ENHANCING ACCURACY OF DPI TRAFFIC CLASSIFIER

Pooja Mehta
ME Student
GTU PG School
Gandhinagar

Shivi Shukla
Project Engineer
C-DAC
Pune

Abstract

Deep Packet Inspection (DPI) techniques are considered extremely expensive in terms of accuracy and processing costs and therefore are usually deployed in edge networks, where the amount of data to be processed is limited. This paper demonstrates that, in case the application can't tolerate any compromises in terms of accuracy, the processing cost can be reduced while even improving the classification precision, making DPI suitable with SSLi for high-speed networks.

Keywords: *Traffic analysis, deep packet inspection, network monitoring, SSLi, traffic classification*

---

1 Introduction

The usage of the internet changed dramatically in the previous couple of years. The internet transports traffic generated by using many exclusive users and programs which include financial transactions, e-enterprise, leisure and more, which is truly one of a kind from the site visitors. We had two decades ago whilst the network changed into engineered for e- mail, telnet and ftp. Frequently, the best way to keep the pace of the new visitors trends is to have an effective infrastructure for actual-time measurements within the community, which permit to find out adjustments inside the site visitors pattern as soon as they seem and adapt to them quick.

The capability to recognize which software generated the visitors is perhaps one of the most essential challenges in network measurements. Several technology was proposed to this point. The most important criticism to DPI isn't related to its difficulties in classifying encrypted or tunneled traffic, but to its meant excessive processing value. In truth, DPI is appreciably utilized in security applications along with ids and firewalls that have strict necessities in terms of precision. In different phrases, a single misclassification in such those programs should permit an attacker to compromise even an entire network, and is consequently a threat that humans do now not want to run. So one can decrease the chance of misclassifications, a majority of these DPI implementations generally tend to privilege the accuracy, without taking the processing cost into plenty consideration. The basic intention is to boost up the content material inspection and decrease the overall required processing of the detection engine. The proposed answer similarly gives flexibility and scalability with a purpose to satisfy the growing desires of community protection. On this paper, we recognition on deep packet inspection with emphasis on enhancing the performance of the required content inspection and minimizing the packet processing load.

2 Related Work

As new net applications began to apply obfuscation methods (port masquerading, tunneling, and encryption) to stay away from traffic manipulate and restrictions, easy inspection of port numbers is not a dependable classification mechanism. Furthermore, payload encryption effortlessly thwarts traditional payload based classification based on pattern matching.
The second fundamentally different institution of payload-impartial tactics use flow-based features including average packet sizes, packets inter-arrival times, or flow durations. A current hybrid method tries to perceive TLS/SSL encrypted application layer protocols with a combination of a signature-based totally and a flow-based statistical evaluation scheme. The method is intently related to our concept; but its goal is confined to the classification of encrypted application layer protocols, whilst we concentrate extra on an in-intensity analysis of the classifier and revealing application flows.
Payload based methods are normally characterized by their excessive accuracy protecting the largest scope of detected packages. But, non-payload based totally tactics are nevertheless preferred due to privacy protection and to their capacity to come across encrypted programs.
Ml based tactics are the most reputed in traffic classification. Unsupervised ml strategies offer rapid classification that does not rely on schooling sets and has the potential to categories unknown packages. These techniques are typically as compared in keeping with their capacity for producing a minimum quantity of clusters that comprise the general public of the classification gadgets, with the highest predictive strength of a single traffic magnificence. Supervised studying algorithms are commonly compared according to the rate of getting to know and classification, tolerance to mistakes in attributes, etc. Each set of rules has specific strengths and weaknesses. For example, Bayesian classifier may be preferred for simplicity and memory saving. Context-based classifier are premiere in describing inter-elegance dependencies.
SVM offer complex classification fashions which are suitable for big units of traffic attributes. Latest assessment works emphasize at the preference for supervised decision trees for actual-time and encrypted traffic classification, specially, the C4.5 choice tree, which is capable of outperform both Bayesian and SVM classifier.

2.1 Limitations with vendor classification engines

Right here is the assessment of classification techniques utilized in industrial merchandise (e.g. Juniper, supplyfire) used for traffic-management and community-protection purposes. Those consist of routers, firewall, intrusion prevention system (IPS), secure net gateways and traffic shapers. Regrettably, there's little or no information available approximately the protocol classification carried out in most of those structures.

Challenge: Even though counting on not unusual techniques, business merchandise often relies on proprietary methods.

An instance of a proprietary algorithm used in tipping factor systems is protocol identification through statistical evaluation (PISA). Pisa creates a ten-dimensional representation of each fingerprint for each protocol, based on a schooling set of captured traffic. Pisa uses simple average and preferred-deviation values of fashionable flow attributes (packet size and inter arrivals) in both instructions, further to the Shannon entropy of the information at the software layer. It uses ok-method to cluster flows for preferred and P2P packages along with Skype. But, one of Pisa's main barriers is the required wide variety of packets to be analyzed before a flow is identified. For instance, Skype results stabilize after the six hundred packets.

