RESEARCH ARTICLE OPEN ACCESS

Deep Packet Inspection in the days of Encryption

Manasvin A*

*(Department of ISE, RVCE, Bangalore
Email: manasvina.is17@rvce.edu.in)

--************************----------------------------------

Abstract:

Deep Packet Inspection (DPI) refers to a technique of analysis, performed on the packet payloads being sent over a network, that can be used for obtaining information, intrusion detection, classifying, rerouting, filtering, etc the packets. Such analysis holds a lot of value in the real world due to its ability to convey details about the traffic seen on the network, which can be of high use to people like the Internet Service Providers (ISP). ISPs can use such details to get a clearer understanding of their subscriber traffic usage and make smarter decisions. However, a problem encountered in the real world is that most traffic being sent over the internet is encrypted (i.e., HTTPS). Hence there is a crossroads between not breaking the privacy and performing the analysis.

This paper explores the ways, such as using neural networks and using new encryption schemes, that DPI can be extended to work on top of encrypted traffic, thus protecting the data while still being able to classify and analyze. The paper also looks at the impact of certain legal frameworks on DPI.

Keywords **—Deep Packet Inspection, Encryption, payload, machine learning, regular expressions**----

---------------------------------************************------------------------------------

I. INTRODUCTION

The global Deep Packet Inspection market is projected to rise to $16620 Mil. By 2026, from the $4350 Mil it's at as of late 2020.[15] This growing demand of DPI, even when there a bunch of challenges facing implementation of it is what sparked the need for this paper, to look at relevance of DPI in today's world.

Deep Packet Inspection plays a vital role in various network traffic data applications such as detection of intrusion [1], application visibility under VPN, preventing leakage of sensitive data, etc. The main common process amongst all these is the inspection of packet payload [2].

One area where use of DPI has gained a lot of importance, as the internet grows more and more is network classification [3]. There are various methods and techniques that are being used to perform this, some of which shall be discussed later in this paper.

However, an important point to note is that parallel to the rise of DPI in analysing network traffic is the rise in usage of encryption protocols to transfer data over the internet. Whilethis provides a greater amount of confidentiality and security, it provides a major roadblock to packet payload inspection.

There are a variety of ways that proposed solutions tackle this hurdle of encryption. Some systems involve decryption of the network traffic before it reaches receiver and performing the analysis. While it enables DPI, it exposes the system to various issues [4], and there is also, from the users' point of view, the prospect of this decrypted private information being sold to 3rd party marketers.[5]

The more acceptable solutions involve maintaining of both the features: encryption to provide degree of confidentiality, and functionality of the DPI module.

This paper explores the various methods ranging from inspecting the encrypted payload directly [2], to using CNNs to identify malware traffic [6], to using autoencoders along with CNNs to perform network traffic classification [3].

The paper also looks at the effect that legal frameworks like GDPR [7], implemented by the European Union (EU) have had on the implementation of DPI.

II. LITERATURE SURVEY

Anat Bremler-Barr et al [10] talk about treating DPI as a service, from the perspective of the middleboxes that are present on the network. This study suggests looking at DPI as a service, thus these middle boxes scan the incoming traffic just one time, but in context of the data of all such boxes that use the service. The information or results of this service are then passed onto next middleboxes. This study showed that such an architecture would provide great improvements in terms of scalability, performance and robustness.

Kumar et al [11] talks about usage of regex in classifying the network traffic. The internal working of parsing these regexes is usually by DFAs. This study however, argues that to overcome the large memory usages of DFAs and the low throughputs of other automatas like D2FA or NFA, a new form of automata Content Addressed Delayed Input DFA (CD2FA). CD2FA has throughput comparable to D2FA while providing a compact way of regex representation. (uses ~10% of space as DFA).

Sherry et al [2] propose a system called BlindBox that does packet payload analysis directly on the encrypted trafficpacketsss. They achieve this via new encryption schemes and protocols.

