International Journal of Computer Science Trends and Technology (IJCST) – Volume 6 Issue 2, Mar - Apr 2018

RESEARCH ARTICLE OPEN ACCESS

A Review of Techniques for VOIP Classification

Anjali, Max Bhatia
Department of Computer Science
Lovely Professional University
Jalandhar - India

ABSTRACT

Communications has evolved in current years which have given the raise of application for the voice over IP (VoIP), the voice communication on the internet the service is easily provided. It provides both standards-based proprietary and VoIP services alternative such as which earns a huge popularity such as Google, Skype, since they were very easy to use and they cost low. On the other hand, for different reasons, like custom strategy enforcement, assurance of quality, safety or particular business benefit, it becomes very important to observe and identify the VoIP traffic presence on data network. The major purpose of this review paper is to provide a survey of techniques for categorization of traffic for VoIP protocol. This survey explains the most important techniques and problems identified VOIP traffic analysis. Finally, the survey helps us to draw the reader's concern to open research topic in the fraction of traffic analysis and function recognition and make a number of final remarks.

Keywords:-*machine learning; traffic classification; VOIP*

I. INTRODUCTION

VoIP stands for Voice over Internet Protocol that makes use of internet or additional data network slightly than using usual 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 [3]. The VoIP be a solid expertise to allow people to communicate through voice via IP protocol instead of telephone lines. The property standard, high cost tag, partial addition with presented telephony environments are some of the factors that have assigned this technology in a niche market. Now a day's situation has been changed due to advent of after risk as well as low-cost VoIP telephone adapters open source tools. This has become easy and ordinary for internet provider to give their clients VoIP calls at very low price, if some in addition to usual xDSL connectivity. Improvement in VoIP additionally coordinates the development of focalized systems that maintain both video and voice administrations not accessible through usual PSTN. Although VoIP is minimal effort or about free ability still unique telecom administrator attempt to cover up VoIP movement purposely to maintain a strategic distance from acknowledgment and run away from excise i.e. Access Promotion Charge (APC) through varying unusual parameter into VoIP packet [4]. The key rewards of VoIP are flexibility and ease. Accordingly, amount of PSTN provider have previously fulfilled or are inside the method of transitioning as of circuit-switched networks to VoIP-friendly packet-switched backbone. Suitable to these advantages, VoIP have see quick uptake into equally the endeavor as well as customer market.

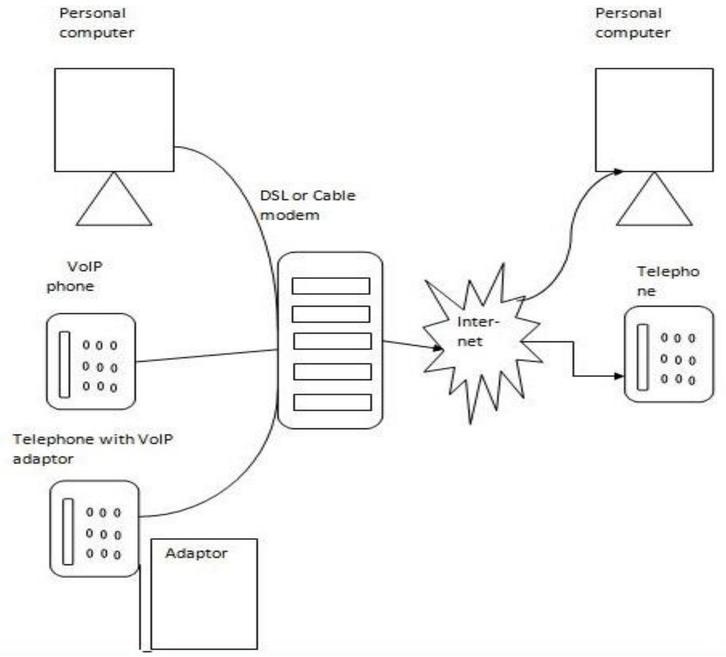

Fig. 1: Working of VoIP

The above shown Fig.1 is the diagram of VoIP this is showing the working of VoIP. Usual VoIP protocol such as SIP and H.323 are extremely accepted in the carrier atmosphere and in several additional fields not partial to VoIP such as messenger and chat. During addition to these standards-based applications, there are other applications such as Skype or voip stunt that as an alternative are base on proprietary communication protocols, codec's and further hybrid application partially base on open standard such as Google talk and gizmo [5]. The network traffic is need to be monitored and analyzed by law enforcement agencies to avoid frauds. Provider wants to categorize the kind of traffic transported during their network mainly VoIP calls. The biggest branch of usual earnings basis of provider is VoIP that's why main focus is on it. That is the reason they get less earnings as of their most important or mainly business endeavor consumers because a most important element of the traffic go invisible or uncharged. To constrain dark movement there is a require to expand technique to examine IP traffic, recognize or identify grey VoIP calls and after that basically building block them and accuse them.

