A Survey on Tor Encrypted Traffic Monitoring

Mohamad Amar Irsyad Mohd Aminuddin[1],
Zarul Fitri Zaaba*[2], Manmeet Kaur Mahinderjit Singh[3]
School of Computer Sciences
Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia

Darshan Singh Mahinder Singh[4]
Centre for Drug Research
School of Computer Sciences
Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia

***Abstract*—Tor (The Onion Router) is an anonymity tool that is widely used worldwide. Tor protect its user privacy against surveillance and censorship using strong encryption and obfuscation techniques which makes it extremely difficult to monitor and identify users' activity on the Tor network. It also implements strong defense to protect the users against traffic features extraction and website fingerprinting. However, the strong anonymity also became the heaven for criminal to avoid network tracing. Therefore, numerous of research has been performed on encrypted traffic analyzing and classification using machine learning techniques. This paper presents survey on existing approaches for classification of Tor and other encrypted traffic. There is preliminary discussion on machine learning approaches and Tor network. Next, there are comparison of the surveyed traffic classification and discussion on their classification properties.**

***Keywords*—Encrypted traffic monitoring; Tor; machine learning; security; survey**

I. Introduction

Tor (The Onion Router) is a well-known anonymity network globally [1]. The popularity of Tor is undeniable as the Tor's users increase up to two million just in year of 2017 alone [2] and currently there are more than four million users worldwide. Similar with other anonymity project such as I2P [3] and FreeNet [4], Tor primary goal is to provide users with privacy protection and anonymous access to the Internet. This will facilitate the Internet users with mechanism to hide their activity thus protecting their privacy to some extent, i.e. to hide the source, the destination, and the nature of the communication, other than encrypting the content itself [35]. Although there are lots of good usage practice, Tor however are also dual-use networks just like other technology such as BitTorrent (where users are not only use it to share free materials, they also share copyrighted materials) since it has been exploited for illegal activities purposes [5]-[8].

Due to the complexity nature of the Tor encrypted traffic, the research community has put considerable effort on analyzing the Tor security especially on the possibility of deanonymizing the Tor users [11]-[15]. These researches have focused on decoy traffic [12], exit router logging [11], attack on identifying Tor relays [14] and investigation on exit relays trustworthiness [15]. Although the result is promising, but these techniques lack of security monitoring proficiencies on the Tor network. In security monitoring perspective, the Tor traffic in general should be monitored and analyzed to obtain useful knowledge such as information on the website that Tor's user access. Therefore, there are few has focused on the privacy disclosure based on identifying application information of traffic in the Tor network [10]. This approach would allow for the Tor traffic to be monitored in large scale and real-time. Although it is not directly deanonymized the user activity on the Tor network, learning the application information that being used by the users in the Tor network is a part of the Tor privacy concern [33]. This security monitoring capabilities could be achieved using machine learning classification technique similar to the classification of encrypted traffic on the surface web.

Even though machine learning classification for encrypted traffic has been studied intensively [17], [48]-[50], the process of applying these studied to classify the Tor traffic is remarkably challenging due to valuable traffic features that obtained from previous study are irrelevant in the context of Tor network and Tor itself developed with strong anonymity protection [31].

The main objective of this paper is to survey the application of machine learning techniques for encrypted Tor traffic classification and several latest clearnet encrypted traffic classification studies. Based on the results of the investigation, we will discuss and compared comprehensively on those machine learning techniques and its operation.

The rest of the paper is organized as follows. In Section 2, will be discussion on the technical background of traffic encryption on the Tor network. In Section 3, we discuss the fundamental of machine learning in traffic classification. Section 4 contain the discussion of studies on machine learning techniques. Section 5 contains the comprehensive comparison and consideration on these surveyed techniques. Finally, the paper is concluded in Section 6.

II. Tor Background

A. Onion Routing

As Tor created to provide anonymity services that allow people to improve their privacy and security on the Internet, it is run by group of volunteer-operated servers around the world. The main backbone of Tor network is the distributed relay server (Tor node) which providing the onion routing capabilities. Onion routing is a concept of anonymous communication over a computer network where the messages are encapsulated in multiple layers of encryption [18]. The encrypted data is transmitted through series of Tor nodes (current Tor implementation use three nodes [1]) which is called as a Tor circuit. Each of the node will decrypt a layer of

encryption to uncover the next destination of the traffic without the knowledge on whether the source of the traffic is coming from a Tor client or from another Tor node. Only the exit (third) node know the true destination of the messages. Hence, there are no node that has both information on the source and destination of the messages.