Some other instance is juniper's DPI mechanism that suits the patterns inside the first packet of a consultation using deterministic finite states automata. It has the capacity to chain signatures and to specify a maximum range of transactions in which the signature must occur to be a healthy.
Network-based totally application reputation (nBAR), used by Cisco routers, relies on DPI and lots of utility-specific attributes. Its miles a state-oriented classification mechanism that helps programs with dynamically negotiated port numbers, along with rtp. It is able to help sub classification, including http consumer agent, content material-kind and uniform aid locator (URL). Nbar2 is a prolonged version of nBAR that helps evasive packages which includes Skype and tor, cloud-based totally applications along with office365, or even mobile programs which includes FaceTime. Cisco service manipulate engine (sce) is a devoted hardware DPI appliance that consists of protocol country evaluation collectively with behavioral and heuristic analysis.
In most of these business merchandise, the increasing requirement for content material awareness and alertness visibility explains the DPI integration with statistical and ml techniques collectively with SSL decryption. Particularly, DPI helper techniques protected port heuristic in juniper behavioral in IPoque and SVM in Websense. The future requirements which are riding the market for subsequent-era merchandise and the important thing capabilities of future classifier are taken into consideration next.

3 Methodology

After the literature survey, a new improved encrypted traffic detection method is proposed. The proposed method can be roughly divided into three parts, namely decryption module, analysis module and re-encryption module. Packet flow can be either classified as P2P class or non-P2P class. However, the classification accuracy is improved by using both DPI method and SSLi tool. In the proposed system, first thunder SSLi will decrypt the packet, then DPI will be performed, and re-encrypted packet will be sent out of the network. By this, we have tried to improve the accuracy to some extent. Figure shows our proposed system to classify P2P packet flows.

After the literature survey, a new advanced encrypted visitor's detection technique is proposed. The proposed method may be more or less divided into three components, particularly decryption module, analysis module and re-encryption module. Packet go with the flow can be both classified as P2P class or non-P2P elegance. But, the class accuracy is advanced by using both DPI technique and SSLi tool. Inside the proposed system, first thunder SSLi will decrypt the packet, then DPI can be executed, and re-encrypted packet might be dispatched out of the community. By way of this, we've got tried to enhance the accuracy to a point. Figure suggests our proposed device to categories P2P packet flows.

