A Skype ML Datasets Validation and Detection Mechanism Using Machine Learning Approach

Hamza Awad Elkarim Hamza Ibrahim
Faculty of Electrical Engineering
Universiti Teknologi Malaysia-UTM
hamysra76@hotmail.com

Sulaiman Mohd Nor
Faculty of Electrical Engineering
Universiti Teknologi Malaysia-UTM
sulaiman@fke.utm.my

Izzeldin Ibrahim Mohamed Abdelaziz
Faculty of Electrical Engineering
Universiti Teknologi Malaysia-UTM
izzeldin@fke.utm.my

***Abstract*— Internet traffic classification is an area of current research interest. Identification of real time applications such as Skype has gained more attention in the last few years. Skype traffic classification is challenging because Skype uses encrypted traffic and uses no well-known port number. Several methods which used both signature-based and statistical approaches were proposed. However, the training and testing datasets validation have not been formally addressed. This work highlights the problem of machine learning (ML) datasets validation and proposes a mechanism based on ML statistical approach to identify Skype traffic. Two different networks environment are considered for Skype traffic to gain insight into the statistical features of Skype traffic. Ten algorithms within Weka are used to examine the best algorithm for the given datasets. Random Forest was found to be the best resulting in more than 99.8% accuracy.**

Keywords: Traffic classification, machine Learning, Skype, ML algorithms

I. INTRODUCTION

Over the last few years Skype has gained significant attention and has become one of the most popular forms of VoIP software. According to the Skype website [1], Skype users in the last year spent 1.8 billion hours making video calls. Also, at certain times, more than 22 million users were logged onto Skype at the same time. Skype is easy to use and provides a wide range of services such as voice and video calls, data transfer, video conference, instant message, online number, sharing screen etc.

Skype consists of several elements which are responsible for providing the connection between the two communication parts. Skype Client (SC) is a term for the machine and software which runs the Skype application. This includes computers and smart phones. The second element is Super Node (SN) which is a node with public address and adequate specification (CPU, RAM, .etc.). SNs establish networks among themselves, while SC tries to select an SN. Another two elements are Skype Login Server (LS) and Skype Update Server. The first is responsible for authentication checking, the second make checks to update users' versions with each login. More explanations and details of how Skype elements communicate can be found in [2], [3], [4], and [5]

Internet Server Provider (ISP) and network operators are usually interested to know the traffic carried in their networks for the purposes of optimising network performance and security issues. Therefore, Internet traffic classification is something important, particularly interactive traffic such as Skype. Based on a review of the literature, we can divide Skype classification methods into three groups namely, (i) ML methods [6] [7] [8] [9] [10] [11] [12] [13]which calculate flow or packets statistical features by using ML algorithms, (ii) algorithms methods [2] [5] [14] [15] [16] which develop or update an algorithm or model depending on features collected from Skype login or connection analysis and (iii) mixed methods [17] which combine ML statistical values, payload signature and header information (port number). The evaluation metrics depend on the different methods that were used, for example, ML works can be evaluated by True Positive (TP), True Negative (TN), False positive (FP), False Negative (FN), Precision or Recall. Another evaluation method is to consider standard Skype traces (ready data sets) [18] as testing data, and test the classifier as to whether the classifier can classify all the Skype packets in the test data as Skype traffic. An additional mechanism to validate Skype classification methods is to compare the results with commercial classifier results (such as Packet Shaper/Packeteer).

In this paper we aim to achieve two goals, firstly; to discuss the validity of using training and testing datasets for ML Skype classification. This we do by answering the question, are the statistical features of Skype traffic the same or different in different network environment. The second objective is to compare of several ML algorithms to reach optimum classification accuracy. In this paper, we consider full Skype session (calls) datasets to comparing statistical features of different network environments. Then four algorithms within Weka [19] are used to identify Skype traffic agonist all other non-Skype traffic.

Several mechanisms and methods were proposed to detect Skype traffic, but none of these succeeds correctly in

classifying all the Skype traffic cases[2]. Identifying Skype traffic is not always easy for the following reasons:

- Skype is a P2P network, so each user can act as a client or server for other users
- There is no well-known port number for Skype
- Skype has a non-fixed protocol
- Skype uses encrypted payload
- Skype continuously releases new software versions
- The values of the statistical traffic are different depending on Skype services (voice/video/data) [20] and version[6].
- The communication between two end SCs includes other channels in between (SNs), which pose some difficulty on Skype classification.

