Data Analysis and Summarization to Detect Illegal VOIP Traffic with Call Detail Records

Uzma Anwar
Department of Telecommunication Engineering UET Taxila, Pakistan

Ghulam Shabbir
Department of Telecommunication Engineering UET Taxila, Pakistan

Malik Ahsan Ali
Electrical Engineering Department NUST (MCS), Rawalpindi, Pakistan

ABSTRACT
Voice over Internet Protocol (VoIP) is an advanced area for researchers. Many different methods are used to send voice over IP networks. With the development of modern telecommunications equipments and softwares telecommunication's malpractices are growing rapidly. Hence there is always a need for monitoring communications and guarantee both security and proper usage. This underlined research work stresses on the analysis of IP traffic and proposes an algorithm for detection mechanisms to control and limit VoIP's grey traffic. The algorithm emphasizes primarily on Session Initiation Protocol (SIP) but it can be modified and used for all VoIP protocols like H.323 and Inter-Asterisk eXchange protocol (IAX2). The suggested method is based on analyzing the pcap files. These files are used to filter VoIP traffic from network's total IP traffic by reading the header of each packet. The algorithm then extracts different parameters for generating call logs. VoIP packets of the same call are correlated to produce a Call Detail Record (CDR).

The produced CDR contains the IP addresses of source and destination that make the calls. For identification of grey traffic these IP addresses are used. If the source IP address in the CDR is of a legal registered operator the user/call is declared as legal, otherwise the user/call is illegal.

Keywords
Inter-Asterisk eXchange protocol (IAX), Internet Protocol (IP), Session Initiation Protocol (SIP), VoIP-Voice over IP (VoIP), Call Detail Record (CDRs)

1. INTRODUCTION
For voice communication, Voice over IP (VoIP) uses Internet (or other data networks) rather than conventional Public Switched Telephone Network (PSTN). From last few years, using Internet for voice communications has observed a rapid growth, which results in reducing the cost of equipment, operation and maintenance. Advancement in VoIP also directs the development of convergent networks which support both video and voice services not presented by conventional PSTN.

Though VoIP is low cost or almost free technology still various telecom operators try to conceal VoIP traffic intentionally to avoid detection and escape from taxes i.e. Access Promotion Charge (APC) by altering different parameters in VoIP packets. The motivation behind doing so (hiding the communication for illegal usage) is to get rid of government taxes, which permits lowest call rates to their users. Illegal telecom traffic is regarded as a threat to the national security and it results in a huge loss to the national exchequer and to the existing operators as well. Grey traffic is an illegal telecom traffic, in which calls from outside the country are brought to the country which are treated and charged as local calls by hiding the identity.

In order to overcome frauds, law enforcement agencies often need to monitor and analyze the network traffic. Providers want to classify the type of traffic transported through their network, particularly VoIP calls. Main focus is on VoIP as it utilizes the largest part of traditional income source of providers. That is why they get less profit from their major and most commercial enterprise customers as a major part of the traffic goes undetected and uncharged. To limit grey traffic, there is a need to develop techniques to analyze IP traffic, identify and detect grey VoIP calls and then simply block them or charge them.

A Call Detail Records (CDRs) containing the details of a phone call, is a computer record produced by a telephone exchange for each call that passed through it. Network management and operation tasks rely on the call details record (CDR), in both traditional PSTN and IP telephony networks, to efficiently troubleshoot problems and monitor trends. Call logs contain complete information regarding telephone calls like source and destination identity (numbers and IP addresses), date and time of call, duration of call, and reply codes. As operators spend much of their financial resources on monitoring systems to detect grey traffic, call logs collected for other purposes mainly for billing, identify a comprehensive and useful source of information. Furthermore, these call logs are used for traffic management, security and other engineering purposes.

Grey traffic affects IP networks badly as they do not use any mechanism to keep a check for illegal communication. As variety of applications is being developed due to tremendous usage of VoIP, monitoring and detection of VoIP applications (especially illegal) is becoming cumbersome. This research facet is a vital need of all telecommunications service providers to limit illegal traffic. 40% of international calls go down into illegal class according to a rough approximation [1]. As VoIP calls utilize a large bandwidth and due to the use of proprietary protocols and/or proprietary encryption techniques almost 70% of them go undetected. As a result telecom operators and service providers get high economic and monetary loss.