II. RELATED WORK

There are a numeral of studies have been acknowledged by a great deal of analysts in this field. These sections provide general idea of several considerable research papers. In [6], author analyzes VoIP function with OPNET via unusual routing protocol, although they had not considered the consequence of queuing technique in the analysis method. In [7] author evaluate VoIP function on the remote LAN (802.11)via OPNET, though, they utilize only four levels of fundamental concern to categorize the traffic as QoS assess for a 2 minute time distance. At the same time as the utilize of classifiers for study or recognition of network application, into exacting VoIP traffic is not a current improvement, they represent a dynamic research field thank toward the current development into conditions of the recognized mechanism so as to maintain these technique, with as well suitable toward the nonstop development of VoIP protocols or application. While a consequence, here is several unusual technique or recognition mechanism developed via educational or business organization.

III. VoIP TRAFFIC CLASSIFICATION TECHNIQUES

Several different techniques have been planned for VoIP traffic categorization. From the perspective of feature extraction, VoIP traffic categorization techniques can be separated into two categories: VoIP traffic classification by payload inspection and VoIP traffic classification without payload inspection. There has been a large amount current effort into the field of VoIP traffic categorization. This part will study the unusual technique accessible in the literature.

A. *VoIP Traffic classification by payload inspection*

This type of technique is based on inspecting the payload data in multiple network layers [8]. In this way, the payload data is checked in a bitwise manner to find specific bit sequence that represents the signature of a certain protocol. The protocol type of a packet can be accurately identified if such a bit sequence is found. Practically, VoIP traffic payload-based technique is usually exploited to detect abnormal traffic and VoIP traffic flows [9]. Although having particularly high accuracy, payload-based technique has many disadvantages because (1) it is a complicated operation that demands a large amount of computation and storage capacity, and usually specific hardware is necessary for inspecting the signatures in the payload data by monitoring the traffic passively or initiatively [10]; (2) it can fail to identify the protocol type if the tunnel or encryption techniques are applied to traffic transmission, and it may become completely ineffective if the new applications with new signatures are encountered; and (3) it has a very serious problem that is related to user privacy violation and laws issues.

B. *VoIP Traffic classification without payload inspection*

In contrast with payload-based technique, there are many techniques that classify VoIP traffic without inspecting the payload data. Early on technique examines the data in the network layer and transport layer, even as additional techniques analyze the transmission patterns of the applications.

1) *Port based technique*

The easiest method for distinguish network application is to study the pre-characterized ports that the application utilize. This method is base on top of track an arrangement of static TCP/IP port numbers use through a function and protocol. These methods moreover apply toward VoIP base application. Port-based approaches typically classify Internet traffic according to the 16-bit port numbers in transport layer header, which are utilized by computers to identify different communication sessions. The mappings between port numbers and services are synchronized through the Internet Assigned Numbers Authority (IANA) [11] Early times on researchers have relied on the simple concept that all services utilize the standard port numbers that are registered by IANA for traffic transmission [12]. However, this method proved to be inefficient because the definitive mapping is no longer effective. More and more newly emerging services violate the port number assignments of IANA. Furthermore, standard services can cover up themselves by use port numbers that are typically connected with unusual services, to circumvent the access control restrictions of proxies and firewalls [13]. Thus, port-based approaches can fail to categorize the traffic if no corresponding port number is establish or generate false outcome but an function use a further well-known port number. A research study conducted by Moore et al. [14] reported to facilitate the overall precision achieved by port-based approach is only among 50% or 70%.

