# Efficient Detection of VoIP Calls Hidden in Web Traffic

P.S. Abi[1]

[1] *Lecturer, Department of Computer science, University College of Engineering Arni.*
*psabi84@gmail.com*

***Abstract*** **- Network managers face nowadays a challenging problem to detect traffic from Skype. Skype can operate behind many firewalls and network proxies without user configuration. Behind restrictive firewalls, Skype uses Web TCP ports (80 or 443) as a fallback mechanism to delude firewalls and other network elements. We propose the adoption of nonextensive entropy to detect anomalies in network traffic within an Autonomous System (AS). We show that our approach based on chi-square test outperforms previous ones based on nonextensive entropy while providing enhanced flexibility, which is enabled by the possibility of the detection mechanism.**

***Index Terms*** **- P2P VoIP, Anomaly Detection, Entropy based Detection.**

## I. INTRODUCTION

Recent years, voice over IP (VoIP) applications have faced a huge increase in popularity, in particular those based on the peer-to-peer (P2P) communication paradigm for scalability purposes. In comparison with services from the traditional Public Switched Telephone Networks, which normally use a per-minute charge for long-distance calls, VoIP calls rely on services from the TCP/IP protocol stack provided by the Internet infrastructure, which in turn are usually charged with a fixed flat monthly fee disregarding the transmitted traffic volume [8]-[9].

Network anomalies denote significative and unusual changes in traffic patterns of one or multiple network links in an Autonomous System. We are interested in detecting volume anomalies in network traffic (i.e. significative changes in the volume of network traffic) using statistical detection methods, more specifically entropy-based methods [1].

Among the P2P VoIP applications that adopt the strategy of disguising their flows as Web traffic to delude firewalls and other network elements. Skype is a very popular VoIP application with a proprietary closed-source protocol and encrypted end-to-end traffic [2]. In order to optimize the use of network resources, very restrictive firewalls are commonly adopted by network managers in many organizations using port numbers to select some applications to receive priority treatment or to be blocked [3].

This paper, deals the problem of detecting VoIP calls hidden in Web traffic. Traffic from P2P-based VoIP systems is composed of signaling flows and the media flows. The former refers to the traffic to establish and maintain the overlay P2Pnetwork of the VoIP system as well as to the traffic to signal the call establishment and release.

We have investigated a method to detect Skype calls hidden in Web traffic using metrics taken from two Goodness-of-Fit tests, the entropy based detection and the chi-square $\chi 2$ value.

It is organized as follows: in section 2 entropy based anomaly detection is explained and in section 3, our proposed work is explained. In section 4, the conclusion is stated.

## II. ENTROPY-BASED ANOMALY DETECTION

In this paper, we refer to ingress and egress points of an AS for a given network traffic as origin and destination, respectively. We establish four basic traffic patterns by combining the classification of origin and destination as concentrated or dispersed, according to the entropy computation of network traffic at the PoPs [1]

- CC:Concentrated origin and concentrated destination;
- CD: Concentrated origin and dispersed destination;
- DC: Dispersed origin and concentrated destination;
- DD:Dispersed origin and dispersed destination.

These traffic patterns may represent volume anomalies in network traffic. For instance, a “CD” traffic pattern might represent a bulky multicast transmission, a “CC” traffic pattern might represent a DoS attack, and a “DC” traffic pattern might represent a DDoS attack.

### *A. Background on classic entropy*

Given a distribution of probabilities P {p1,p2,...,pN} with N elements, where $0 \leq pi \leq 1$and $\Sigma_i$ pi = 1, Shannon entropy HS is defined as [3]

$$H_S = -\sum_{i=1}^{N} p_i \log_2 p_i. \tag{1}$$

In our case, N identifies the number of PoPs in an AS and pi the probability of incoming/outcoming traffic at PoP i. On the one hand, the minimum entropy value

$$H_S^{\min} = 0$$

Indicates maximum flow concentration, i.e. considering the incoming traffic, all flows enter the AS through a single PoP. On the other hand, the maximum entropy value indicates maximum dispersion, i.e. for the same case of incoming traffic, flows are uniformly distributed among ingress PoPs with 1/N probability.