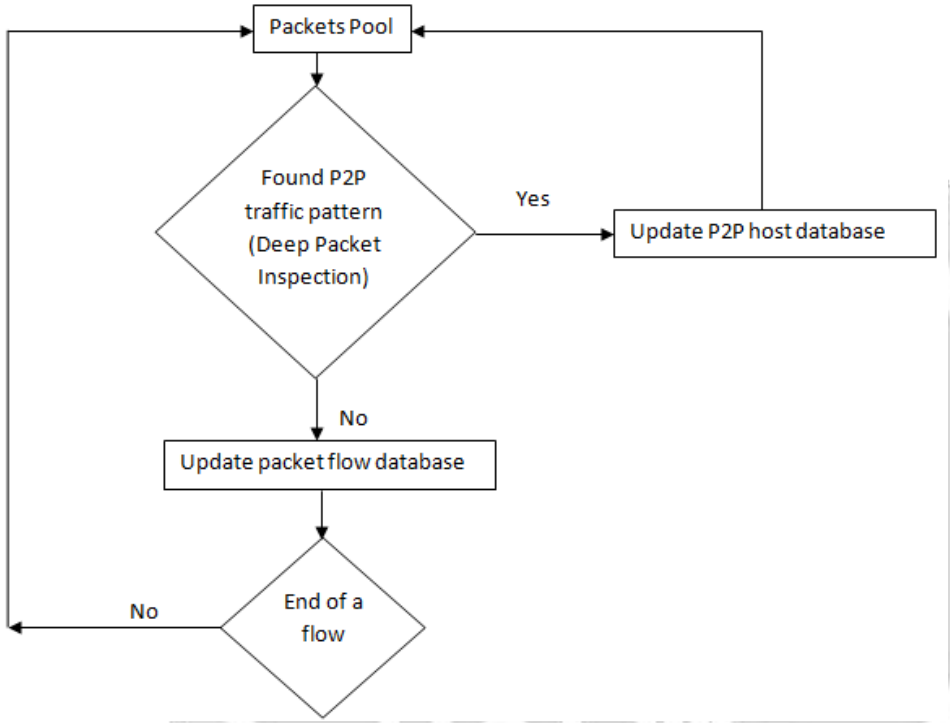


Figure 1 proposed system to classify P2P packet flows

3.1 Architecture of Thunder SSLi

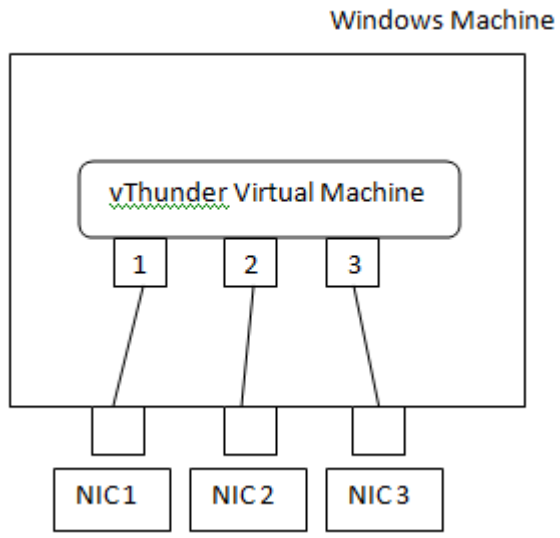


Figure 2 architecture of vThunder

3.2 Network Topology

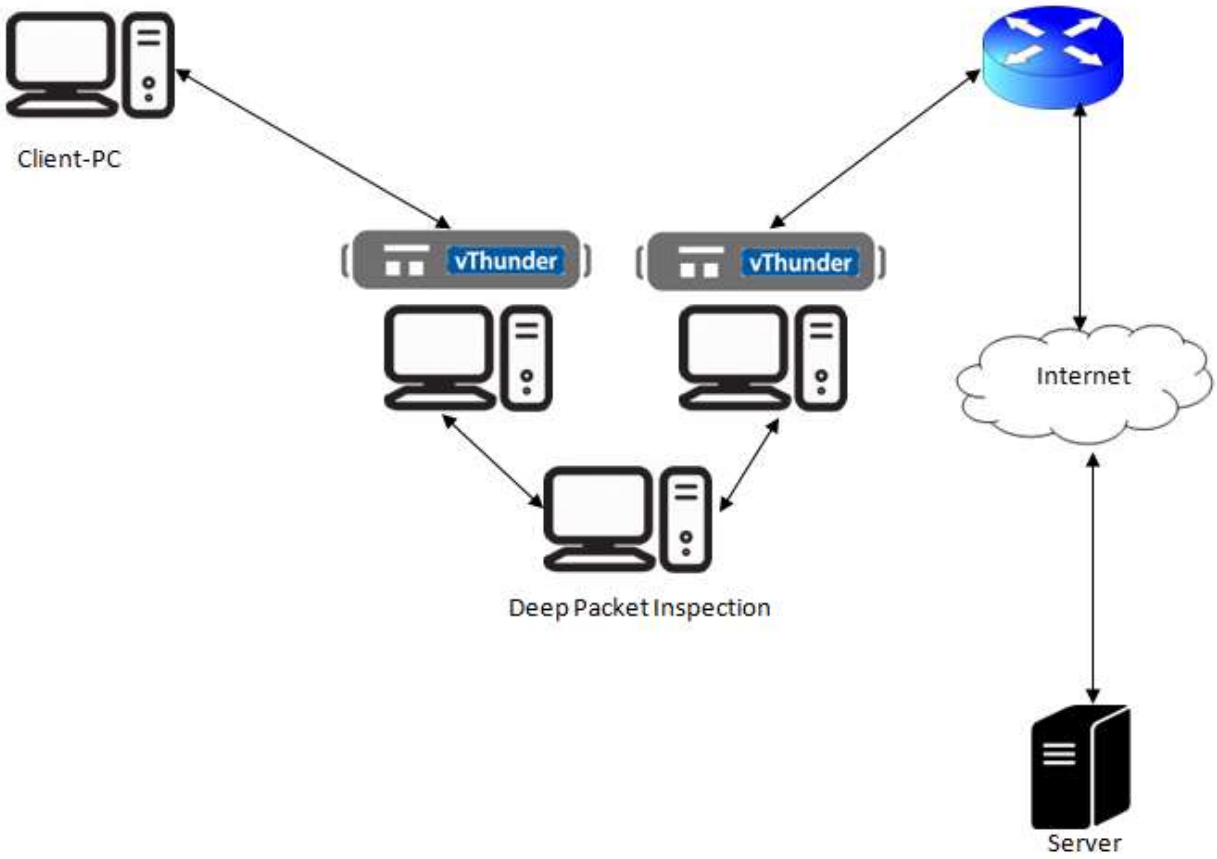


Figure 3 Network topology for experiment

An encrypted packet will first be decrypted with the aid of vThunder SSLi set up in VM of the windows device. After that, to seize P2P packets, here in particular the BT packets, we manually force the client to use a

unmarried TCP port (i.e. 1200) for statistics transfer. For this reason, all the BT site visitors should go through this TCP port. Then, we start a pattern torrent document and the customer will routinely start downloading/importing the contents. On the identical time, we start our packet shooting software to attain the packets. In addition, to capture non-P2P packets, we begin our packet shooting software whilst we had been creating non-P2P network activities along with http, ftp and SSH. But, as the wide variety of BT packets growth, the classifier can be saturated sooner or later. After that, even extra packets is provided, the accuracy will increase notably.