Dainotti et al [8] talk about a simple traffic classification system based around port and port numbers. Such systems use the TCP/UDP headers present in the packets to identify the port number. This port number which will be associated with a particular application.

Lotfollahi et al [3], propose a framework that uses Deep Learning (DL) to extrapolate network traffic features and classify traffic. The proposed solution was also tried out on application identification, and showed results that were better than most systems. The algorithms used include a stacked autoencoder and a CNN, which took in the network traffic once certain baseline preprocessing such as data link header removal, byte conversion, normalization was done.

Wang et al [6] proposed a solution that used a 2D CNN for the purpose of a binary classification of normal or malware for the encrypted network traffic packet.

Lopez-Martin et al [12] looked at impact of using recurrent neural network-based solutions. The combo of LSTM and 2D-CNN working on traffic in time series form provided an accuracy of 96.32% for encrypted traffic classification.

Gil et al. [13] used attributes such as the flow duration, flow rate (bytes /s), inter-arrival time, etc. to characterize the network traffic using k-nearest neighbour (k-NN) and C4.5 DT algorithms. Results were 92% recall for classifying web browsing, email, chat, streaming, file transfer and VoIP. Results were 88% recall when the same data was fed but tunnelled through a VPN.

Parish, Wang [14] used probability density on the packet size to classify and identify encrypted network traffic protocols such as FTP/IMAP/SSH/TELNET. Results were around 87% accuracy.All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

III. COMPARISION OF DPI SOLUTIONS

The main streams along which most DPI solutions fall under are port based, traditional payload inspection and machine learning based.

Port based solutions, as mentioned, work by identifying the port number from the TCP/UDP headers and matching these numbers to the application. Since they don't have to worry about whether traffic is encrypted or not, they result in fast retrieval tomes. Hence, they are ideally used in firewalls and access control lists. [9]. The problem, however, is that most modern network connections have features like port forwarding, obfuscation, NAT, etc greatly impact this method in the negative sense. Studies have shown that when worked onmodern internet traffic, only around 30% of traffic can be accurately classified.[16]

Traditional DPI solutions are those that are entirely based on REGEX and signatures [17][19]. The way these solutions work is that payload is checked against the patterns and any matches result in the packet being classified appropriately. This has a major disadvantage in that these signatures are static, thus any change or update in the protocol results in the requirement for new regex/signatures. These also have a low scope for work on encrypted data. The previously mentioned solution, BlindBox [2] a solution that comes under this category but is able to handle encrypted traffic.

The type of DPI solutions where a lot of studies are being conducted is ML based solutions. These solutions utilize some mathematical function/stats to take advantage of the supposedly unique features that each applications traffic possesses.

Studies have been conducted using algorithms such as Bayesian neural network, naïve bayes classifier, ANN, etc. Some of these solutions have been discussed under literature survey.

The popular consensus is that machine learning based solutions for DPI on encrypted traffic are far more effective than port based or traditional DPI solutions [18].

The main advantages that a ml/dl model-based solution has over a signature-based solution is that the ml solution learns the unique pattern by inference and does not require a human to have analysed the data streams, thus it is much less expensive. Furthermore, since ML models can keep learning while running/analysing, thus making them more flexible and quicker to adapt.

On the other hand, a disadvantage that ML based DPI solutions have is that they are susceptible to producing false positives, especially if model focusses on traffic related metrics over inspection on payload.

IV. COMPARISION OF DPI SOLUTIONS

Internet traffic is increasingly being encrypted before being communicated. This clear trend has provided end users of the internet with a greater sense of security and privacy. But what has been an unintended consequence of this is that malware and malicious activities are now behind this layer of security, thus making their identification and prevention that much harder.

In today's world around 80% of the internet traffic is encrypted. This represents a huge channel for attackers to try to send their malicious attack.

More than 67% of all malware attacks are deployed by means of email. These emails which are often encrypted with HTTPS protocol.