### *B. Nonextensive entropy*

In some fields of physics, such as mechanics and thermodynamics, microscopical phenomena may be statistically investigated to estimate their macroscopic properties. Such a study may be performed using Shannon entropy, also known in this context as extensive entropy. Nevertheless, some physical systems present many challenges if an approach based on extensive entropy is adopted to analyze them. These systems typically present characteristics such as long-range dependence (in both time and space) and fractal behavior. To deal with these systems, Tsallis proposes the concept of nonextensive entropy, generalizing the extensive Shannon entropy. Nonextensive entropy is defined as

$$H_q = \frac{1-\sum_{i=1}^{N} p_i{}^q}{q-1}. \tag{2}$$

Note that Eq. (2) with q → is equivalent 1 to Eq. (1). Tsallis entropy is actually a one-parameter generalization of the Shannon entropy, q being called the entropic parameter. The value of Tsallis entropy may range from, which represents maximum concentration, to Hmax q , which indicates maximum dispersion, where

$$H_q^{\max} = \frac{1-N^{1-q}}{q-1}. \tag{3}$$

In order to better understand the role of q, note that events with very high or very low probabilities associated with them do not exert much influence on the results of Shannon entropy. In contrast, in Tsallis entropy, in the case q>1, high probability events contribute more to the resulting entropy value than low probability events, whereas the opposite holds for q<1. Therefore, the variation of q modifies the relative contribution of a given event to the whole, enabling one to take a closer look onto events with higher or lower probabilities.

## III. PROPOSED WORK

The detection process can be subdivided in two steps. First, we define a HTTP workload model and capture real Web data to build empirical distributions of some relevant parameters. We then capture Web traffic with VoIP calls hidden in it, calculate the same relevant parameters for each flow and use metrics taken from two Goodness-of-fit tests to decide whether the computed parameters are compatible (or not) with the empirical distributions derived in the previous step, classifying each flow as legitimate Web traffic or not. Figure 1 illustrates this methodology in the general case of distinguishing "normal" Web traffic from "anomalous" traffic hidden in the Web traffic aggregate [2].

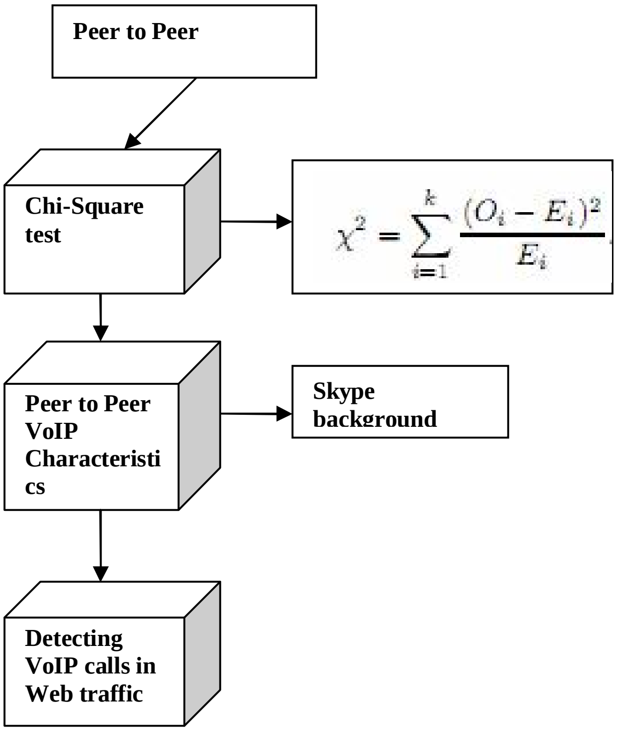
In this paper, we focus on the case of detecting anomalous traffic within the Web traffic aggregate caused by VoIP call from a P2P VoIP application such as Skype or Google Talk. Along this section, we briefly review the considered HTTP workload model for Web traffic and the adopted statistical tests.

### *A. HTTP Workload Model*

A first step towards HTTP traffic characterization is the selection of some HTTP relevant parameters. We are interested in finding VoIP calls hidden among Web flows. Since we avoid relying on program signatures or patterns that can be easily changed, we must define a model to evaluate Web "normal" behavior. This paper is mainly based on the particular subject of HTTP workload characterization [4]. Contrasting with related work in the characterization of a HTTP workload model, to distinguish anomalous flows within Web traffic, in particular those generated by ongoing VoIP calls to distinguish VoIP calls from legitimate Web traffic. This model has the following parameters:

- Web request size;
- Web response size;
- Inter-arrival time between requests;
- Number of requests per page;
- Page retrieval time.