2) *Statistical classification technique*

This method depend on packet or flow level assets of VoIP activity, like packet size, TCP or UDP ports use, etc; which can be utilize independently or cooperatively used for computation of statistical systems, for example, normal, variance or likelihood density function.. Freire et al. in [15] or [16] planned method for classify VoIP calls secreted into Web traffic through analyze a number of properties of network information, these are: size of Web asks for or reaction, numeral of per page requirements, inter-arrival point among requirements or rescue time of page. They assess their technique on VoIP information of Google-Talk and Skype which was gather from ISP and college connects or accomplish help rates to remember concerning 90 % for VoIP calls or 100 % for VoIP calls hid in Web traffic. VoIP administrations return high entropy esteem though general client– server traffic returns continually littler esteem. Sun and Chen [17] planned a new method appropriate base on C4.5 decision tree to identify function linked through a TCP flow, by means of two uniqueness: the ACK-Len ab and ACK-Lenba; which are the information measure first sent by communicate party constantly. The tentative results achieve average true positive and false positive rate of 97.22 or 2.78 % in a specific order. The arranged method does not matter complex statistical highlights of traffic or machine learning calculations. The outcomes of these studies have inspired researchers to take a new classification method for statistical properties. To achieve the actual categorization based on statistical characteristics, classifiers need to employ data mining techniques, specifically ML algorithms, because they want to contract among different traffic pattern from huge datasets. ML algorithms are very lightweight and less computationally expensive than payload-based classification techniques, because they do not depend on DPI but rather utilize the information from flow-level analysis. The effectiveness of the classifier in statistical classification depends on the features extracted from the flow, which require extensive knowledge due to their complexity. However, these techniques outperform payload-based techniques since they do not deal with packet contents, and thus can analyze encrypted traffic without any difficulty. The earlier methods are partial through their belief on top of the conditional semantics for the data gather during profound assessment of packet substance (payload and port numbers). Newer approach relies on top of traffic's statistical distinctiveness toward classify the function. The statement essential some method is so as to traffic on the network layer have statistical property (for example, the circulation of flow length, flow unused time, or packet inter-arrival instance and packet length) so as to are exclusive used for assured module of application or permit unusual cause application to be real illustrious as of each other. The correlation among the set of traffic or its experiential statistical property have been illustrious inside [27] (somewhere the author analyze or construct experimental model of link distinctiveness – for example bytes, duration, arrival periodicity - for a numeral of particular TCP application), and inside [28] (wherever the author analyzed Internet chat system via focus on top of the distinctiveness of the traffic in conditions of flow interval, packet inter-arrival point and packet size or byte profile).The outcome of these mechanism have encouraged new classification technique base on top of traffic flow statistical property. The require toward contract through traffic pattern, huge datasets and multi-dimensional spaces of flow or packet attribute it results intended for the preface of ML methods in this field.

3) Heuristic-based methods

Heuristic technique classify the traffic via observe the behavioral pattern of VoIP traffic utilize pre-characterized set of heuristics, for example, hosts substitute together as receiver or sender, numeral of associations prepared through host, numeral of different addresses and ports a host be associated toward, hosts use together TCP or UDP used for announcement, and so on. The groups of heuristics are analyzed in sequence and the packets and flows are classify while belong toward a exacting group depending ahead the outcome obtain. Here are several study to compose make use of heuristics to classify VoIP traffic. Reddy and Hota [18] planned a novel group of heuristics to classify VoIP host base on top of its link pattern or they do not need some payload signatures. Datasets used for sensible into environment and consisted of application i.e. Http, FTP, Dropbox, SMTP, eMule, Frostwire, Skype, uTorrent and Vuze. Author confirmed their techniques in real time or just 0.2 % of VoIP traffic remain uncategorized .Because their methods consisted of minimum heuristics, it can be use intended for concurrent classification; other than it can only classify large VoIP application relatively than unusual VoIP application. Bashir et al. [19] planned an approach base upon heuristics to classify BitTorrent actions via Netflow report through observe 3 most important segment of traffic: a) traffic since peers contact by the use of DHT, b) TCP traffic since peers contact through trackers or c) UDP traffic since peers contact through trackers. The technique be experienced on top of 5 genuine life datasets have combination of application consisting of BitTorrent, Skype, SopCast or PPStream. The tentative outcome achieve the byte precision range as of 91.3 to 95.4 % in identify BitTorrent activity.

IV. MACHINE LEARNING BASED APPROACHES

The machine learning have a broad collection of application, with look for engines, medical diagnosis, text or writing identification, image broadcast, load forecasting, marketing and sales diagnosis, etc. Machine learning methods usually consists of 2 parts: model building and classification. The model is first build by training data. These models are after that inputted into a classifier after that classifies a data group.