This paper presents a general algorithm to inspect and analyze network IP traffic and efficiently identify the VoIP illegal traffic using CDRs. This work mainly focus on VoIP calls generated by SIP (can be extended to other VoIP protocols).

2. BACKGROUND
2.1 Session Initiation Protocol (SIP)
In this research work main focus is on Session Initiation Protocol (SIP) a VoIP protocol. Below is a little introduction to SIP, its network elements and messages.

2.2 SIP Network Architecture

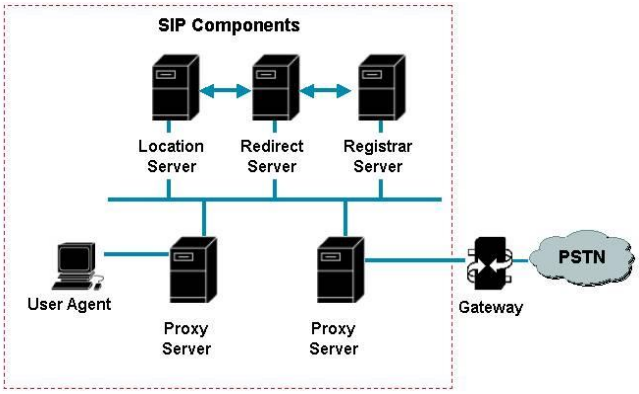

Figure 1 Network Architecture of SIP

Here is a little introduction of SIP network elements.

- **SIP User Agent (UA**) can be a soft/hard phone having SIP calling capabilities.UA (end Points) can set off and receive calls.
- **SIP User Agent Client (UAC**) is a logical element which sets off SIP requests and accepts responses of SIP.
- **User Agent Server (UAS)** is a logical element in SIP network that accepts requests of SIP and in return send SIP responses.
- **Proxy server** in SIP network sends traffic to and receives from UAs and other locations or devices.
- **Registrar** accepts UAs requests and registers and authenticates them on a network. Like GSM HLR SIP Registrar maintains user's location on a Location Server.
- **Redirect server** in SIP network receives SIP requests. In the case of numerous locations for SIP, to complete the initial request UAs returns the address that should be contacted.
- SIP **Location Server** maintains a database of registered users and locates them, who are registered through a SIP server.

2.3 Process Establishing Communication

Table 1.Steps in the Process of establishing SIP communication

1	Registering, initiating and locating the user.
2	Determine the media to use-involves delivering a description of the session that the user is invited to.
3	Determine the willingness of the called party to communicate- the called party must send a response message to indicate willingness to communicate- accept or reject.
4	Call setup.
5	Call modification or handing – example, call transfer (optional).
6	Call termination.

2.4 SIP Messages

SIP messages are of two types

- Request-send by clients to servers
- Response-send by server to a client

SIP messages have the same basic format however the syntax varies in specifics and character set.

Generic-message = start-line
*message-header
CRLF
[message-body]

Start-line = Request-Line / Status-Line

Figure 2 Generic Format of SIP Message

1. Start-line (Status-line/Request Line)

2. Header

3. Body (optional)

Fig. shows the layout of a SIP message along with an example for an INVITE message

Request line	Method	Request URI	SIP Version	INVITE sip:011923218282002 @sip.velocitydial.com SIP/2.0
Message Header	Header Fields			Via: SIP/2.0/UDP 192.168.66.6:55960 To:<sip:011923218282002@sip.velocitydial.com> From: "MTA"<sip:7758377@sip.velocitydial.com> Call-ID:MGVmY2ZhNmFhNDhiODZlNTk CSeq: 1 INVITE Content-Type: application/sdp Content-Length: 262
Empty Line				
Message Body	Body			

Figure 3 SIP Message Layout

2.5 SIP Request Methods

SIP methods are used to initiate calls, transfer information about endpoints and terminate calls.

Start-line of a SIP request is Request-Line which differentiates it from SIP response. Request-Line comprises of

- a method i.e. Invite, Register, Ack and Cancel etc
- a Request URI
- SIP version

All these are separated by SP (single space) character. The Request-Line is terminated by CR/LF.

Method	Request URI	SIP version

There are six methods used in SIP.

- REGISTER registers contact information
- INVITE, ACK, and CANCEL sets up communications.
- BYE ends the communications.
- OPTIONS inquire servers regarding their abilities.

Invite	Ack	Bye	Cancel	Options	Register

Request URI is a SIP or SIPS URI. It specifies the service/user to which this request is intended to be sent. SIP Version: identifies the version of SIP and is used in both request and response.