Fig. 1 shows the Tor circuit example. Notice that the connection between the exit node and destination server does not encrypted. This is because Tor provide traffic encryption mechanism while the messages are in the Tor network [19]. The moment that messages go out of the Tor network, it is up to the users whether the traffic is encrypted or not. As an example, if a user accesses a HTTPS (Hyper Text Transfer Protocol Secure) websites, not only the communication is encrypted in the Tor network, but also encrypted outside of the Tor network through the HTTPS. If the user accesses an unsecure HTTP website, the communication is encrypted only in the Tor network and no encryption provided outside of the Tor network (situation in Fig. 1).

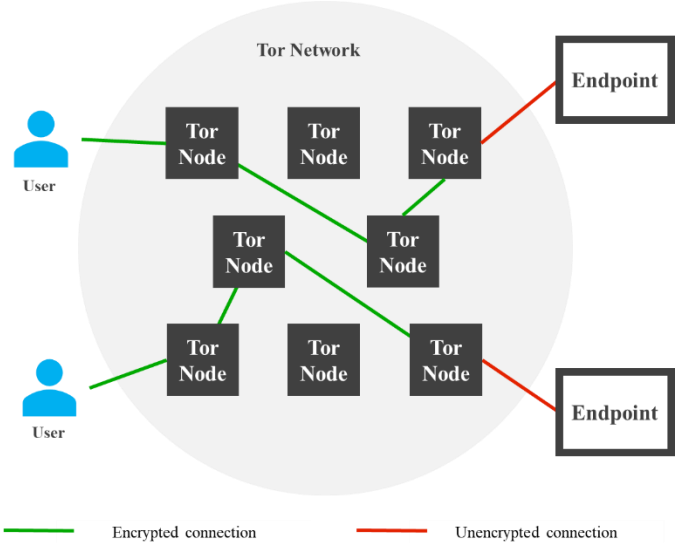

Fig. 1. Tor circuit example.

B. TLS

The encrypted communications inside the Tor network heavily relied on the TLS (Transport Layer Security). (TLS) is a cryptographic protocol designed to provide secure communication over the network [20] which could be considered as the successor of the Secure Sockets Layer (SSL) [21]. TLS is located between the Transport Layer and the Application Layer in the TCP/IP Model Layer. Other than encryption, the fundamental of TLS is the used of X.509 certificates as the verification features of with whom they are communicating, asymmetric cryptography to authenticate those entity and symmetric session key as the key to encrypt the data transfer between the entity. In addition, TLS use Record Protocol [20], which acts as a wrapper that responsible for dividing messages into several fragments. These fragments than will be paired with its corresponding Message Authentication Code (MAC). These procedures are important to ensure that the messages transferred from source to the destination accessible by the right party and no third party could involve or do modification to these messages without the right party awareness. Other than Tor, TLS is widely use in the network environment to secure the data transfer; for example, as web browsing, email, instant messaging, and voice-over-IP (VoIP).

The knowledge on TLS protocol mechanism is very crucial in Tor traffic classification. Fig. 2 shows the Tor traffic layer. Tor will divide communication messages into several fixed size packets which called as cell [22]. Then, these cells are processed and transformed into TLS records. These records than will be fragmented into the TCP packets before it will be send to another Tor node. These traffic layer mechanism enable features extraction of the traffic at three different levels for the machine traffic classification process as each layer has certain header information that are not encrypted [33].

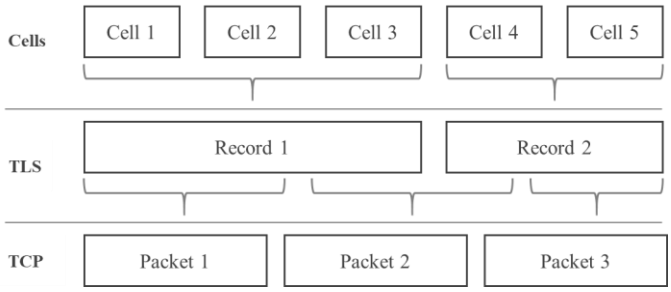

Fig. 2. Tor traffic layer [22].