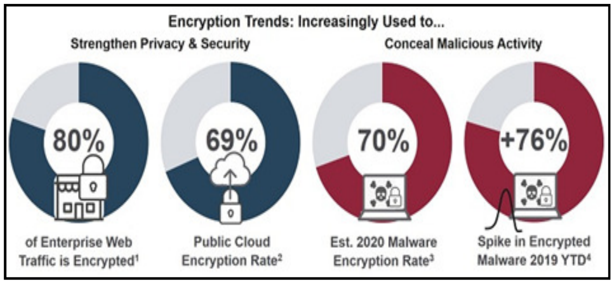

Fig 1. Encryption Trends

Thus, there is strong need to extend DPI for IDS, malware detection. In the recent years, companies such as JPMC, Equifax, the IRS have been breached costing them huge amounts in damages.

As discussed, studies like Wang et al [6] have shown a promising scenario wherein machine learning and DPI are used in conjunction to build a malware detection system which works on encrypted traffic and shows results with great accuracy.

V. IMPACT OF LEGAL REGULATIONS

General Data Protection Regulation (GDPR) is a 2018 EU law based on data protection and privacy. This put data protection at the forefront for many companies' agendas. Collection or reading of information such as IP addresses, user ID, etc that are considered as Personally Identifiable Information (PII) has been barred. This has made DPI more complex whether it be for stopping attacks or for traffic analysis.

While this has made DPI for intrusion detection harder, such legal frameworks have brought into conversation the need for using DPI in another field, that of data loss prevention.[20][21][23]

GDPR is explicit in stating that cloud providers will not be exempt from GDPR violations. Thus, they have a demand to put in place checks and rules to prevent any accidental or willing upload of confidential end user data such as phone no., address, card details, etc which fall under PII, to a public server or personal mail.

This can be overcome by using DPI for data loss prevention. There are products in the market [22] which utilize DPI on SSL/TLS encrypted traffic with signatures/regex to inspect the encrypted traffic to identify the GDPR PII violations in realtime. These identified violations are in real time blocked and reported.

VI. CONCLUSION

Deep Packet Inspection is a vital field of network traffic analysis that is expected to grow even further in the coming years. Whether the use case be about identifying and blocking malware, or about ISPs identifying network subscriber internet usage details and tailoring internet services to better service them, or about targeted advertising, or preventing sensitive data from being leaked, or parental content regulation, etc DPI provides a lot of people with a lot of advantages.

However, in today's world there is just as big a need to ensure that user details aren't accessed by 3rd parties or middleboxes to either gain access to private information or to sell such data to marketers. Thus, there is a need for DPI solutions to be built with encrypted traffic in mind, and provide functionality without having to break the encryption layer.

The various legal frameworks concerning user data privacy vary by region, and the solutions built must be compliant of all the appropriate frameworks.

This paper hoped to look at the state and relevance of Deep Packet Inspection in today's world, and explore the various solutions available currently, and compare their advantages and use cases.

REFERENCES

[1] R. T. El-Maghraby, N. M. Abd Elazim and A. M. Bahaa-Eldin, "A survey on deep packet inspection," 2017 12th International Conference on Computer Engineering and Systems (ICCES), 2017, pp. 188-197, doi: 10.1109/ICCES.2017.8275301.

[2] Justine Sherry, Chang Lan, Raluca Ada Popa, and Sylvia Ratnasamy. 2015. BlindBox: Deep Packet Inspection over Encrypted Traffic. SIGCOMM Comput. Commun. Rev. 45, 4 (October 2015), 213–226. DOI:https://doi.org/10.1145/2829988.2787502

[3] Lotfollahi, M., Jafari Siavoshani, M., Shirali Hossein Zade, R. et al. Deep packet: a novel approach for encrypted traffic classification using deep learning. Soft Comput 24, 1999–2012 (2020). https://doi.org/10.1007/s00500-019-04030-2