2.6 Status Codes/Responses in SIP

SIP responses are the answers to SIP requests. Responses are identified through a 3-digits number. Start-line of a SIP response is status-line which distinguishes it from SIP requests. Status-Line comprises of

SIP version

status-code of 3-digit

Reason Phrase (textual phrase associated with status code)

Each element is separated by a SP character. No CR or LF is permitted apart from the final CRLF string.

SIP version	Status code	Reason phrase

SIP version identifies SIP version used. 3-digit status codes classify SIP responses. Reason phrase gives a short textual explanation of the status code.

The type of response is identified by first digit of Status Code. Second and third digits do not have any classification role. Six different values for first digit of status code are supported by SIP version 2.0.

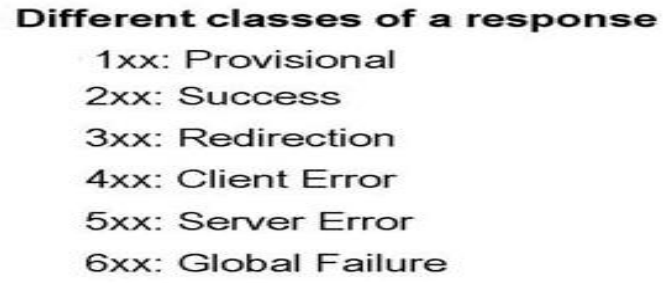

Figure 4 SIP Status Codes and corresponding Response classes

2.7 SIP Message Headers

```
INVITE sip:bob@biloxi.com SIP/2.0
Via: SIP/2.0/UDP pc33.atlanta.com;branch=z9hG4bK776asdhds
Max-Forwards: 70
To: Bob <sip:bob@biloxi.com>
From: Alice <sip:alice@atlanta.com>;tag=1928301774
Call-ID: a84b4c76e66710@pc33.atlanta.com
CSeq: 314159 INVITE
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 142
```

Figure 5 SIP Message Header

```
Session Initiation Protocol
  Request-Line: INVITE sip:011923218282002@sip.velocitydial.com SIP/2.0
  Message Header
    Via: SIP/2.0/UDP 192.168.66.6:55960;branch=z9hG4bK-d8754z-df2f791cde60ca29-1---d8754z-;rport
    Max-Forwards: 70
    Contact: <sip:77583779@203.99.191.76:55960>
    To: "011923218282002"<sip:011923218282002@sip.velocitydial.com>
    From: "MTA"<sip:77583779@sip.velocitydial.com>;tag=a859b659
    Call-ID: MGVmY2ZhNmFhNDhiODZlNTk1Y2JhNDlhOTA1MzZmYmE.
    CSeq: 1 INVITE
    Allow: INVITE, ACK, CANCEL, OPTIONS, BYE, REFER, NOTIFY, MESSAGE, SUBSCRIBE, INFO
    Content-Type: application/sdp
    User-Agent: X-Lite release 1104o stamp 56125
    Content-Length: 262
  Message Body
```