III. Machine Learning

In computer science, machine learning is a technique that enable computer to learn from experience with data using statistical techniques rather than explicitly being programmed. In computer security, machine learning has been utilized in lots of area such as traffic classification, anomaly detection, spam detection, malware identification and entity classification [23].

Fig. 3 shows traffic classification taxonomy. Below is the discussion of broad categories of machine learning approaches, classification input features, classification output classes and evaluation metric of classification algorithms.

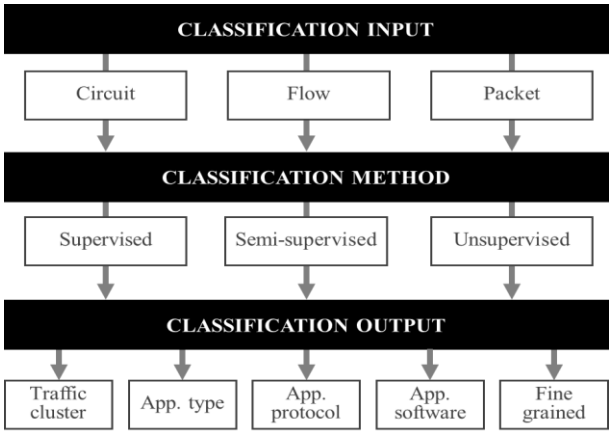

Fig. 3. Traffic classification taxonomy.

A. Machine Learning Approaches

There are three broad approaches of machine learning in traffic monitoring which are supervised, semi-supervised and unsupervised. In traffic classification, different approach produces different result, effectiveness and reliability depend on its target and the dataset that being used as the input.

Supervised learning used sets of pre-training data (data that are labelled based on certain traffic features) to train the algorithms on classification of the traffic. The example of supervised learning algorithms is K -Nearest Neighbors (k-NN) [24], Bayesian Network [25], Decision Tree [26] and Support Vector Machine (SVM) [27]. The main advantage is that it produces low false rate of classification. Yet, it might become challenging in providing complete sets of pre-training data and might require long training time.

Semi-supervised learning is similar to the supervised, but it does not utilize complete pre-training data. Instead the pre-training data is only partial labelled data. The main advantage is that there are less sets of pre-training data need to be provided. However, the accuracy might be a huge issue especially in classification various kind of encrypted application traffic.

Unsupervised learning in the other hand use mainly for data clustering without any pre-training data. The example of unsupervised learning algorithms is Fuzzy C-means [28] and K-means [29]. Despite the ease of usage as no training data is required, it commonly utilized for traffic clustering process rather than for encrypted traffic classification. One of the main usage of unsupervised traffic clustering is in anomaly detection which is work well with unlabelled traffic [29], [47], [51].

B. Classification Input

The pre-training data that will be used to train supervised or semi-supervised algorithms are fundamentally important in receiving the best output of traffic classification. Thus, selection of input data need to be done carefully and thoroughly. The Tor traffic classification input data could be segmented into three categories.

- **Circuit** – Circuit lifetime, Cell inter-arrival times, Cells per circuit life time, Uplink Cells and The Rate of the Downlink Cells to the Uplink Cells.
- **Flow** – Flow segment size, round trip time, duration.
- **Packet** – Packet length, frequency, header.

C. Classification Output

Based on the classification input, there are several types of classification output could be obtained. The best output is the fine-grained level which has the most detailed information on the classified traffic.

- **Traffic cluster (TC)** – Tor, Bulk/small transactions,
- **Application type (AT)** – Streaming, browsing, torrenting.
- **Application protocol (AP)** – HTTPS, FTP, P2P
- **Application software (AS)** – Web browser, Mail client
- **Fine-grained (FG)** – Facebook site, YouTube video, Skype call, Windows update.

D. Evaluation Metrics

As lot of researches has been carried on machine learning techniques for traffic classification, these researches need to be evaluated and compared with established findings. To evaluate the outcome of each machine learning method for traffic classification, several important metrics are used [16] which are accuracy, F-Measure, True Positive Rate (TPR), False Positive Rate (FPR), precision and recall.