3.3 Performance Evaluation Matrix

Based totally at the literature survey that we have accomplished, normally talking there are two types of overall performance evaluation matrix utilized by the authors of the papers. The first kind really makes use of the time period "accuracy". Its miles defined because the variety of successfully classified objects divided via the whole number of items. As the name implied, higher the accuracy represents the better the proposed algorithm. A few other papers used a alternatively formal definition of the statistical equations (equation 1 to 4) to evaluate the overall performance. The goal becomes to maximize (i.e. 1) the true positive rate (TPR) and true negative rate (TNR).

$$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \qquad \text{EQUATION 1}$$

$$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{TN+FP}} \qquad \text{EQUATION 2}$$

Where,

TP = number of correctly identified objects for a given P2P class

TN = number of correctly rejected objects for a given non-P2P class

FP = number of objects falsely identified as P2P class

Fn = number of objects from P2P class that are falsely rejected

4 Experiment

Installing vThunder advanced traffic manager on VMware ESXi

```
Starting up ...

Decompressing Linux... Parsing ELF... done.
Booting the kernel.

SoftAX login: admin
Password:

[type ? for help]

SoftAX>enable
Password:
SoftAX#config
SoftAX(config)#enable-password enpwd1
SoftAX(config)#interface management
SoftAX(config-if:management)#ip address 192.168.2.228 /24
SoftAX(config-if:management)#ip default-gateway 192.168.2.1
SoftAX(config-if:management)#show interface management
GigabitEthernet 0 is up, line protocol is up.
  Hardware is GigabitEthernet, Address is 000c.293f.436a
  Internet address is 192.168.2.228, Subnet mask is 255.255.255.0
  Internet V6 address is ::/0
. . .
```

Figure 4 Installation of vThunder in VM

```
SoftAX(config-if:management)#ip default-gateway 192.168.2.1
SoftAX(config-if:management)#show interface management
GigabitEthernet 0 is up, line protocol is up.
  Hardware is GigabitEthernet, Address is 000c.293f.436a
  Internet address is 192.168.2.228, Subnet mask is 255.255.255.0
  Internet V6 address is ::/0
  Configured Speed auto, Actual 1000, Configured Duplex auto, Actual fdx
  Flow Control is disabled, IP MTU is 1500 bytes
  166 packets input,  11349 bytes
  Received 0 broadcasts,  Received 0 multicasts,  Received 166 unicasts
  0 input errors,  0 CRC  0 frame
  0 runts  0 giants
  6 packets output  468 bytes
  Transmitted 0 broadcasts  0 multicasts  6 unicasts
  0 output errors  0 collisions
SoftAX(config-if:management)#ip control-apps-use-mgmt-port
SoftAX(config-if:management)#exit
SoftAX(config)#admin admin password adminewpwd
SoftAX(config-admin:admin)#write memory
Building configuration...
Write configuration to default startup-config
[OK]
SoftAX(config-admin:admin)#
```

Figure 5 vThuner CLI

4.1 P2P Packets Classification Software

Development environment

The P2P packet class software is built with Cygwin below home windows surroundings. Cygwin affords a Unix like surroundings with the overall GCC supported for software improvement and home windows provides user friendly environment without demanding about the community card drivers. This combination supplied a person

pleasant improvement environment and additionally become able to utilize open supply libraries. Packets seize is made feasible through Winpcap library. Win cap is an open source library that allow user to set his/her network interface card (NIC) to operate in "promiscuous" mode. Consequently, all of the packets going through the network will be captured. Parent shows the gadget level of structure of the software written in this venture.