Figure 6 SIP Message Header fields opened in Wirshark

2.8 SIP Call Flow

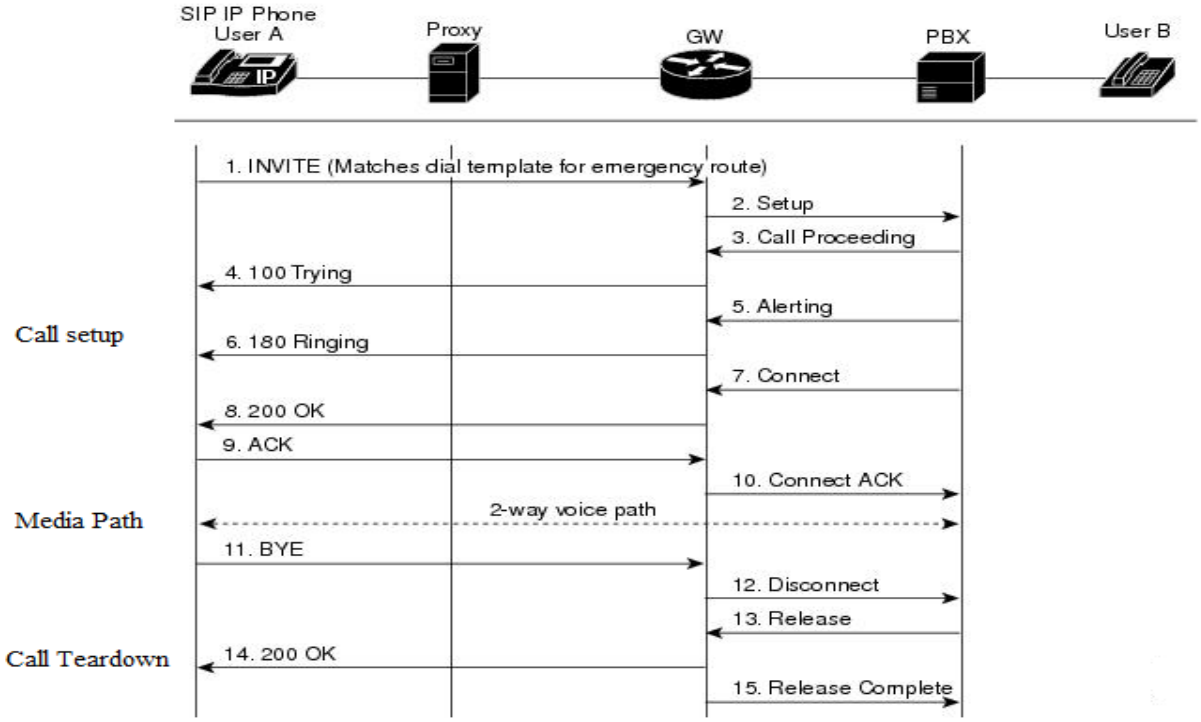

Figure 7 SIP Call Flow using SIP Requests and Responses

3. PREVIOUS WORK

Terminating illegal VoIP call has turned into a major problem for several countries. Governments need to develop and implement mechanisms that will permit them to keep an eye on network traffic including VoIP usage. An efficient Monitoring System is needed to monitor and limit the Illegal/Grey traffic. Internet traffic problem has been studied extensively by using various techniques to identify illegal Voice traffic. Different researchers and commercial organizations had developed various solutions for identification of VoIP communications. To overcome the limitations of previous techniques it is essential to develop an efficient methodology that accurately detects VoIP applications.

Most commonly used and simple method called as Port-based analysis is used to detect P2P users in network traffic. The concept implies such that majority of the P2P applications use default ports on which they work. VoIP uses nonstandard ports for communication. To avoid detection voice packets accomplish communication via dynamic ports. Furthermore, various applications hide their functional ports and pretend they are using well-known application ports for example http port number 80. That is why when detecting applications using ports numbers, there is a greater possibility to get false positives. Hence these are the main reasons which make port based analysis less effective [2]. Protocol based analysis emphasis on packet contents and provides alerts only on an exact match. VoIP developers also encrypt the traffic to avoid detection, making the analysis much difficult [2].

QOS specific approach states that "It is essential to apply controls to a traffic for example flow control and filtering by precisely identifying the Legal VoIP traffic from the forbidden (illegal) traffic to provide a reliable VoIP service." The method is used to analyze the packet exchange patterns incorporating the exchange of signaling messages (flow) and exchange of media data (interaction). However this method is limited to Skype, Netmeeting and SIP phone VoIP [3].

VoIP monitoring based on standard and proprietary protocols was also carried out. Two open source applications named Ntop (a portable traffic monitoring and measurement tool) and nprobe (captures full speed packets with very less (< 1%) packet loss) were developed for this purpose. But these methods are again protocol specific [4]. Detection of Skype VoIP is based on analysis and monitoring of the packet exchange patterns which are limited to previous version of Skype. A lot of research has been done on Skype specific protocol. This is only limited to Skype detection and cannot be applied for detection and monitoring of other protocols used for VoIP [5]. VoIP identification based on flow-level characteristics, for example packet rate, packet inter-arrival time and packet lengths is quite efficient. As these characteristics cannot be altered by applications so detection based on these features is far more effective. The algorithm effectively identifies the presence of voice data on the network without the knowledge of the protocol used [6]. The scheme based on human conversations pattern for VoIP flow identification has been worked upon. This scheme is useful because flow detection relies on human conversations instead of packet timings. Thus, the whole packet stream detection is based on a sequence of voice activities [7]. Research work has been done on VoIP CDRs from different aspects. CDRs are collected basically for billing but they are used for a variety of other purposes. Originally CDRs are collected in PSTN and they represent the most widely deployed form of telephone logs. VoIP CDRs contain plenty of information and are very valuable for monitoring, security and management of the systems. The intention of the research work [8] is to provide support to the researchers already working on telephone logs. The way CDRs are gathered as well as the formats, may differ among various operators. Without a common format of data logs used for security, traffic management, and other purposes can be cumbersome

