Based and Machine Learning Based technique. The most common technique

used these days is Machine Learning (ML) technique. This is used by many researchers and got very effective accuracy results. In this paper, authors have discussed step by step techniques of network traffic classification and develop a real time internet data set using network traffic capture tool. At that point the highlights are removed from the catch activity utilizing devices of highlight extraction the connected a Support Vector Machine, C4.5 choice tree, Naive Bays and Bayes Net machine learning classifiers. The experimental and simulation results show that C4.5 classifiers prove to be good in terms of accuracy as compared to other existing classifiers.

Aboagela Dogman, et.al, (2014), have introduced overseeing nature of administration (QoS) as a vital system operation for the most part in cross breed wired and remote sight and sound systems. In this paper , creators given an explored and built up an approach in view of two phases to astutely oversee QoS for sight and sound activity. As a run of the mill sight and sound application they have considered VoIP and connected a versatile factual examining system in starting stages. It helps in deciding the statics of movement and afterward utilized them in a fluffy surmising framework that aides in deciding ideal interim between each two traffics tested sequential segments. A fluffy c-implies (FCM) grouping was utilized to pre-process deferral, jitter and parcel misfortune proportion like QoS parameters in second stage that are acquired from the concocted inspecting plan. Keeping in mind the end goal to survey the VoIP accommodated QoS the FCM data is utilized by multilayer perceptron (MLP) neural system. The reenactment comes about demonstrate that movement are spoken to more accurately by created versatile measurable inspecting than the deliberate, stratified and irregular non-versatile testing strategies. The mix of factual inspecting took after by FCM and MLP are all the more precisely showed the QoS for VoIP.

Jaiswal Rupesh Chandrakant, et.al, (2013), have examined that web movement acknowledgment procedures has turned out to be critical for specialists on the grounds that these methods are free of TCP or UDP port numbers. The movement is characterized utilizing new methodologies by perceiving factual examples in remotely activity

noticeable characteristics. The principle objective of analysts is to bunch or order stream of web activity into indistinguishable measurable properties gatherings. The explanation for presentation of Machine Learning (ML) procedures is the need to manage activity designs, extensive datasets and multidimensional spaces stream and bundles characteristics. In activity acknowledgment ML procedures has been utilized which are the subset of counterfeit consciousness. The Classification, grouping, Numeric forecast and Association are the four sorts of Machine Learning. In this paper, creators have actualized activity acknowledgment through arrangement process. They have utilized diverse standard datasets then a diminished measurable element dataset has been created utilizing standard trait choice calculations. AdaboostM1, C4.5, Random Forest tree, MLP, RBF and SVM are six ML calculations that are utilized for IP movement grouping with Polykernel work classifiers. The reproduction and execution comes about demonstrate that Tree based calculation are more powerful ML strategies for web activity grouping as far as accomplished precision of 99.7616%.

Riyad Alshammari, et.al, (2015), have analyzed the performance of C5.0, AdaBoost and Genetic programming (GP) like three different machine learning algorithms that generate robust classifiers to identify VoIP encrypted traffic. One of the scientists has utilized a machine learning based novel approach that create hearty mark for ordering VoIP encoded information. A factual figuring is connected stream of system that aides in removing set of highlight without including payload data, data in view of the port of source, goal and IP addresses. The results show that performance of classifying VoIP significantly can be improve by employing most suitable sampling and machine learning technique. In this paper [21], authors have found it very challenging to find robust rules specifically to detect encrypted VoIP Skype network traffic. The authors have investigated how to form a training set when machine learning based approach is used for classifying network traffic without including port numbers, IP addresses, or payload information. The results indicates that a priori information is resulting in ‘‘over learning’’ on our data sets. Given the outcomes got in this examination paper, one without bounds bearings which can be taken after is investigate whether a comparable pattern for other system applications.

M. Mazhar, et.al, (2015), have introduced the development in viability cost, sensational usefulness over the conventional phone organize. The similarity with open exchanged

phone arrange (PSTN) has been seen. The business utilization of VoIP has been denied in a portion of the nations like Pakistan. Network access suppliers (ISPs) and media transmission experts are keen on identifying VoIP calls to either square or organize them. So discovery of VoIP calls is essential for the two sorts of experts. Mark based, port-based, and design based VoIP discovery strategies are wasteful because of mind boggling and classified security and burrowing components utilized by VoIP. In this paper [22], creators have proposed a non specific, vigorous, proficient and for all intents and purposes implementable factual examination based answer for recognize encoded, non-scrambled or burrowed VoIP media (voice) streams utilizing edge estimations of stream measurable parameters. The recreation aftereffects of proposed procedure have been contrasted and existing methods and enhanced outcomes has been found as far as exactness and productivity. The accomplished direct rate is 97.54% and regarding false positive rate it is .00015%.