IV. Traffic Classification Techniques

This section surveys on machine learning approaches of traffic classification. The discussion will be divided into two main categories which are encrypted traffic classification that work on the Tor network and latest encrypted traffic classification on non-Tor network. Table I shows summary of cite papers and machine learning method on traffic classification.

A. Traffic Classification on Tor

Alsabah et al. [31] proposed and evaluated DiffTor that classify real-time Tor circuit using machine learning algorithms. The main intention of this study is to improve the performance of the Tor network using classification process that assigns distinct classes of services on each application traffic in the Tor circuit. Based on their observation, different applications have different time and throughput requirements. Therefore, the chosen attributes are circuit lifetime, amount of data transfer, cell inter-arrival times and number of recent cells sent are selected to classify the Tor traffic. The authors confirmed that their experiments managed to classify Tor circuit which being generated on the live Tor network with extremely high accuracy.

Similar classification technique carried out by [32] that focus on circuit and traffic flow classification. The circuit classification retrieved data at Tor's relay server and the flow classification retrieved data that are transmitting anywhere between relay server and Tor's user. This flow classification approach does not require any access on the relay server which make it much more flexible that the circuit classification approach. For circuit level classification, attributes such as cells per circuit lifetime, Uplink cells, rate of Downlink cells to Uplink cells and Exponentially Weighted Moving Average (EWMA) are chosen. For flow level classification, the authors use two flow exporting tools (Tranalyzer2 and Tcptrace) which able to generate flows and extract the attributes of the flows. The author managed to perform study on four machine learning algorithms which are Naïve Bayes, Bayesian Networks, C4.5 and Random Forest.

There is study on application classification attack on Tor network [33] which identify and classify the application types inside the Tor network traffic. The identification is based on application behaviour that characterized by traffic flow features such as burst volumes and directions. The authors define burst as successive packets in between two packets sent on the opposite direction. Based on Tor's design features

which utilize round-robin fashion for Tor scheduling, the relation between burst volumes and burst directions could be utilized as the critical features in classifying the Tor traffic. This research utilized unsupervised K-means and Multiple Sequence Alignment (MSA) to pre-treat sample data and Profile Hidden Markov Model to build model and classify the application type of the Tor traffic.

Lingyu et al. [34] proposed hierarchical classification that exploits decision tree algorithm for Tor traffic identification and Tri-Training algorithm for Tor traffic segmentation. Tri-Training algorithm is a semi-supervised machine learning algorithm which utilize co-training technique [52]. It has certain advantages such as it requires low number of training data than supervised methods, does not require cross-validation and no restriction on base classifier. Rather than focus on cell-based attributes like in [31] and [32], this study had focus on packet-based attributes which are packet length entropy, 600-byte packet frequency, zero data packet frequency (first 10) and average packet interval time. The result shows high accuracy classification which could be achieved due to the hierarchical instrument.

There is also classification study based on Anon17 dataset [30] that uses four classifier approaches (Naïve Bayes, Bayesian Network, C4,5 and Random Forest) [35]. The public dataset contains traffic from three popular anonymity services (Tor [1], I2P [3] and JonDonym [30]). The authors performed three level of classification beginning with Anon Network (Tor, I2P, JonDonym), Traffic Type (Normal, Tor Apps, I2P Apps) and Application (Tor, Streaming, Torrent, Browsing). The result of these experiment shows that all classification levels on these anonymity services could be identified and distinguished with high accuracy. There are 81 extracted features for classification including flow direction, packet length, inter-arrival time, IP header features and number of connections during traffic flow lifetime. This experiment is unique from others since it uses dataset that are publicly available. However, due to the dataset is only recently available, there lack of studies that utilize the same dataset currently.

Soleimani et al. [38] has focus on identification of Tor pluggable transports using machine learning techniques. Tor pluggable transport is a bridge from the Internet into the Tor network which considered as the technique to bypass the worldwide Tor censorship operation [45]. This experiment proceeds on three plugin techniques which are Obfs3, Obfs4, and ScrambleSuit. Using supervised learning, the identification of these plugins could be executed with only first 10-50 packets inspection in real-time. The authors utilize statistical flow features such as flow size (both direction), mean size of packet sent (both direction) and standard deviation of packet sizes (both direction).