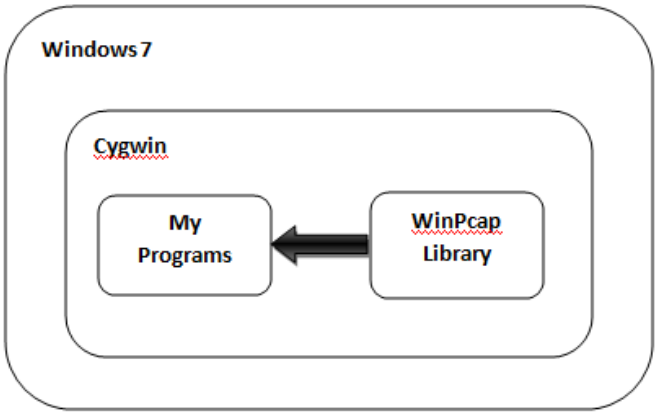


Figure 6 Classification software

```
68.215663 72868.216128 68.850134 53.861556 43.850662 44.420519 35.138308 28.430733 28.467183 19.314173 1
532.470130 69664.343277 544.308862 383.708930 341.550180 351.261990 260.596566 195.370164 197.611818 126.780714 2
516.103221 62886.006197 531.656331 337.220048 381.773563 392.726630 263.260030 138.660572 143.421809 77.658853 2
1948.389678 33556.950720 2532.906581 700.766049 1856.519826 2413.475773 704.938492 96.227096 125.095225 5.522217 2
24.986129 35706.988044 25.469245 10.913349 18.420252 18.782618 6.643760 10.591033 10.659118 4.471425 1
132.374301 2098.472340 135.348809 6.512157 122.121150 124.991519 5.100712 14.294605 14.451497 1.591970 2
45.718235 90122.970090 46.201968 45.472543 10.766651 10.973982 9.181654 39.037583 39.070388 36.530961 1
15.481357 67209.900697 15.655305 13.358433 14.525806 14.525806 10.825768 4.968247 4.968247 2.791157 1
8.057950 8547.906709 8.388532 1.241782 7.250857 7.357929 0.659162 4.818884 4.914886 0.702662 2
7.967405 7356.155120 7.989822 1.090285 7.990841 7.995401 0.666593 3.986616 3.987759 0.543818 2
33.473361 7276.841199 34.913649 3.618781 23.473239 24.513530 1.685568 14.033118 14.634538 2.067368 2
346.319264 63440.468426 349.719944 229.084440 192.220488 195.021273 130.231227 158.457293 159.108830 100.746329 2
6.942565 31422.766958 7.189673 3.588390 9.179041 9.355561 3.285239 1.768880 1.910391 0.458864 1
19.724875 16293.448939 19.836525 4.364258 15.632284 15.743943 2.003068 8.112168 8.140141 2.499712 2
0.000000 4592.186429 0.000000 0.181390 4.001120 4.001120 0.296420 0.000000 0.000000 0.000000 2
33.921678 28017.220378 35.113406 11.430812 21.878466 22.606330 4.565280 16.080503 16.579553 7.039413 1
20.804161 10167.051481 22.118709 3.066230 13.946762 14.358306 1.427050 10.882826 11.432695 1.769786 2
812.562513 50081.410548 812.562513 425.806310 583.363342 583.363342 336.765033 233.751654 233.751654 93.741221 2
92.212617 73484.036302 94.869459 72.378286 62.464092 64.608382 50.757285 33.827405 34.399630 22.424614 1
17.196422 18275.029185 18.937121 4.301120 13.056287 14.088155 2.145721 8.159608 8.750476 2.295835 2
485.915365 44261.739579 485.915365 227.553704 424.118753 424.118753 221.704314 65.973880 65.973880 8.975310 2
```

Figure 7 Content inside the trained database

4.4 Online Classification Module

Figure shows the online classification module top level block diagram.

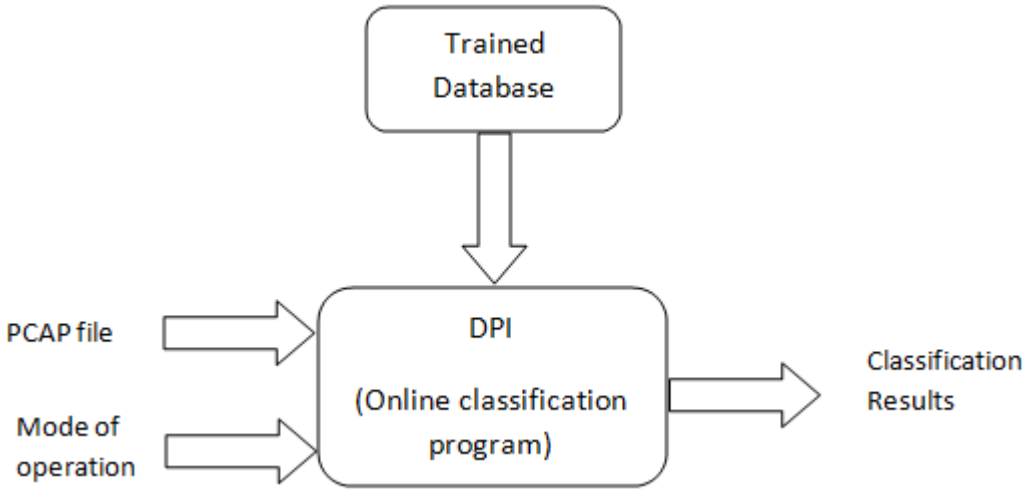


Figure 8 Online classification program block diagram

There are three enter parameters required by this module, specifically packet capture document, skilled database and the modes of operation. The online type software supports deep packet inspection mode of operation. This mode requires customers to provide a Pcap report as source of the packet flows. The DPI mode is based totally at the string (i.e. " BitTorrent protocol") assessment to determine if the encounter packet is BT type.

```
C:\Users\PM>Pooja
$ ./dpi.exe ../dat/center.dat ../packets/P2P.pcap dpi
DPI mode enabled
Report for ../pcap/P2P.pcap
Total numbr of IP:408
Total number of packets:31000
Total number of packet flows processed:0
(Results with DPI method) BT IP:391, NONBT IP:17, BT percentage:95.83%
Total time spent:0.013400
C:\Users\PM>
```