and can result in misjudged conclusions. An overview has been provided how to handle and deal with huge amount of CDRs from various operators. In another approach [9], the author has addressed a new method of achieving optimized network planning for VoIP services by using historical data. The work shows how to obtain a set of service and use the clusters based on calling behaviors by applying an algorithm named as Simple Expectation Maximization (EM). Successful evaluation of this methodology with real data extracts useful knowledge that can result in an improved network planning. The real-time analysis of calling log is one of the most useful methods to detect problems. The characters of Deterministic Dendritic Cell Algorithm (dDCA) are used for the real-time analysis of large time-ordered data to detect faults in Internet Protocol Private Branch Exchange (IP-PBX). dDCA has very low CPU processing requirements and does not need extensive training period and it has good results for this kind of data [10]. The increased number of services being offered and available functionality by Telecom companies are resulting in an ever growing volume of CDRs. Most of the services including pre-paid, the CDRs are to be analyzed in real-time due to several reasons like analysis of predicting subscriber usage and also preventing fraudulent activity. An approach to address the challenges which are related to real-time data transform and load (RTL) of CDR data for supporting the operational and business intelligence needs of telecommunication services have also been studied [11]. To survive in the telecommunication market, the telecom companies are not only competing depending on the price but they are also expanding their services based on the customer's needs through the use of CDR. This work presents that an Online Analytical Processing (OLAP) system will be useful for telecommunication companies to get better insight of its customer's behavior and therefore can improve its marketing campaigns and pricing strategies [12]. A method to prevent VoIP attacks through the analysis of call logs has been delivered. A structured privacy engineering approach is used when analyzing call logs to ensure privacy of the call participants. The proposed technique prevents VoIP attacks on VoIP systems and preserves the privacy of the call participants as well [13]. The method used for fraud analysis is based on the Naive Bayes model which reveals that useable CDR lies in rejection proportion [14]. Two algorithms based on the analysis of CDRs for monitoring the Quality of Service (QoS) and based on this monitoring to detect the failures of voice communication have also been proposed [15].

4. METHODOLOGY

To provide a cost effective solution for precise detection of illegal VoIP, an efficient methodology is needed. The previous methods of detection were not always accurate, as the attributes on which they base either got changed or masked by the applications. In addition, these solutions are very expensive to be implemented in networks. No such technique has been invented until now that had identified illegal VoIP using CDRs.

The method discussed in this paper for illegal VoIP traffic detection uses the pcap files. These files are analyzed, filtered and call logs are generated by extracting the required parameters for CDRs. These CDRs are then used to detect either the user is white or grey. The method entails generation of VoIP calls using soft phones i.e. 3CX and X-lite. VoIP packets are captured using Wireshark tool and saved in a pcap.

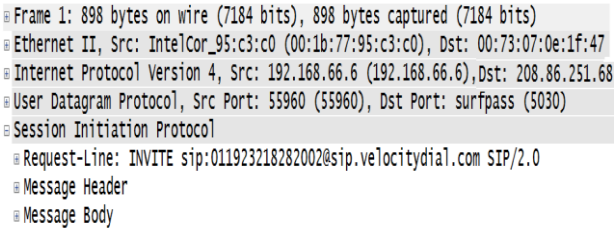

Figure 8 SIP Packet opened in Wireshark

Using pcap programming extract, the required parameters from the saved packets header up to IP layer i.e. source and destination MAC addresses (data link layer), destination and source IP addresses (IP layer) and destination and source port numbers (Transport layer) are read as shown below:

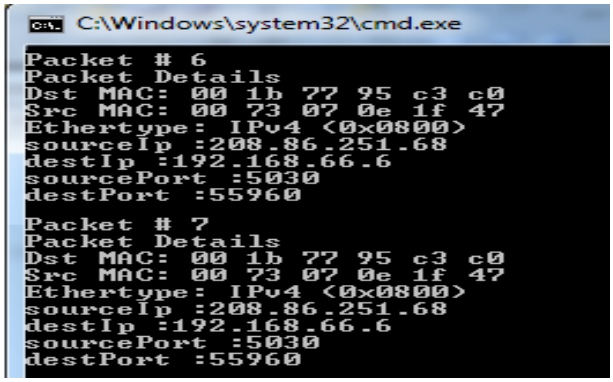

Figure 9 Parameters Extracted from a SIP packet up to IP layer using PCAP programming