Based on local network observer of Tor traffic dataset, [43] has analysed and found that standard HTTPS traffic (related to top monitored sites on Alexa) and Tor network has variations that could be classified using the machine learning technique. The authors generate traffic using virtual machines with two different instances. One with HTTPS traffic and the other is Tor-based traffic (both access similar website). The authors use 40 features including total packets, total bytes, smallest packet size, largest packet size, minimum (including maximum and mean) amount of time between two packet and duration of flow. Due to the proven variation of traffic features of HTTPS and Tor network, the study outcome is a fine grained output of traffic classification (classify which traffic related to which website).

Cuzzocrea et al. [44] also presented technique that identifies Tor-related traffic that generated on a host. Hence, it could detect whether a user is using Tor application. The identification process uses supervised classification based on traffic flows features. Similar to others Tor network classification [35] and [43], the chosen attributes are 23 including flow duration, flow bytes per second, flow inter-arrival time and flow active time. This study had been carried out on six machine algorithms with the most accurate result is J48 (C4.5) approach.

B. Traffic Classification on Non-Tor Network

Fu et al. [39] has developed CUMMA, a system that use machine learning for in-App service usage classification on encrypted traffic in mobile messaging apps. This system learning model is based on temporal dependencies, user behavioural patterns and network traffic characteristics. It also works closely on time series classification and features segmentation. The traffic features extraction including packet length related features (such as descriptive statistics, variances in directions, and hopping counts) and time delay related features (such as time interval for consecutive packet). The outcome of this project shows that CUMMA enable service usage identification and application usage behaviour detection (text, picture, audio note, stream video call) based on encrypted traffic classification.

Another supervised research for encrypted traffic from [40] has proposed an attribute-aware classification that utilized second-order Markov Chain algorithm. Second-order Markov Chains is required in order to determine state transition probabilities. This study also proposes modelling process using application attribute bigram that able to increase second-order Markov Chains state diversity. The study attained better discernment accuracy and diverse application fingerprints through the leverage of the attribute bigram (Certificate and first Application Data packets). This experiment manages to classify the encrypted traffic based on the detection of website traffic-attribute which could be considered as fine grained output.

Sun et al. [41] has studied on incremental SVM (ISVM) model that focus on enabling quick and high-frequency of classifier update with reduced training cost of memory and CPU. The essential different on ISVM is that the original training data is removed and only holds the Support Vectors (SVs) produced in latest updating process which overcome the traditional SVM weaknesses. They also proposed AISVM, an ISVM with attenuation factor through the use of weight on each SVs to maximize the usage of information on SVs updating process. The outcome of this study illustrates that both proposed model shows similar classification accuracy with traditional SVM but with significant reduced updating process.

Fan et al. [36] has study on the machine learning classification that emphasizes traffic in Software Define Networking (SDN). SDN [46] is a new network paradigm that provide simplification of network management and support for exponential traffic growth on mobile cellular network. The feature selection (such as port number, number of unique data bytes, maximum segment size and initial window bytes) techniques had been use as the manipulated variable in this study as different combination of selected features produced different classification accuracy. This study employs SVM and K-means clustering for the traffic classification process. The outcome shows that K-means are effectively clustering new type of traffic; however, it has lower accuracy than supervised SVM approach.

Another SVM learning approach has been carried out by [37] that focus on real-time traffic classification. The author proposed SSP-SVM based on principal component analysis (PCA) and scaling dataset to extract and verify traffic features. It adapts the reduction on feature dimension, lower features redundancy and higher features generalization. They also utilize improved particle swarm optimization algorithm to automatically produce optimal parameter for kernel function. The outcome performance shows that two-class and multi-class classifier work effectively on traffic classification compared to the traditional SVM.

Another novel fingerprinting technique (through sampling of Application Protocol Data Units exchange patterns) for traffic classification has been proposed by [42]. The studied technique is simple to be implemented with minimal resource requirement. The principal of this technique is to provide high efficient sampling strategy that applied by single Content Addressable Memory (CAM) filtering rule based on zero-length TCP packet flows. This classification technique isn't affected with network transmission issues such as fragmentation, loses and congestion. The author also suggested that for UDP traffic, analogous fingerprinting scheme should be utilized to attain the same accuracy in TCP.