Before everything, an unknown elegance of packet flows can be inputted to the DPI module in which it is able to be determined if the packet is a BT packet. If it's far a BT packet, database of BT hosts can be updated at once. Otherwise, primarily based on the packet statistics, the corresponding packet go with the flow information may be up to date (i.e. Range of packet in drift, common packet length, etc) inside the packet drift database. If that packet is at the stop of a waft, identical technique could be repeated. Parent shows an example run of the net type software for deep packet inspection mode of operation with a pattern BT Pcap report. As we can see from the sample run, there are 408 BT IP addresses in the record. The DPI method is able to discover ninety six% of BT IP.

```
C:\Users\PM>Pooja
$ ./dpi.exe ../dat/center.dat ../packets/NONP2P.pcap dpi
DPI mode enabled
Report for ../pcap/NONP2P.pcap
Total numbr of IP:459
Total number of packets:9000
Total number of packet flows processed:0
(Results with DPI method) BT IP:441, NONBT IP:18, BT percentage:96.07%
Total time spent:0.013400
C:\Users\PM>\
```

Figure 10 Captured non-BT packets

As we can see from figure, there are 459 BT IP addresses in the file. The DPI method is able to detect 96% of BT IP.

5 Result And Analysis

The table gives the client and server IP address, port to be targeted and time taken by SSLi to decrypt and re-encrypt the packet.

Table 1 Details of the port targeted

Client address	Method used	Server address	Server port	Time taken for decryption	Time taken for re-encryption
200.198.136.76	Post	10.41.30.98	6436	0.123	0.125
192.168.2.56	Get	10.41.45.67	5434	0.064	0.067
192.168.2.140	Post	10.41.30.34	146	0.004	0.004
200.198.136.76	Get	10.41.45.56	5093	0.097	0.096
1.127.39. 34	Options	10.41.30.12	4678	0.34	0.29
192.168.2.36	Post	10.41.45.92	351	0.023	0.01
200.198.136.76	Options	10.41.30.27	3454	0.67	0.70
1.127.39.67	Get	10.41.45.48	2363	0.49	0.38
192.168.2.152	Options	10.41.30.50	1272	0.035	0.30
1.127.39.89	Post	10.41.45.73	281	0.29	0.30
200.198.136.76	Get	10.41.30.44	80	0.009	0.04

The figure shows the graphical representation of packet and bytes captured.

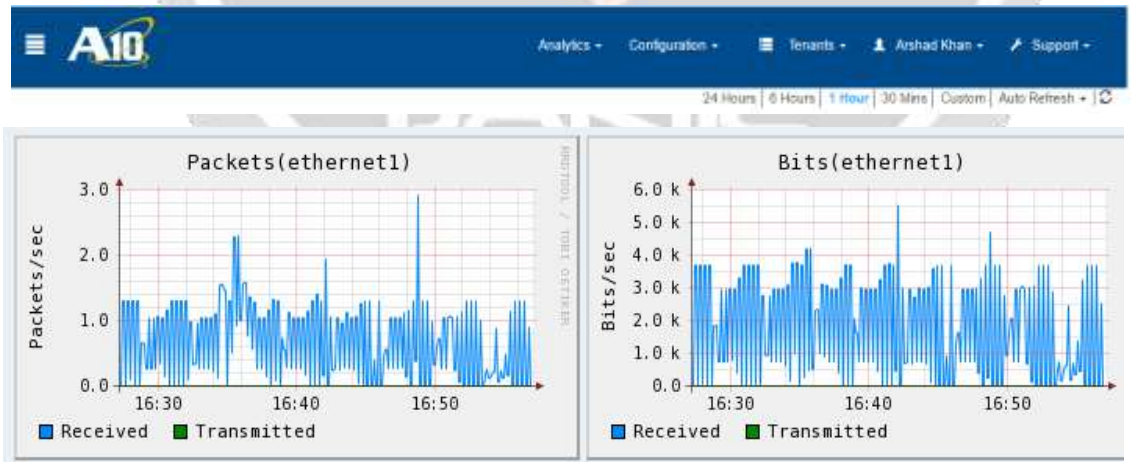


Figure 11 Packets and bytes captured

5.1 Classification Accuracy

Four statistical tests had been used to assess the classifiers in special modes of operation. They are the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rate (FNR). As we discussed in previous section, this performance evaluation metric could be very famous as it turned into utilized by many authors in their papers. So as to check the classifier, the classifier became first trained with 8000 TCP packet flows wherein greater than 3500 of them are BT TCP packet flows. Table show the class consequences for DPI (proposed algorithm) strategies for 2 simulation assessments. These experiments had been executed with Pcap documents with packets type recognized. The first test Pcap report contained packets with 408 P2P IP