Section 2 describes and analyses ML datasets validation. Related works were viewed in section 3. Some experiments to answer the question "Are the statistical features of Skype the same in different network scenarios?" are discussed in section 4. Overviews of classification mechanisms, system stages, the proposed six Skype Snort rules, experiments work are discussed in Section 5. Finally, Section 6 provides the conclusion and limitations of the mechanism.

II. ML DATASETS VALIDATION

The problem encountered in the machine learning Internet traffic classification is the validation of training and testing datasets. Normally, the datasets criterion's is assumed to be similar to the real network environment. The challenge in Skype classification is the difficulty to ensure the similarity of training traffic characteristics (packets/flows features) and the traffic to be tested which can only be guaranteed if both are taken from a single machine with the same network environment. Thus there is likelihood that the traffic features values will differ depending on the network factor. This may imply that ML classification for Skype will be only be accurate when all the training and testing datasets are collected in real time from the same network environment. This is further explained from the figure 1 below.

According to [21] and [22], many Internet applications change their statistical properties over time. [23] did a good comparison between classification accuracies when Skype datasets were collected from different networks as well as over different years. The training dataset (Univ07) used was collected in 2007 from a university in Canada. Three testing datasets were considered. The first is where the training and testing datasets were from the same network. The second testing dataset is from the same network as training datasets but for different year (Univ10). The last group is from different country (Italy). The results show that the Detection Rate (DR) is high and False Positive (FP) is low when the training and testing datasets comes from the same network and at the same time.

The requirement of using real data (traffic packets) is essential in ML classification. We look to use valid Skype ML datasets and then use a statistical approach to distinguish between Skype and non-Skype traffic (Skype classification). Another motivation is to identify features that are able to classify Skype traffic in specific networks scenarios. In general, Internet traffic classification can help to increase security issues and decrease malicious users [7] [24]. In particular, when Skype traffic is identified, real characteristics of Skype can be defined. This identification also helps campuses and organizations to manage their internet traffic, and also reveal applications which use non-well port number to hide themselves. Operators are usually interested to know the traffic carried by their networks for the purposes of optimising network performance [25].

III. RELATED WORK

This section discusses some research work which uses the different ML datasets for the training and testing stage. The shortcomings of this approach are reviewed and studied further in our work.

[13] aims to classify encrypted traffic and take SSH and Skype as the case study. The authors developed classifier trained data from one network to test on data from an entirely different network. The testing data is collected from three different places (Dalhousie traces, public traces and DARPA99 traces). Each of these traces is trained from the Dalhousie network. However, the question here is how to ensure the validity of output of the testing stage when the training and testing data for both sets of data is totally different.

In [9] the authors mentioned that Skype traffic can be identified by observing five seconds of a Skype traffic flow. The classifier achieved more than 98% accuracy and succeeded in identifying suitable traffic features to classify Skype. However, the method and datasets are used only for offline classification. The offline detection has several shortcomings such as the different environment from the online classification.

The authors in [11] used AdaBoost and C4.5 to classifying the traffics into Skype and non-Skype. The Skype traces were collected as labeled data and taken from the campus network. The data were separated into UDP and TCP and classified independently. The classification results are 98% and 94% for UDP and TCP respectively. However, the labeled datasets were collected at different classification times. This causes a difference between classification environment and datasets collection. Moreover the classifier did not identify all Skype traffics.

The researchers in [6] uses ML in their work. They focused on Skype classifications for versions 2, 3 and 4. The work used ten folds cross-validation with 100 packets sub-flow. The results showed about 98% precision and 86% recall. As has been the case with the previous works, the problem is the use of training and testing datasets which were collected from two different environments. The first group is Skype traces collected in real-time using Tcpdump (of unknown origin). The second group comprises the offline pcap files which were obtained from University of Twente (saved files). Again the question of how to train the classifier by datasets

collected from some network and to examine this classifier by datasets

Skype calls	TCP rate	UDP rate	Avg.pckt/S	Avg.pckt/size
Between two NAT Networks- call 1.1	99.89%	0.11%	138.244	128.170
Between two NAT Networks- call 1.2	98.73%	1.24%	112.738	130.039
Between two NAT Networks- call 1.3	83.16%	16.84%	195.569	70.633
Between two NAT Networks- call 1.4	99.43%	0.56%	162.724	75.128
Between two Real IPs - call 2.1	5.29%	90.66%	20.782	147.465
Between two Real IPs - call 2.2	1.08%	98.92%	84.411	120.204
Between two Real IPs - call 2.3	0.30%	99.70%	87.290	121.991
Between NAT and Real IP – call 2.4	1.83	98.12	61.613	120.158