At the application layer there is no standard technique or algorithm to detect protocols or identify traffic. An algorithm to identify VoIP protocols and detect the illegal traffic has been proposed. The main focus of interest is SIP which is the widely used VoIP protocol. SIP is an application layer (layer7) signaling protocol that sets, modifies and tears down multimedia and all the communication sessions. SIP is also a text based protocol just like HTTP. For detection of HTTP protocol some algorithms already exist. The proposed algorithm for identification of SIP will work almost like that of HTTP as both have almost same headers and messages format. The data from IP layer is encapsulated in UDP header and extracted from payload. As there is no standard protocol number which can be used for SIP or VoIP filtering so there arises a need for a standard method.

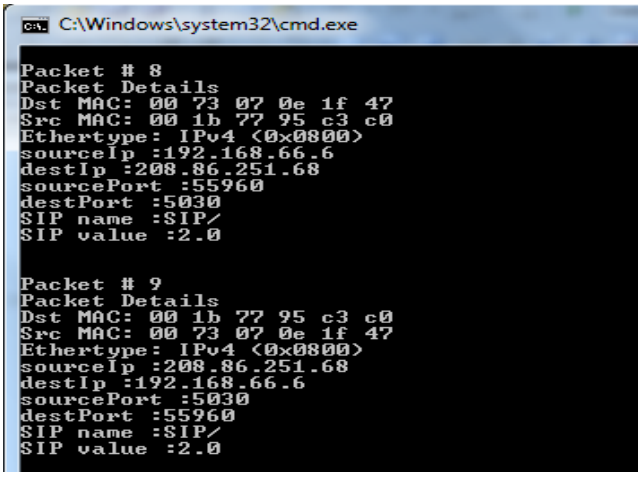

Figure 10 Identification of SIP Packets

Based on the analyses and simulation carried out, a general algorithm is designed for the detection of grey VoIP traffic using CDRs which is summarized below.

1. Open .pcap file
2. Read all packets
3. Identify VoIP Packets (SIP/H.323/IAX2)
4. Filter VoIP packets
5. Open and read headers of VoIP packets
6. Identify call parameters
7. Correlate packets and generate CDRs
8. Save CDRs
9. Compare IPs with the list
10. Legal/Illegal

The flow diagram of the algorithm is shown below which shows the clearer picture that how this method is achieved.

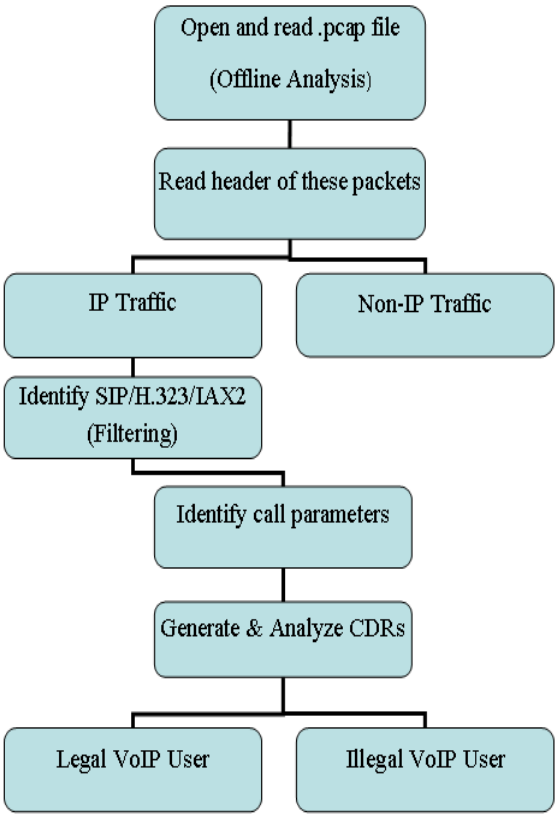

Figure 11 Flow Chart

4.1 Algorithm

PT: Packet Type

PD: Payload

If (PT = = UDP) // Filter UDP Packets

Open (PD)

If [{(request-line == '1')&&(request-line-end=='SIP/')}|| {(status-line== '1') && (status-line-start== 'SIP/')}] // a sip packet is detected and filtered if this is true.