Machine learning (ML) methods can be separated interested in the category of unsupervised and supervised. McGregor et al. assume capability of using an unsupervised method to cluster flows based upon connection-level (i.e., transport layer) statistics to identify traffic [20]. ML-based approach is a extremely capable technique for VoIP traffic categorization. Further, it do not rely on top of payload inspection, which has drawbacks of low accuracy and high computational overhead. On the other hand, it exploits many TLS features that are directly extracted from traffic, which can provide better classification results than port-based technique. To calculate the efficiency of the algorithms 3 methods were used: precision, recall, and overall accuracy. They procedures include extensively use in the data mining literature to assess data cluster algorithms [21]. For a particular class, the numerals of suitably identified objects are referred to as the True Positives. The numeral of substances incorrectly classified because a class is referred to the False Positives. Numeral of objects from a group of class so they are falsely label like a further class is referred to the False Negatives.

- Precision is the proportion of relative amount of True Positives to True and False Positives. These determine the correction of number of classified objects.

$$\text{Precision} = \frac{TP}{TP+FP}$$

- Recall is the proportion of relative amount of True Positives to the number of True Positives and False Negatives. These verify the number of objects in a class which are misclassified as something else.

$$\text{Recall} = \frac{TP}{TP+FN}$$

- Overall accuracy is defined as the sum of all True Positives to the computation of all the True and False Positives for all classes. This procedures the general correctness of the classifier. Note that precision and recall are per-class procedures.

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)}$$

Where n is the numeral of classes. Precision and recall are interrelated with one another [29,30,31,32]. Precision for other classes are caused if the recall for one class is lesser because the algorithms used here always categorize the objects into a class. The overall accuracy is associated with precision where the average precision of all classes is measures.The limitation of port numbers has motivated the consideration of exploiting new TLS features to classify Internet traffic. This consideration is based on the assumption that application usually sends data in several type of pattern, which can be utilized as a communication way to differentiate network links. It is notable that port numbers are no longer exclusively utilized to classify VoIP traffic, although they are still important features for network communication. Generally, these TLS features can be obtained by directly extracting the information from network layer headers. In practice, traffic flows can be collected by built-in network devices. Because 1000 MBs of traffic lead to TLS information in the order of kBs, ML-based approaches are extra efficient for traffic classification in high-speed network than payload-based techniques. Among an increase numeral of TLS features to be analyzed, substantial effort has been made on the function of ML algorithms to TLS features for VoIP traffic categorization [22]. In 2004, Roughan et al. [23] and McGregor et al. [24] first extract TLS features, e.g., the packet range statistics and flow duration, for traffic classification. In 2005, Moore et al. [25] accessible 248 TLS features to fully describe the Internet traffic characteristics. In recent years, more and more TLS features were presented for characterizing network transmission behavior [26]. So extreme, hundreds of TLS features can be able to utilize for traffic classification.

V. CONCLUSION

The categorization of these applications depends on the network traffic data and it is essential matter that concerns with the entities which includes the telecommunications operative or service providers of VoIP, for some causes: like safety, service value or policies for business and network. In this, the review paper present general idea of current techniques and both the legacy in which helps in the recognition and categorization of VoIP based application. Legacy methods consist of procedure depends on port or protocol recognition, in which the precise level has decreased due to the individual development of new VoIP applications and protocols and evasion methods change. In this paper, we have identified the key challenges associated with VOIP traffic. An important task of machine learning is data classification. It is identified with develop computer programs ready to gain from label data set and, in this way, to forecast unlabeled instances. Because of the huge amount of applications, numerous data categorization systems have been developed.

REFERENCES

[1] JeroenHoebeke, Ingrid Moerman, Bart Dhoedt and Piet Demeester, "An Overview of Mobile Ad Hoc Networks: Applications

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

[3] Uzma Anwar, GhulamShabbir, 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] Mohammed S. Islam, Adnan Riaz and Mohammed Tarique: Performance analysis of routing protocols of mobile ad hoc networks for VoIP applications, In Journal of Selected Areas in Telecommunications, pp. 26-33, June 2012.

[7]A. Al-Naamany, H. Bourdoucen and W. Al-Menthari: Modeling and Simulation of Quality of Service in VoIP Wireless LAN, In Journal of Computing and Information Technology - CIT 16,4, pp.131–142, 2008.

[8] T. Karagiannis, K. Papagiannaki, M. Faloutsos, BLINC: Multilevel traffic classification in the dark, SIGCOMM Computer

Communication Review 35 (4) (2005) 229 240..