Table 1: Skype calls some statistical values

from another network, where the characteristics of the two networks may be different.

[12] is a flattering work which proposes an online method based on SVM-ML to classify Skype traffics. This work has an advantage over others in data collection. All datasets (covering both training and testing) were collected from the campus network. However one question remains as to how to evaluate the classifier due to the lack of comparison to ensure classification results. We did not find a paper in the literature review that combined Snort rules and ML to identify Skype traffics.

IV. ARE THE STATISTICAL FEATURES OF SKYPE THE SAME IN DIFFERENT NETWORK SCENARIOS?

As discussed earlier, one important issue in ML is to use valid training and testing datasets. We consider full Skype session (call) datasets to answer the question, is Internet application (Skype in particular) traffic features the same when the traffics are collected from different network environments. All data was collected by Wireshark [26], as well, the statistical values are summarized from the same software.

Eight different Skype calls were considered and divided into two groups. The first group includes four different calls (calls 1.1-1.4). In this group, the Skype sessions are full Skype session (call) between two SCs located in two different countries. This means, both Skype clients are located behind firewall and thus using NATed IP. The second group includes other four calls (calls 2.1 – 2.4) of Skype session between two SCs located inside our campus area. This group calls configures with no firewall between both clients (the two clients used real IPs). We aim to generate two different datasets to study Skype traffic features in two different network scenarios.

Table 1 and figure 1 show statistical features results of the two groups. When we compare the features values of group one together, the TCP rate, UDP rate, average packets per second and average packets per size (bytes) were seen to be have near values. This means, the same network environment, generate same traffic features. However, when any call of group one compared with other call from group two, clear differences appear in the TCP rate, UDP rate, and average packets per second. This means the statistical features of

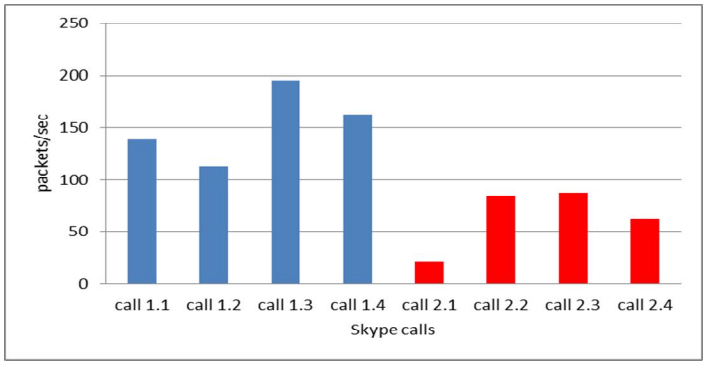

Figure 1 Average packet per second

Skype traffic are not the same when the network environments are different.