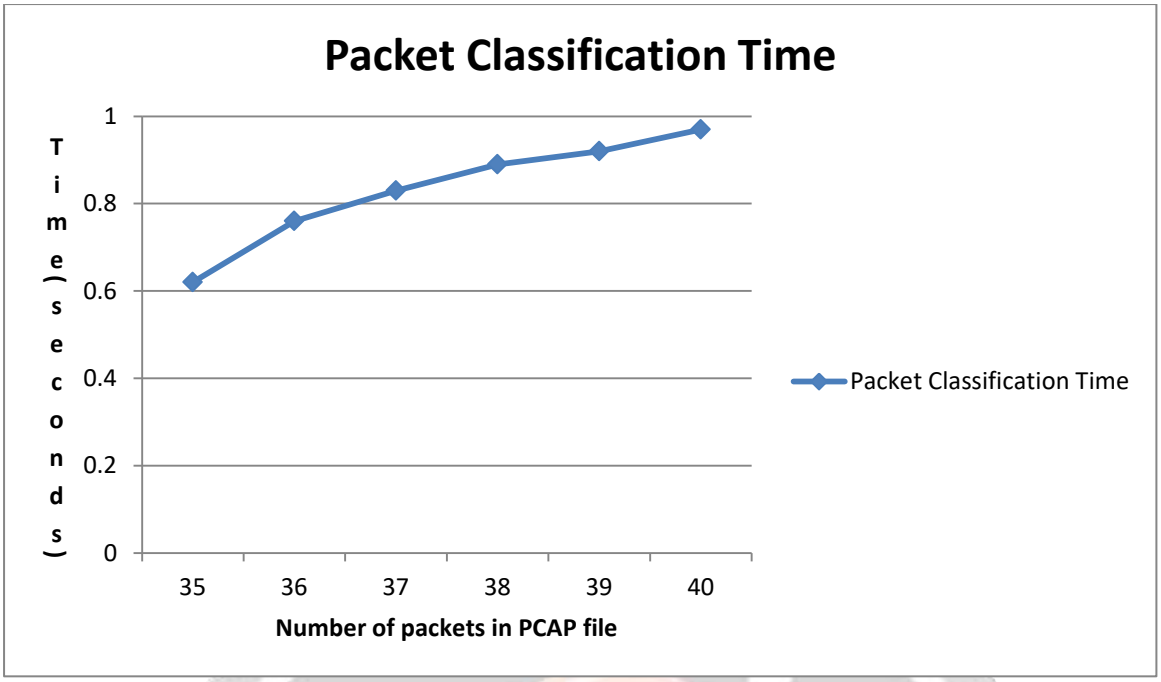
addresses; amongst them 391 had been BT IP copes with and 17 have been NON-BT IP addresses. The second one test Pcap record contained packets with 459 P2P IP addresses, 441 BT IP addresses and 18 NON-BT IP addresses. Based at the statistical consequences, the higher the TPR and the TNR, the better the classifier may be. There are multiple observations from the experiments. The first remark is that the DPI technique has 100% accuracy to stumble on http, ftp and SSH as NON-BT protocol. It's far due to the fact the DPI approach searches for the BT sample string ("Bittorrent protocol") explicitly inside the packets. Because the NON-BT packets rarely have specific BT pattern string inside the packets, 100% accuracy changed into anticipated. Then again, the DPI approach isn't always continually capable of discover BT packets. It is because the BT pattern string best occurs within the handshaking messages and it cannot seem throughout the BT facts transfer.

Table 2 Analysis summary

Method	Strength	Limitation
Port based approach	Easy to implement, no computation power required	Eliminate well-known non-P2P application (ftp, e-mail)
Payload based approach	High accuracy rate, robustness	No encryption supported, privacy issues
Flow based approach	Work with encrypted data packets	Difficult to determine threshold
Behavioral approach	Characterized by behavior e.g. Packet inter-arrival time, jitter, packet size	Required pre classified traffic trace, limited temporal validity of the training set due to network reconfiguration
Our approach	Higher accuracy compared to simple DPI methods, easier implementation	

5.3 Speed Performance Comparison

Discern shows the packets category time for numerous type methods. The DPI has the fastest type time because of the classification is only based totally at the string evaluation. This method is rapid as it handiest takes one packet so that you can determine the packet class (i.e. BT or NON-BT). The gain of this technique is that we are able to identify some of the packet flows elegance earlier. As soon as we decide the elegance of a packet drift, we do not need to hold track of the packet waft's records. Primarily based at the experiments performed, it appears that our approach is set 15%-20% quicker than conventional method.