[9] T. Karagiannis, A. Broido, M. Faloutsos, K.C. Claffy, Transport layer identification of P2P traffic, in: 4th ACM SIGCOMM Conference on Internet Measurement, 2004..

[10] S. Yi, B.-K. Kim, J. Oh, J. Jang, G. Kesidis, C.R. Das, Memory-efficient content filtering hardware for high-speed intrusion detection systems, in: ACM Symposium on Applied Computing, 2007, pp. 264-269.

[11] The Internet Assigned Numbers Authority, IANA, [Online], Available: http://www.iana.org/.

[12] D. Moore, K. Keys, R. Koga, E. Lagache, K.C. Claffy, The Coral Reef software suite as a tool for system and network administrators, in: 15th USENIX Conference on System Administration, LISA01, pp. 133-144.

[13] M. Crotti, F. Gringoli, P. Pelosato, L. Salgarelli, A statistical approach to IP-level classification of network traffic, in: IEEE International Conference on Communications, ICC06, 2006, pp. 170-1765

[14] A.W. Moore, D. Zuev, Internet traffic classification using Bayesian analysis techniques, in: Proc. ACM SIGMETRICS International

[15] Freire EP, Ziviani A, Salles RM (2008) Detecting Skype flows in web traffic. In: Proceedings of the IEEE network operations and management symposium (NOMS 2008), IEEE, pp. 89–96

[16] Freire EP, Ziviani A, Salles RM (2008) Detecting VoIP calls hidden in web traffic. IEEE Trans Netw Serv Manag 5(4):204–214

[17]SunM, ChenJ (2011) Research of the traffic characteristics for the real time online traffic classification. J China Univ Posts Telecommun (Elsevier) 18(3):92–98

[18] Reddy JM, Hota C (2015) Heuristic-based real-time P2P traffic identification. In IEEE international conference on emerging information technology and engineering solutions (EITES), Pune, pp. 38–43

[19] Bashir A, Huang C, Nandy B, Seddigh N (2013) Classifying P2P activity in netflow records: a case study on BitTorrent. In IEEE International Conference on Communications (ICC), Budapest, pp. 3018–3023

[20] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, “Flow Clustering Using Machine Learning Techniques,” in PAM 2004, Antibes Juan-lesPins, France, April 19-20, 2004.

[21] I. Witten and E. Frank, (2005) Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[22] R. Yuan, Z. Li, X. Guan, L. Xu, An svm-based machine learning method for accurate internet traffic classification, Information

[23] M. Roughan, S. Sen, O. Spatscheck, N. Duffield, Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification, in: 4th ACM SIGCOMM Conference on Internet Measurement, IMC04, 2004, pp. 135-148.

[24] A. McGregor, M. Hall, P. Lorier, J. Bruuskill, Flow clustering using machine learning techniques, in: Proceedings of the 5th Passive

[25] A. Moore, K. Papagiannaki, toward the accurate identification of network applications, in: Passive and Active Network Measurement, 2005, pp. 41–54.

[26] W. Li, M. Canini, A.W. Moore, Efficient application identification and the temporal and spatial stability of classification schema,

Computer Networks 53 (6) (2009) 790–809.

[27] V. Paxson, "Empirically derived analytic models of wide-area TCP connections," IEEE/ACM Trans. Networking, vol. 2, no. 4, pp. 316–336, 1994.

[28] C. Dewes, A. Wichmann, and A. Feldmann, "An analysis of Internet chat systems," in ACM/SIGCOMM Internet Measurement Conference 2003, Miami, Florida, USA, October 2003.

[29] Aditya Khamparia and Babita Pandey. 2017. A novel method of case representation and retrieval in CBR for e-learning. *Education and Information Technologies, Vol* 22, 1 (January 2017), 337-354. DOI: https://doi.org/10.1007/s10639-015-9447-8, 2017.

[30] Khamparia, A. & Pandey, B. Springer Plus 5: 446, 2016 https://doi.org/10.1186/s40064-016-2101-0.

[31] Khamparia, A. & Pandey, B. Educ Inf Technol 22: 3121. https://doi.org/10.1007/s10639-017-9574-5, 2017

[32] Aditya Khamparia and Babita Pandey. Knowledge and intelligent computing methods in e-learning. *Int. J. Technol. Enhanc. Learn.* 7, 3 (November 2015), 221-242. DOI=http://dx.doi.org/10.1504/IJTEL.2015.072810, 2015.