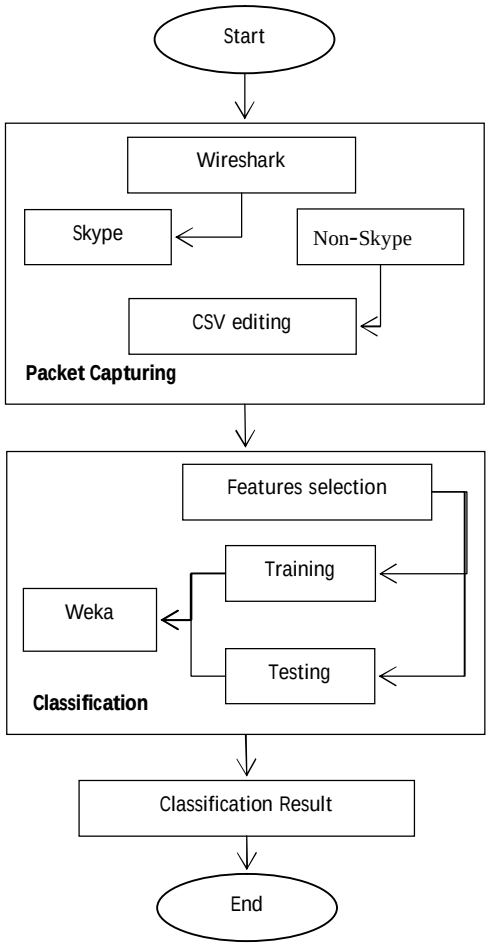

figure2. Classifier process steps

V. ML SKYPE CLASSIFICATION

A. *Classifier description*

In order of finding interactive applications classification mechanism, ten of ML algorithms were used to classify the

traffic into Skype and non-Skype. Figure 2 shows experimental process steps, which start by using Wireshark to captures Skype and non-Skype. The Skype datasets are real sessions (calls), which collected from some monitored IPs in our campus. As well non-Skype traffic as collected by the same manner, which include almost all Internet applications such as e-mail, games, youtube, facebook, HTTP, etc. Then, the two pcap files was editing by Excel to remove unneeded parameters and prparign the data in Weka strcure. Two capturing devices was provided, the first device equipped with Intel core (TM) i3-2330M CPU 2.2 GHz 2 core and memory

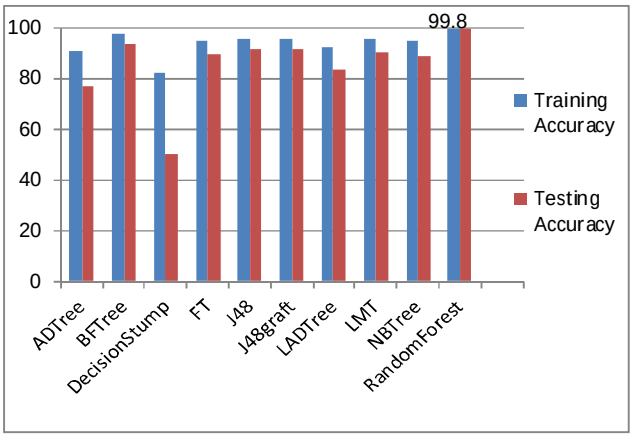

figure3. ML algorithms accuracy

6.00 GB, and the second device is Server of 4 CPUs, Intel(R) XEON(TM) 2.00 GHz . Our aim is to study and analysis interactive applications (Skype). To do so, real Skype sessions (calls) of some monitored IPs are managed and captured.

B. Features and datasets

In the first step of classification, Packet Interarrival time and packet length were selected as features for machine learning classifier. The advantage of selecting of only two features is reducing of classification complexity, particularly we deals with real time application. Each capturing pcap file was divided into two parts. The upper portions (T1 and S1) of the files were used as training dataset and the lower portions (T2 and S2) of the same files were used as testing dataset. By this we aims to take the training and testing datasets from the same place and at near time.

The total number of training instances are19,995 packets, which include the upper portion of Skype capturing files (S1) and the upper portion of non-Skype capturing file (T1). 2,002 packets (S2 + T2) were used as testing dataset. In the same manner, the total numbers of testing are 2002 packets, which include the lower portion of Skype file (S2) and the lower portion of non-Skype files. Three benefits were gained when collect data by this way, first, ensure of similarity between training and testing data, second; reducing of classification errors rate (this type of error discussed on [7]), three; easy in data processing.

C. Results and Analyzing

To classify one of most popular real time applications (Skype), ten algorithms within Weka Explorer were used. These algorithms were applied to both training and the testing, which are ZeroR, PART, DecisionStump, J48, J48graft, LADTree, NBTree, RandomForest, RandomTree, and REPTree. Figure 3 and 4 illustrate training and testing accuracy and time taken to build models. It's clear that Tree.RandomForest algorithm provided optimal results of 99.8% accuracy, compared with other algorithms. In addition to that, DecisionStump provides the shortest time (0.05 Seconds) when compared with other algorithms. Because we dealing with Skype traffic (interactive applications), time to