V. Discussion and Comparison

We have provided the summarize overview on the encrypted traffic classification using machine learning approaches. Most surveyed studies use flow and packet features as the input for the traffic classification technique. The authors of [35] manage to do classification based on both flow and packet features using public dataset [30] that are focus on the anonymity services traffic. There is also traffic classification based on circuit features [31], [32] that could be considered one of specialized machine learning process on the Tor network. This is because, the Tor circuit is only exist in the Tor network, thus it could be experimented and analysed entirely in the Tor network.

TABLE I. Summary Table of Cite Papers and Machine Learning Method on Traffic Classification

Reference	Publication Year	Traffic Properties			Learning			Methods												Real-time	Data set		Tor compatible	Output
		Circuit	Flow	Packet	Supervised	Semi-Supervised	Unsupervised	AdaBoost	K-Means	Naïve Bayes	Bayesian Networks	Decision Trees	C4.5	Random Forest	Hidden Markov model	Tri-Training	Support Vector Machine	Second-order Markov Chain			Public	Private		
[31]	2012	✔			✔					✔	✔	✔								✔		✔	✔	AT
[32]	2014	✔	✔		✔					✔	✔		✔	✔								✔	✔	AT
[43]	2015		✔		✔					✔			✔	✔			✔					✔	✔	TC,FG
[33]	2015		✔		✔		✔		✔						✔							✔	✔	AP
[39]	2016			✔	✔										✔							✔		AS,FG
[34]	2017			✔		✔						✔				✔						✔	✔	TC,AP
[35]	2017		✔	✔	✔					✔	✔		✔	✔							✔		✔	TC,AT
[36]	2017		✔		✔		✔		✔								✔				✔			AT
[37]	2017			✔	✔												✔			✔	✔			AP
[40]	2017			✔	✔													✔				✔		FG
[44]	2017		✔		✔						✔		✔									✔	✔	TC
[38]	2018		✔	✔	✔			✔					✔	✔			✔			✔		✔	✔	AP
[41]	2018		✔		✔												✔			✔	✔			AP
[42]	2018		✔		✔								✔							✔		✔		FG

Due to the complexity of encrypted traffic, most of the surveyed research focuses on supervised machine learning which required pre-training dataset to classify the traffic. The studies by [33], [36] utilized unsupervised machine learning solely for traffic clustering rather than traffic classification. Despite study by [34] managed to classify encrypted traffic using semi-supervised learning, this learning practice could be achieved with the presence of hierarchical classification approach.

The most popular machine learning methods are the C4.5 followed by the SVM. These two methods generally produce most accurate result. However, it might not be the best algorithm on most of encrypted traffic environment. Naïve Bayes, Bayesian Networks and Random Forest also produce satisfactory accuracy when the features are adequate and reliable.

Traditional non-Tor encrypted traffic classification has the privilege of accessing important attribute such as source and destination IP address and port number to learn and classify the traffic accurately. However, in Tor network, these attributes are impractical as the source and destination info are no longer the accurate attribute as it has been hidden by the onion routing mechanism. Hence, a lot of other attributes are used for classifying the Tor encrypted network traffic.

The real-time column in Table I shows that whether the classification process could be applied on real-time data. Although accurate detection is very important, some traffic classification such as anomaly detection [47] required real-time classification process to make it useful most of the time. Hence, heavy computational and slow traffic classification is impracticable in real-time environment.

The dataset column in Table I shows the source of data whether it was publicly available or privately collected. Traffic classification on public dataset would allow researchers to closely compared their accuracy and performance result with other published work. However, due to limited availability of dataset that involve Tor network, most of the studies that focus on Tor network produced their own dataset.

Compared to the traditional machine learning research [17], current researches on encrypted traffic classification using machine learning approach produced various type of output that could be further analysed and refined. Study by [39], [40], [42], [43] managed to perform encrypted traffic classification that identify traffic with fine granularity output of information.

To this end, there is no algorithm that performs the best in all condition. Different algorithm has different capabilities and efficiency depending on its classification objective, implementation strategy and training dataset. Therefore, there are several considerations factor that need to be taken while choosing the right machine learning algorithm [9].