CHAPTER-3

Scope of Study

The grouping is the procedure which is connected to characterize the info information as indicate by the portrayed classes.The network traffic classification technique is been applied which will classify the network traffic according to user activities. The classes which are used to classify the network traffic is malicious and non-malicious. In this work, the dataset of wire shark is taken as input for the classification. In the process of classification three steps has been followed in which in the first step, the data is pre-processed in which redundancy from the data is removed, in the second step technique clustering is applied in which similar and dissimilar type of data is clustered together. In the technique of k-mean clustering, the data is taken as input, the arithmetic mean of the whole dataset is calculated which will be the central point. The Euclidian distance from the central point is calculated and points which is having similar distance is clustered in one cluster and other in the second cluster. The technique of SVM classification is applied which will classify the data into malicious and non-malicious class. In this work, improved technique will be applied for the classification which increase accuracy of classification, reduce execution time of the algorithm.

CHAPTER-4

Objectives

1. To study and analyze various network traffic classification for the data mining

2. To propose improvement in the existing SVM based classification technique for the network classification

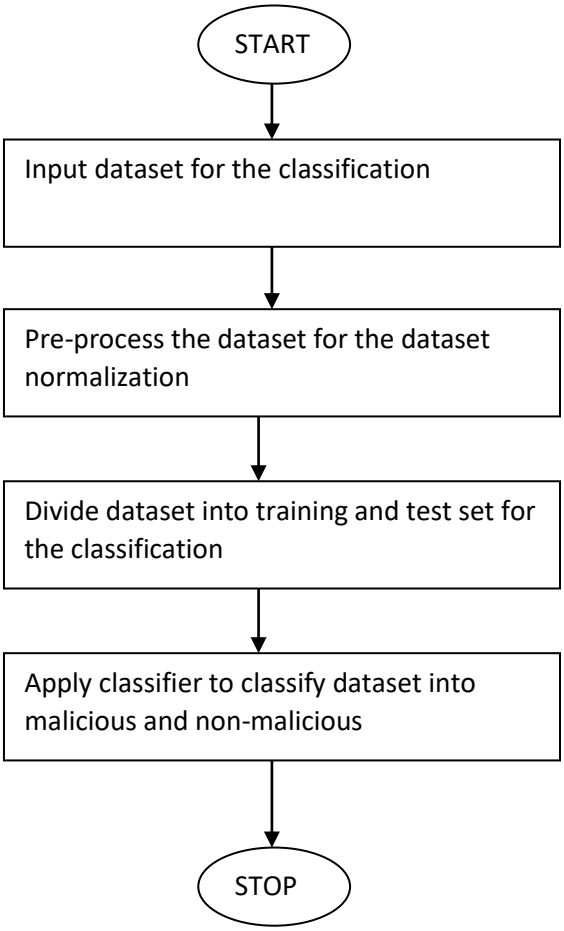
3. The proposed improvement will be based on the KNN approach for the traffic classification

4. Implement proposed technique and compare results with the existing in terms of accuracy, execution time

CHAPTER-5

Research Methodology

This work is based on the network traffic classification to classify the traffic into malicious, non-malicious. The network traffic analysis is the technique which is applied to predict the malicious activities of the users which are active on the network. To classify the network traffic three steps has been followed in the methodology, in the first step technique of k-mean clustering is been applied in which similar and dissimilar type of data will clustered. The dataset which is taken as input will be refined by removing redundancy and missing values. In the second step, technique of k-mean clustering is applied in which arithmetic mean of the whole dataset is calculated which will be the central point of the dataset. The Euclidian distance from the central point is calculated which define the similarity and dissimilarity of the points. The points which are similar will be clustered in one cluster and other in the second cluster. In the last step of classification technique, SVM classifier will be applied which classify the data into two classes. To enhance the execution of the current framework system of Knn classifier
will be applied which will cluster the uncluttered points and increase accuracy of classification. The Knn classifier the nearest neighbor classifier in which Euclidian distance is calculated and points which have similar distance will be clustered in one class and other in the second class.