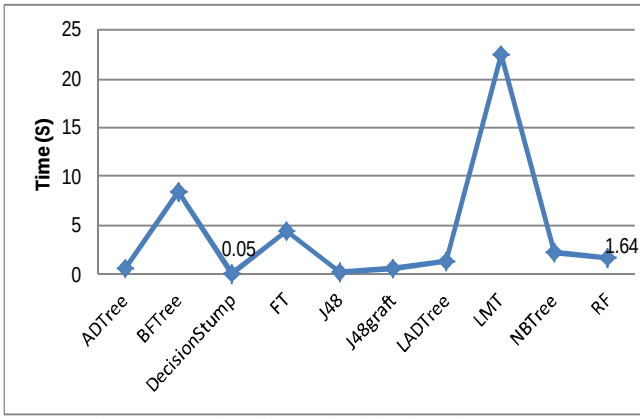

figure 4 ML Algorithms Classification time

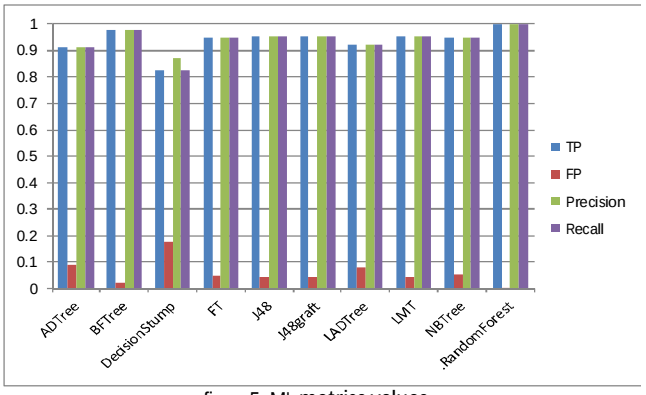

figure5. ML metrics values

build a model is very important factor. Figure 4 analysis the ML metrics True Positive (TP), False Positive (FP), Precision, and Recall, which are used to compare between algorithms.

VI. CONCLUSION AND FUTURE WORK

Interactive applications such as Skype and online games use several approaches to prevent from being detected. Supplying ML with real and valid training datasets is an important issue for IP traffic classification. In this paper, two network scenarios are considered to check Skype statistical features. We conclude that some of Skype statistical features such as packets per second are varying when network environment is different.

Because of the important of the relation between training and testing ML datasets, we collected both from the same

place and at near time. In this paper we comparing between a ten of ML algorithms to classify Skype traffic. Real dataset was collected from campus environment to give fact inputs to classifier. The comparison result shows, Tree.RandomForest algorithm was obtained high accuracy 99.8%. Moreover and from time point of view, the work has acquired that some of ML algorithms were suitable to classify interactive real time applications. The method has limited as offline classification, which is material for future work for online classification. To do so, how to collect known trained dataset from near real time at the same traffic (Switch/Router) without manual IPs monitoring?. Another question can we find some Skype traffic features, which are suitable to identify Skype traffic in all kinds of networks environments.

References

1. Skype.website. www.skype.com (2012). Accessed accessed at 6/5/2012

2. Adami, D., Callegari, C., Giordano, S., Pagano, M., Pepe, T.: Skype-Hunter: A real-time system for the detection and classification of Skype traffic. Int J Commun Syst **25**, 386–403 (2012).

3. EzineArticles.com: How Skype work. http://ezinearticles.com/?How-Skype-Works&id=496462 (2012). Accessed 8/5/2012 2012

4. Baset, S.A., Schulzrinne, H.G.: An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol. In: INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings, April 2006 2006, pp. 1-11

5. Dongyan, Z., Chao, Z., Hongli, Z., Hongliang, Y.: Identification and Analysis of Skype Peer-to-Peer Traffic. In: Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on, 9-15 May 2010 2010, pp. 200-206

6. Jesudasan, R.N., Branch, P., But, J.: Generic Attributes for Skype Identification Using Machine Learning. Technical Report **100820A** (2010).

7. Ibrahim, H.A.H., Nor, S.M., Mohammed, A., Mohammed, A.B.: Taxonomy of Machine Learning Algorithms to classify realtime Interactive applications. International Journal of Computer Networks and Wireless Communications **Vol. 2, No. 1, 2012** (2012).