- **Accuracy** – Despite the most accurate algorithm is very important, the processing time might be a huge bottleneck. In some cases, approximation of traffic classification is acceptable.
- **Training time** – When data set is huge, the training time of this data might be an important consideration as it might result the training time taken to be from several minutes to few hours. Encrypted traffic classification with shorter training time is much more preferred.
- **Computational resource** – Very accurate algorithm that utilized very high computational resource might not be suitable in certain traffic classification circumstances especially in real-time classification environment. Therefore, encrypted traffic classification algorithm needs to have acceptable computational requirement for more efficient processing resources.
- **Number of features** – Commonly, the more extracted features that available, the better accuracy of encryption traffic classification. However, it might become a bottleneck on the algorithm training time process. There are also features that will reduce the accuracy of the encrypted traffic classification which need to be avoided.
- **Number of parameters** – Number that affect algorithm behavior, such as number of iterations, error tolerance and options between variants. These parameters are very important in getting the effective and accurate result since different settings could significantly impact the encrypted traffic classification outcome.

VI. Conclusion

In the past, traffic classification using machine learning approaches was not important. With the emerging traffic encryption and anonymity services such as Tor, machine learning technique for encrypted traffic classification should be considered as the prominent approaches on identifying this Tor traffic.

In this paper, we presented an overview of machine learning classification for Tor encrypted traffic. We begin with discussion on Tor technical background and machine learning background. Then there is discussion on surveyed papers, summarize table of the discussed papers and comparison on the machine learning approaches. To sum up, there still lots of things could be further investigated and improved in the machine learning classification process to discover the truth of privacy protection on the Tor network.

References

[1] R. Dingledine, N. Mathewson, and P. Syverson, “Tor: The Second-Generation Onion Router,” SSYM’04 Proc. 13th Conf. USENIX Secur. Symp., pp. 21–22, 2004.

[2] Tor Project, Users – Tor Metrics, 2018. Retrieved 26 March 2018, from https://metrics.torproject.org/userstats-relay-country.html?start=2017-01-1&end=2018-01-1&country=all&events=off

[3] I2P: The Invisible Internet Project. Retrieved 26 March 2018, from https://geti2p.net/

[4] FreeNet. Retrieved 26 March 2018, from https://freenetproject.org

[5] M. Spitters, F. Klaver, G. Koot, and M. Van Staalduinen, “Authorship Analysis on Dark Marketplace Forums,” Proc. - 2015 Eur. Intell. Secur. Informatics Conf. EISIC 2015, pp. 1–8, 2016.

[6] FTR Team, 2016. Cybercrime and the Deep Web. Trend Micro Security News. Retrieved 1 March 2018, from https://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-cybercrime-and-the-deep-web.pdf

[7] J. Aldridge and R. Askew, "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement," Int. J. Drug Policy, vol. 41, pp. 101–109, 2017.

[8] J. Broséus, D. Rhumorbarbe, M. Morelato, L. Staehli, and Q. Rossy, "A geographical analysis of trafficking on a popular darknet market," Forensic Sci. Int., vol. 277, pp. 88–102, 2017.

[9] How to choose algorithms for Microsoft Azure Machine Learning, Microsoft Azure, 2018. Retrieved 30 March 2018, from https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

[10] I. Sanchez-Rola, D. Balzarotti, and I. Santos, "The Onions Have Eyes: A Comprehensive Structure and Privacy Analysis of Tor Hidden Services," Proc. World Wide Web Conf., pp. 1251–1260, 2017.

[11] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker, "Shining light in dark places: Understanding the tor network," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5134 LNCS, pp. 63–76, 2008.

[12] S. Chakravarty, G. Portokalidis, M. Polychronakis, and A. D. Keromytis, "Detecting traffic snooping in tor using decoys," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6961 LNCS, pp. 222–241, 2011.

[13] Y. Okada, S. Ata, N. Nakamura, Y. Nakahira, and I. Oka, "Application identification from encrypted traffic based on characteristic changes by encryption," 2011 IEEE Int. Work. Tech. Comm. Commun. Qual. Reliab. CQR 2011, 2011.