START
Input dataset for the classification
Pre-process the dataset for the dataset normalization
Divide dataset into training and test set for the classification
Apply classifier to classify dataset into malicious and non-malicious
STOP

5.1 Expected Outcomes

Following are the various expected outcomes of this research

1. The proposed change will be founded on to recognize the IP deliver which are mindful to do vindictive exercises.
This leads to increase security of the network .
2. The proposed algorithm will be based on classification which can classify malicious and non- malicious locations

CHAPTER 6

CONCLUSION:

Information characterization is a critical undertaking in machine learning. It is related to create PC programs prepared to pick up from named informational indexes and, along these lines, to foresee unlabeled occurrences. In light of the immense number of uses, various information characterization frameworks have been produced. A segment of the outstanding ones are choice trees, example based learning, e.g., the K-closest neighbors calculation (KNN), manufactured neural systems, Naive-Bayes, and bolster vector machines (SVM). Everything considered, most of them is profoundly needy of suitable parameter tuning. Cases incorporate the certainty factor and the base number of cases to segment a set in C4.5 choice tree; the K esteem in KNN; the stop paradigm, the quantity of neurons, the quantity of concealed layers, and others in manufactured neural systems; and the delicate edge, the piece work, the bit parameters, the halting basis, and others in SVM.

References:

[1] Jeroen Hoebeke, Ingrid Moerman, Bart Dhoedt and Piet Demeester, "An Overview of Mobile Ad Hoc Networks: Applications and Challenges", IJSER, 2005.

[2] ZHOU, L., AND HAAS, Z. J. Securing Ad Hoc Networks. IEEE Network 13, vol.6, pp. 24-30, 1999.

[3] Uzma Anwar, Ghulam Shabbir, Malik Ahsan Ali, "Information Analysis and Summarization to Detect Illegal VOIP Traffic with Call Detail Records", International Journal of Computer Applications (0975 – 8887), vol. 89, pp. 1-7, 2014.

[4] Kuan-Ta Chen , Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei, "Distinguishing VoIP Traffic Based on Human Conversation Pattern", Principles, Systems and Applications of IP Telecommunications. Administrations and Security for Next Generation Networks, Springer-Verlag Berlin, Heidelberg, vol. 7, pp. 280 - 295, 2008.

[5] Yoseba K. Penya , Igor Ruiz-Agundez and Pablo G. Bringas, "System PLANNING OF A VOIP CAPABLE PBX", International Conference on Data Communication Networking (DCNET) , SciTePress, Seville, Spain, vol. 8, pp. 85-88, 2011.

[6] Patrick stop, "voice over IP Security ", Cisco Press, September 2008, ISBN-10: 1-58705-469-8

[7] Peter Thermos; Ari Takanen, "Securing VoIP Networks: Threats, Vulnerabilities, and Countermeasures", Addison-Wesley Professional, August 2007, ISBN-10: 0-321-43734-9

[8] Meisel, J.B. furthermore, Needles, M. (2005), "Voice over web convention (VoIP) improvement and open strategy suggestions", information, Vol. 7 No. 3, pp. 3-15.

[9] Olivier Hersent, Jean-Pierre Petit, and David Gurle, "Past VoIP Protocols: Understanding Voice Technology and Networking Techniques for IP Telephony", Wiley; 1 release (March 4, 2005), Edition 1, ISBN-10: 0470023627

[10] Manikandan S., "Measures of focal inclination: Median and mode", J Pharmacol Pharmacother, vol. 5, pp. 212-214, 2011.

[11] Myles PS, Gin T., "Factual Methods for Anesthesia and Intensive Care", Ist ed. Oxford: Butterworth Heinemann, vol. 7, pp. 8-10, 2000.

[12] Binu VS, Mayya SS, Dhar M., "Some fundamental parts of factual strategies and test

estimate assurance in wellbeing science look into", Ayu, vol. 35, pp. 119-123, 2014.

[13] Altman DG, Bland JM., "Parametric v non-parametric techniques for information investigation", BMJ., vol. 7, pp. 316-338, 2009.

[14] R.O. Duda, et.al, "Example Classification second Ed", John Wiley and Sons Inc., 2000.

[15] D.W. Ruck, et.al, "The Multi-Layer Perceptron as an Approximation to a Bayes Optimal Discriminant Function", IEEE Transactions on Neural Networks, vol. 1, no. 4, 1990,

[16] Cortes, C., et.al, "Bolster Vector Networks, Machine Learning", vol. 20, pp. 273-297, 1995.

[17] Mazhar Rathore, Anand Paul, Awais Ahmad, Muhammad Imran, Mohsen Guizani, "Fast Network Traffic Analysis: Detecting VoIP Calls in Secure Big Data Streaming", 2016 IEEE 41st Conference on Local Computer Networks, vol. 7, pp. 595-598, 2016.

[18] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudil Abdessamia, "System Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2016 second IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.

[19] Riyad Alshammari, A. Nur Zincir-Heywood, "Recognizable proof of VoIP scrambled activity utilizing a machine learning approach", Journal of King Saud University – Computer and Information Sciences, vol. 27, pp. 77– 92, 2015.