8. Li, J., Zhang, S.Y., Xuan, Y., Sun, Y.F.: Identifying Skype Traffic by Random Forest. 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Vols 1-15, 2841-2844 (2007).

9. Branch, P.A., Heyde, A., Armitage, G.J., Acm: Rapid Identification of Skype Traffic Flows. Nossdav 09: 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video. Assoc Computing Machinery, New York (2009)

10. Alshammari, R., Zincir-Heywood, A.N.: An investigation on the identification of VoIP traffic: Case study on Gtalk and Skype. In: Network and Service Management (CNSM), 2010 International Conference on, 25-29 Oct. 2010 2010, pp. 310-313

11. Angevine, D., Zincir-Heywood, A.N.: A Preliminary Investigation of Skype Traffic Classification Using a Minimalist Feature Set. In: Availability, Reliability and Security, 2008. ARES 08. Third International Conference on, 4-7 March 2008 2008, pp. 1075-1079

12. Hongli, Z., Zhimin, G., Zhenqing, T.: Skype traffic identification based SVM using optimized feature set. In: Information Networking and Automation (ICINA), 2010 International Conference on, 18-19 Oct. 2010 2010, pp. V2-431-V432-435

13. Alshammari, R., Zincir-Heywood, A.N.: Machine learning based encrypted traffic classification: Identifying SSH and Skype. In: Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, 8-10 July 2009 2009, pp. 1-8

14. Weirong, J., Gokhale, M.: Real-Time Classification of Multimedia Traffic Using FPGA. In: Field Programmable Logic and Applications (FPL), 2010 International Conference on, Aug. 31 2010-Sept. 2 2010 2010, pp. 56-63

15. Chen, K.T., Huang, C.Y., Huang, P., Lei, C.L.: Quantifying skype user satisfaction. Comput Commun Rev **36**(4), 399-410 (2006).

16. Adami, D., Callegari, C., Giordano, S., Pagano, M., Pepe, T.: A Real-Time Algorithm for Skype Traffic Detection and Classification

Smart Spaces and Next Generation Wired/Wireless Networking. In: Balandin, S., Moltchanov, D., Koucheryavy, Y. (eds.), vol. 5764. Lecture Notes in Computer Science, pp. 168-179. Springer Berlin / Heidelberg, (2009)

17. Freire, E.P., Ziviani, A., Salles, R.M.: Detecting Skype flows in Web traffic. In: Network Operations and Management Symposium, 2008. NOMS 2008. IEEE, 7-11 April 2008 2008, pp. 89-96

18. Group, T.N.: Skype Traces http://tstat.tlc.polito.it/traces-skype.shtml (2008). Accessed 14/5 2012

19. Witten, I.H., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques. Diane Cerra, (2005)

20. Bonfiglio, D., Mellia, M., Meo, M., Rossi, D., Tofanelli, P.: Revealing Skype traffic: When randomness plays with you. Comput Commun Rev **37**(4), 37-48 (2007).

21. Nguyen, T.T.T., Armitage, G.: Clustering to assist supervised machine learning for real-time IP traffic classification. Ieee Icc, 5857-5862 (2008).

22. Nguyen, T.T.T., Armitage, G.: Training on multiple sub-flows to optimise the use of Machine Learning classifiers in real-world IP networks. Conf Local Comput Ne, 369-376 (2006).

23. Alshammari, R., Zincir-Heywood, A.N.: Is machine learning losing the battle to produce transportable signatures against VoIP traffic? In: Evolutionary Computation (CEC), 2011 IEEE Congress on, 5-8 June 2011 2011, pp. 1543-1550

24. Alshammari, R., Zincir-Heywood, A.N.: Can encrypted traffic be identified without port numbers, IP addresses and payload inspection? Comput Netw **55**(6), 1326-1350 (2011). doi:DOI 10.1016/j.comnet.2010.12.002

25. Molnar, S., Perenyi, M.: On the identification and analysis of Skype traffic. Int J Commun Syst **24**(1), 94-117 (2011). doi:Doi 10.1002/Dac.1142

26. Orebaugh, A., Ramirez, G., Burke, J., Pesce, L., Wright, J., Morris, G.: Wireshark & Ethereal Network Protocol Analyzer Toolkit. Syngress, (2007)