[4] J. Jarmoc. SSL/TLS Interception Proxies and Transitive Trust. Presentation at Black Hat Europe,2012

[5] N. Vallina-Rodriguez, S. Sundaresan, C. Kreibich, N. Weaver, and V. Paxson. Beyond the Radio: Illuminating the Higher Layers of Mobile Networks. In Proc. ACM MobiSys, 2015

[6] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using CNN for representation learning," in IEEE ICOIN'17

[7] gdpr-info.eu

[8] Dainotti A, Pescape A, Claffy KC (2012) Issues and future directions in traffic classification. IEEE network 26(1)

[9] Qi Y, Xu L, Yang B, Xue Y, Li J (2009) Packet classification algorithms: From theory to practice. In: INFOCOM 2009, IEEE, IEEE, pp 648–656

[10] Anat Bremler-Barr, Yotam Harchol, David Hay, and Yaron Koral. 2014. Deep Packet Inspection as a Service. In Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies (CoNEXT '14). Association for Computing Machinery, New York, NY, USA, 271–282. DOI:https://doi.org/10.1145/2674005.2674984

[11] Sailesh Kumar, Jonathan Turner, and John Williams. 2006. Advanced algorithms for fast and scalable deep packet inspection. In Proceedings of the 2006 ACM/IEEE symposium on Architecture for networking and communications systems (ANCS '06). Association for Computing Machinery, New York, NY, USA, 81–92. DOI:https://doi.org/10.1145/1185347.1185359

[12] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, “Network traffic classifier with convolutional and recurrent neural networks for Internet of Things,” IEEE Access, vol. 5, pp. 18 042–18 050, 2017.

[13] Gil GD, Lashkari AH, Mamun M, Ghorbani AA (2016) Characterization of encrypted and vpn traffic using time-related features. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016), pp 407–414

[14] Wang X, Parish DJ (2010) Optimised multi-stage tcp traffic classifier based on packet size distributions. In: Communication Theory, Reliability, and Quality of Service (CTRQ), 2010 Third International Conference on, IEEE, pp 98–103

[15] reports.valuates.com/market-reports/QYRE-Othe-0H237/deep-packet-inspection

[16] Madhukar A, Williamson C (2006) A longitudinal study of p2p traffic classification. In: Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. MASCOTS 2006. 14th IEEE

[17] Yeganeh SH, Eftekhar M, Ganjali Y, Keralapura R, Nucci A (2012) Cute: Traffic classification using terms. In: Computer Communications and Networks (ICCCN), 2012 21st International Conference on, IEEE, pp 1–9

[18] S. Cui, B. Jiang, Z. Cai, Z. Lu, S. Liu and J. Liu, "A Session-Packets-Based Encrypted Traffic Classification Using Capsule Neural Networks," 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2019, pp. 429-436, doi: 10.1109/HPCC/SmartCity/DSS.2019.00071.

[19] R. T. El-Maghraby, N. M. Abd Elazim and A. M. Bahaa-Eldin, "A survey on deep packet inspection," 2017 12th International Conference on Computer Engineering and Systems (ICCES), 2017, pp. 188-197, doi: 10.1109/ICCES.2017.8275301.

[20] Tahboub, Radwan & Saleh, Yousef. (2014). Data Leakage/Loss Prevention Systems (DLP). International Journal of Information Systems. 1. 13-19. 10.1109/WCCAIS.2014.6916624.

[21] Geong Sen Poh and Dinil Mon Divakaran and Hoon Wei Lim and Jianting Ning and Achintya Desai "A Survey of Privacy-Preserving Techniques for Encrypted Traffic Inspection over Network Middleboxes", arXiv:2101.04338 [cs.CR], 2021.

[22] sonicwall.com/products/firewalls/

[23] D.H. Sharma; C.A. Dhote; M.M. Potey, "Managed data loss prevention security service in cloud" 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016), 2016 page (4 pp.)