[20] M. Mazhar, U. Rathore, "Limit based nonexclusive plan for scrambled and burrowed Voice Flows Detection over IP Networks", Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305– 314, 2015.

[21] Aboagela Dogman, Reza Saatchi, "Interactive media movement nature of administration utilizing measurable and counterfeit consciousness strategies", The Institution of Engineering and Technology 2014, vol. 8, pp. 367– 377, 2014.

[22] Jaiswal Rupesh Chandrakant, Lokhande Shashikant. D., "Machine Learning Based Internet Traffic Recognition with Statistical Approach", 2013 Annual IEEE India Conference (INDICON), vol. 7, pp. 121-126, 2013.

[23] D. Lim, Principal Component Analysis using Singular Value Decomposition of Microarray Data, International Journal of Mathematical, Computational, Physical and Quantum Engineering, 7(9) (2013), 859-861.

[24] 75P. Drotár, J. Gazda, Z. Smékal, An experimental comparison of feature selection methods on two-class biomedical datasets. Computers in Biology and Medicine 66 (2015) 1–10.

[25] G. Exarchakos , L. Druda , V. Menkovski , A. Liotta ,Network analysis on Skype end-to-end video quality, Int. J. Pervasive Comput. Commun. 11 (1) (2015) 17–42 .

[26] J. Suh, T. Kwon, C. Dixon, W. Felter, J. Carter, OpenSample: A Low-Latency, Sampling-Based Measurement Platform for SDN, IBM Research Report (2014).

[27] M. Kwon , Z. Dou , W. Heinzelman , T. Soyata , H. Ba , J. Shi , Use of Network Latency Profiling and Redundancy for Cloud Server Selection, in: Proceedings of the IEEE 7th International Conference on Cloud Computing (CLOUD), 2014, pp. 826–832 .

[28] M. Malboubi , L. Wang , C.N. Chuah , P. Sharma , Intelligent SDN based traffic (de)aggregation and measurement paradigm (iSTAMP), in: Proceedings of the IEEE INFOCOM'14, 2014 .

[29] Cisco Visual Networking, Index: forecast and methodology, 2015–2020, 2016. June 6 June 6 http://www.cisco.com/c/en/us/solutions/ collateral/service- provider/ip- ngn- ip-next- generation- network/white _ paper _ c11-520862.html .

[30] A Anand , G Veciana , Invited paper: context-aware schedulers: realizing quality of service/experience trade-offs for heterogeneous traffic mixes, in: 2016 IEEE 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2016, pp. 1–8 .

[31] Y N Dong , L T Yao , H X Shi , Fine grained classification of internet video traffics, in: 2015 21st Asia-Pacific Conference on Communications (APCC), Kyoto, 2015, pp. 580–584 .

[32] N Al Khater , E Overill R , Network traffic classification techniques and challenges, in: Tenth International Conference on Digital Information Management, IEEE, 2015, pp. 43–48 .

[33] D Jiang , L Tao , P2P traffic identification research based on the SVM, in: Wireless and Optical Communication Conference (WOCC) 16-18 May 2013, Chongqing, 2013 22nd., IEEE, 2013, pp. 6 83–6 86 .

[34] ZJ Wang , YN Dong , HX Shi , LY Yang , PP Tang , Internet video traffic classification using QoS features, in: 2016 International Conference on Computing, Network- ing and Communications (ICNC), Kauai, HI,,2016, pp. 1–5 .

[35] MMA Patwary , et al. , PANDA: extreme ecale parallel K-Nearest neighbor on distributed architectures, in: 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2016, pp. 494–503 .

[36] S.S.L. Pereira , J.L.d.C.e. Silva , J.E.B. Maia , NTCS: A real time flow-based network traffic classification system, in: 10th International Conference on Network and Service Management (CNSM) and Workshop, Rio de Janeiro, 2014, pp. 368–371 .

[37] J Zhang , X Chen , Y Xiang , L Zhou W , J Wu , Robust network traffic classification, in: ACM Transactions on Networking, 99, IEEE, 2014, p. 1 .

[38] M Zhang , H Zhang , B Zhang , et al. , Encrypted traffic classification based on an improved clustering algorithm, in: Trustworthy Computing and Services, Springer, Berlin Heidelberg, 2013, pp. 124–131 .

[39] GD Gonçalves , ÍCunha , AB Vieira , Predicting the level of cooperation in a peer–to-peer live streaming application, Multimedia. Syst. (2014) 1–20 .

[40] J Datta , N Kataria , N Hubballi , Network traffic classification in encrypted environment: a case study of Google hangout, in: 2015 Twenty First National Conference on Communications (NCC), Mumbai, 2015, pp. 1–6 .