## IV. ARCHITECTURE


Peer to Peer
Chi-Square test
$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
Peer to Peer VoIP Characteristi cs
Skype background
Detecting VoIP calls in Web traffic

The Web request size is the size in bytes of the HTTP request message. The Web response size is the size in bytes of the HTTP response, sent by some Web server [7]. The time interval between two consecutive requests of the same client for the same Web server is the inter arrival time between requests, if these requests are close enough to be considered parts of the same Web page.

A Web page may have one or many requests and it is important to identify page boundaries. Based on this information, we can compute parameters such as the number of requests per page, page retrieval times, and request inter arrival times. The number of requests per page is the number of HTTP request messages in the same Web page and the page retrieval time is the time elapsed from the first request to the last response. These parameters have been chosen because they are the most discriminant ones to distinguish VoIP calls from legitimate Web traffic [5].

*B. Chi-square test*

The chi-square (χ2) goodness-of-fit test, was first investigated by Karl Pearson in 1900. Basically, it tests a null hypothesis that the observed frequencies of some independent events follow a specified distribution. Suppose we have n observations from a population classified into k mutually exclusive classes and there is some theory or hypothesis which says that an observation falls into class i with probability pi (i =1,...,k), so, the number of events expected in class i is Ei = npi. If Oi is the number of events observed in class i, the chi-square statistic χ2 is the sum over all bins as given by [10]

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

A large value of the sum indicates that is rather unlikely that the Oi values are drawn from the population represented by the Ei.

## IV. CONCLUSION

In our proposed methodology we achieves performance of chi-square test around 100% detection rate with around 5% of false positives.

The experimental results are derived using the data gathered at a commercial Internet Service Provider. Both Skype and Google are used to build the evaluation datasets to verify that the proposed methodology is efficient. In case of a real time detection of VoIP calls, network managers could choose perform immediate action after detection, for example, blocking all traffic identified as VoIP from his network, giving such traffic is a differentiated treatment, or any policy-based management measure a network operator prefers.

As future work, we intend to build and evaluate an optimized version of our tool to perform real-time monitoring in network links. The proposed HTTP workload model can also be seen as a building block to the development of an automatic detection system of other kinds of non-HTTP flows hidden in Web traffic, such as P2P file sharing and media streaming applications. Clearly, to achieve this, further investigations are needed to identify the proper parameters in the HTTP workload model to detect each target disguised application.

## REFERENCES

[1] A. Ziviani, M. L. Monsores, P. S. S. Rodrigues, and A. T. A. Gomes, “Network anomaly detection using nonextensive entropy,” IEEE Commun. Lett., vol. 11, no. 12, pp. 1034–1036, Dec. 2007.

[2] Emanuel P. Freire, Artur Ziviani, and Ronaldo M. Salles, “Detecting VoIP Calls Hidden in Web Traffic” IEEE trans.on network and service management, vol. 5, no. 4, Dec 2008

[3] E. P. Freire, A. Ziviani, and R. M. Salles, “Detecting Skype flows in Web traffic,” in NOMS 2008: Proceedings of the 2008 IEEE/IFIP Network Operations and Management Symposium, 2008.

[4] B. A. Mah, “An empirical model of HTTP network traffic,” in INFOCOM ’97: Proc. 16th Joint Conference of the IEEE Computer and Communications Societies, 1997.

[5] E. P. Freire, A. Ziviani, and R. M. Salles, “On metrics to distinguish Skype flows from HTTP traffic,” in LANOMS 2007: Proc. 5th Latin American Network Operations and Management Symposium,Sept.2007.

[6] Y. J. Won, B.-C. Park, H.-T. Ju, M.-S. Kim, and J. W. Hong, “A hybrid approach for accurate application traffic identification,” in Proc. 4th IEEE/IFIP Workshop on End-to-End Monitoring Techniques and Services, Apr. 2006, pp. 1–8.

[7] H.-K. Choi and J. O. Limb, “A behavioral model of Web traffic,” in ICNP ’99: Proc. 7th International Conference on Network Protocols. IEEE Computer Society, 1999, pp. 327–334.

[8] S. Baset and H. Schulzrinne, “An analysis of the Skype peer-to-peer Internet telephony protocol,” in INFOCOM’06: Proc. 25th IEEE International Conference on Computer Communications, Apr. 2006.

[9] D. Bonfiglio, M. Mellia, M. Meo, N. Ritacca, and D. Rossi, “Tracking down Skype traffic,” in INFOCOM’08: Proc. 2008 IEEE INFOCOM, Apr. 2008

[10] N. Ye and Q. Chen, “An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems,” Quality and Reliability Engineering International, vol. 17, no. 2, pp. 105–112, 2001.