6 Conclusion & Future Work

The promising simulation results show that by combining multiple techniques such as DPI, DFI and learning algorithms, the detection rate of the P2P packets, execution speed can increase significantly. In terms of the future work of this project, the ultimate goal would be applying the proposed algorithm into a live network situation.

References

[1] N. Basher, A. Mahanti, A. Mahanti, C. Williamson And M. Arlitt, "A Comparative Analysis Of Web And Peer-To-Peer Traffic," Proceeding Of The 17th International Conference On World Wide Web, April 21-25, 2008, Beijing, China

[2] H. Chen, Z. Hu, Z. Ye And W. Liu, "A New Model For P2P Traffic Identification Based On Dpi And DFI," Information Engineering And Computer Science, 2009. Iciecswe2009. International Conference On Digital Object Identifier:We10.1109/Iciecs.2009.5366295; Publication Year: 2009, Page(S): 1 – 3

[3] H. Chen, Z. Hu, Z. Ye And W. Liu, "Research Of P2P Traffic Identification Based On Neural Network," Computer Network And Multimedia Technology, 2009. Cnmt 2009.Weinternational Symposium On Digital Object Identifier: 10.1109/Cnmt.2009.5374510;Wepublication Year: 2009, Page(S): 1 – 4

[4] H. Chen, X. Zhou, F. You And C. Wang, "Study Of Double-Characteristics-Based SVM Method For P2P Traffic Identification," Networks Security Wireless Communications And Trusted Computing (Nswctc), 2010 Second International Conference On Volume: 1. Digital Object Identifier: 10.1109/Nswctc.2010.54wepublication Year: 2010, Page(S): 202 – 205

[5] F. Constantinou And P. Mavrommatis, "Identifying Known And Unknown Peer-To Peer Traffic," Proceedings Of The Fifth IEEE International Symposium On Network Computing And Applications, P.93-102, July 24-26, 2006

[6] Wej. Erman, A. Mahanti, M. Arlitt, I.Wecohen, And C. Williamson, "Offline/Real-Time Traffic Classification Using Semi-Supervised Learning," IfIP Performance, October2007

[7] R. Keralapura, A. Nucci And C. Chuah, "Self-Learning Peer-To-Peer Traffic Classifier," Computer Communications And Networks, 2009.

[8] A. Klemm, C. Lindemann, M. K. Vernon, And O. P. Waldhorst. Characterizing The Query Behavior In Peer-To-Peer File Sharing Systems. In IMC '04: Proceedings Of The 4th ACM Sigcomm Conference On Internet Measurement, Pages 55–67. ACM Press, 2004

[9] T. Le And J. But, "Bittorrent Traffic Classification," CAIA Technical Report 091022a, 22 October 2009

[10] B. Liu, Zhitand Li And Zhanchun Li, "Measurements Of Bittorrent System Based On Netfilter," Computational Intelligence And Security, 2006 International Conference On Volume: 2 Digital Object Identifier: 10.1109/Iccias.2006.295304; Publication Year: 2006, Page(S): 1470 – 1474

[11] F. Liu; Z. Li And J. Yu, "Applications Identification Based On The Statistics Analysis Of Packet Length," Information Engineering And Electronic Commerce, 2009. IEEC '09.Weinternational Symposium On Digital Object Identifier: 10.1109/Ieec.2009.38;Wepublication Year: 2009, Page(S): 160 – 163

[12] C. Wang, T. Li And H. Chen, "P2P Traffic Identification Based On Double Layer Characteristics," Information Technology And Computer Science, 2009. ITCS 2009.Weinternational Conference On Volume: 2, Digital Object Identifier:We10.1109/Itcs.2009.298, Publication Year: 2009, Page(S): 593 – 596

[13] B. Xu, M. Chen, F. Lan And N. Wang, "P2P Flows Identification Method Based On Listening Port," Broadband Network & Multimedia Technology, 2009. IC-BNMT '09. 2nd IEEE International Conference On Digital Object Identifier:We10.1109/Icbnmt.2009.5348496 Publication Year: 2009, Page(S): 296 - 300

[14] R. Zhang, Y. Du And Y. Zhang, "A BT Traffic Identification Method Based On Peer Cache," Internet Computing For Science And Engineering (ICICSE), 2009 Fourth International Conference On Digital Object Identifier: 10.1109/Icicse.2009.39wepublicationweyear: 2009, Page(S):We320 - 323

[15] D. Zhang, C. Zheng, H. Zhang And H. Yu, "Identification And Analysis Of Skype Peer-To-Peer Traffic," Internet And Web Applications And Services (Iciw), 2010 Fifth International Conference On Digital Object Identifier: 10.1109/Iciw.2010.36; Publication Year: 2010, Page(S):We200 - 206

[16]http://www.winpcap.org