//if the above condition is satisfied the following statements are processed

{

Read and save ('From' field) //Extract calling number

Read and save ('To' field) // Extract called number

Read and save ('source IP') // Extract source IP

Read and save ('destination IP') // Extract destination IP

Read and save ('date') // Extract date of call

Read and save ('time') // Extract time of call

If (Call ID==same number for no. of packets) // correlates packets to generate CDRs

{Save (CDR.txt)}

Open (legal IP list.text) //compare IPs

If (ip==match)

legal

else

illegal

}

Else

Return=0

5. CONCLUSION

This paper is based on analysis of pcap files captured through Wireshark. An algorithm is proposed for pcap programming to filter SIP packets. Call parameters are extracted to produced call logs (CDRs).The generated call logs are used to classify VoIP traffic as legal or illegal.

The Algorithm can be extended to other VoIP protocols like H.323 and IAX2.This method can also be used to develop/implement software which can be achieved by pacap programming to capture the packets from the nodes without using the tool like Wireshark and do offline analysis for real time data.

6. ACKNOWLEDGEMENT

We are thankful to Engr. Muhammad Taimur Arshad who is the Deputy Director Vigilance Cell PTA, for his encouragement and guidance throughout the thesis. He has been helping throughout this work and we really appreciate the motivation provided by him.

7. REFERENCES

[1] Article, Available: http://telecompk.net

[2] Yiming Gong, Identifying P2P users using traffic analysis available: http://www.symantec.com/connect/articles/identifying-p2p-users-using-traffic-analysis

[3] Kitamura tsutomu (Nec Corp.), Shizuno takayuki (Nec Corp.), Okabe toshiya (Nec Corp.), Tani hideaki (Nec Corp.), (2006) "Traffic Identification for Dependable VoIP", NEC Technical Journal, VOL.1;NO.3;PAGE.17-20

[4] Luca Deri, (2009), Open Source VoIP Traffic Monitoring, Available: http://luca.ntop.org/.

[5] Baset, S. A.; Schulzrinne, H. G., (2006) "An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol", in Proc. INFOCOM 2006. 25th IEEE International Conference on Computer Communications.

[6] Uzma Aslam Khan, Fauzia Idrees , (2008) " A Generic Technique for Voice over Internet Protocol (VoIP)

Traffic Detection" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2

[7] Kuan-Ta Chen , Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei1, (2008) "Detecting VoIP Traffic Based on Human Conversation Pattern" Principles, Systems and Applications of IP Telecommunications. Services and Security for Next Generation Networks, Pages 280 - 295, Springer-Verlag Berlin, Heidelberg ©2008

[8] Nico d'Heureuse and Sandra Tartarelli, and Saverio Niccolini , (2010)"Lessons Learned on the Usage of Call Logs for Security and Management in IP Telephony" IEEE Communications Magazine, Volume 48 Issue 12

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

[10] Lin Lu, Yiwen Liang, Chao Yang,Shiwei Song, (2011) "Detecting Faults in IP-PBX with Deterministic Dendritic Cell Algorithm",Advances in Information Sciences and Service Sciences(AISS), Vol. 3, No. 11, pp. 457 ~ 465

[11] Munir Cochinwala, Euthimios Panagos,(2009) "Near Real-time Call Detail Record ETL Flows" Conference on Business Intelligence for the Real-Time Enterprises - BIRTE , pp. 142-154, Telcordia Applied Research

[12] Dragana Ćamilović , Dragana Bečejski-Vujaklija and Nataša Gospić, (2009) "A Call Detail Records Data Mart: Data Modelling and OLAP Analysis" Journal Article ,2009, Comput. Sci. Inf. Syst 01/2009; 6:87-110

[13] Stef Hofbauer, S.; Quirchmayr, G.; Beckers, K., (2012) "A privacy preserving approach to Call Detail Records analysis in VoIP systems", IEEE 2012 Seventh International Conference on Availability, Reliability and Security

[14] Khairul Nizam Baharim, Mohd. Shafri Kamaruddin, Faeizah Jusof, (2008) "Leveraging Missing Values in Call Detail Record Using Naïve Bayes for Fraud Analysis", ICOIN 2008. International Conference on Information Networking

[15] Gean D. Breda, Leonardo de S. Mendes, (2006) "QoS monitoring and Failure Detection using CDRs" IEEE Telecommunications Symposium, 2006 International, 2006, Page(s): 243 - 248