[14] P. Mittal, A. Khurshid, J. Juen, M. Caesar, and N. Borisov, "Stealthy traffic analysis of low-latency anonymous communication using throughput fingerprinting," Proc. 18th ACM Conf. Comput. Commun. Secur. - CCS '11, p. 215, 2011.

[15] P. Winter et al., "Spoiled onions: Exposing malicious Tor exit relays," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8555 LNCS, pp. 304–331, 2014.

[16] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm", Towards Data Science, 2018. Retrieved 1 April 2018, from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[17] P. Velan, M. Cermák, Pavel Celeda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," Int. J. Netw. Manag., vol. 25, pp. 355–374, 2015.

[18] D. Goldschlag, M. Reedy, and P. Syversony, "Onion Routing for Anonymous and Private Internet Connections," Network, pp. 1–5, 1999.

[19] Phobos (2010). Plaintext over Tor is still plaintext, Tor Blog. Retrieved 10 March 2018, from https://blog.torproject.org/plaintext-over-tor-still-plaintext

[20] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246 (Proposed Standard), Internet Engineering Task Force, Aug. 2008, updated by RFCs 5746, 5878, 6176.

[21] A. Freier, P. Karlton, and P. Kocher, "The Secure Sockets Layer (SSL) Protocol Version 3.0," RFC 6101 (Historic), Internet Engineering Task Force, Aug. 2011.

[22] A. Lazarenko and S. Avdoshin, "Anonymity of Tor: Myth and Reality," Proc. 12th Cent. East. Eur. Softw. Eng. Conf. Russ. - CEE-SECR '16, pp. 1–5, 2016.

[23] R. Marty, AI and Machine Learning in Cyber Security – Towards Data Science, Formulated.by, 2018. Retrieved 15 March 2018, from https://towardsdatascience.com/ai-and-machine-learning-in-cyber-security-d6fbee480af0

[24] S. Manocha and M. A. Girolami, "An empirical analysis of the probabilistic K-nearest neighbour classifier," Pattern Recognit. Lett., vol. 28, no. 13, pp. 1818–1824, 2007.

[25] R. E. Neapolitan, "Learning Bayesian Networks," Prentice Hall, vol. 6, no. 2, p. 674, 2003.

[26] J. R. Quinlan (1993). C4.5: programs for machine learning. Log Altos, CA, Morgan Kaufmann

[27] V. N. Vapnik, "An overview of statistical learning theory.," IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988–99, 1999.

[28] J. Bezdek, 1981. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981)

[29] H. Li, "Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis," 2010 Int. Symp. Intell. Inf. Process. Trust. Comput., pp. 458–462, 2010.

[30] Jondos GmbH. (2018). JonDo – the IP changer. Retrieved 13 April 2018, from https://anonymous-proxy-servers.net/en/jondo.html

[31] M. Alsabah, K. Bauer, and I. Goldberg, "Enhancing Tor ' s Performance using Real-time Traffic Classification Categories and Subject Descriptors," Proc. 2012 ACM Conf. Comput. Commun. Secur., pp. 73–84, 2012.

[32] K. Shahbar and A. N. Zincir-Heywood, "Benchmarking Two Techniques for Tor Classification," Comput. Intell. Cyber Secur., pp. 1–8, 2014.

[33] G. He, M. Yang, J. Luo, and X. Gu, "A novel application classification attack against Tor," Concurr. Comput. Pract. Exp., vol. 27, pp. 5640–5661, 2015.

[34] J. Lingyu, L. Yang, W. Bailing, L. Hongri, and X. Guodong, "A Hierarchical Classification Approach for Tor Anonymous Traffic," IEEE Int. Conf. Commun. Softw. Networks, pp. 239–243, 2017.

[35] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescape, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark," Proc. 29th Int. Teletraffic Congr. ITC 2017, vol. 1, pp. 81–89, 2017.

[36] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," 2017 Int. Symp. Wirel. Commun. Syst., pp. 1–6, 2017.

[37] J. Cao, Z. Fang, G. Qu, H. Sun, and D. Zhang, "An accurate traffic classification model based on support vector machines," Int. J. Netw. Manag., vol. 27, no. 1, pp. 1–15, 2017.

[38] M. H. M. Soleimani, M. Mansoorizadeh, and M. Nassiri, "Real-time identification of three Tor pluggable transports using machine learning techniques," J. Supercomput., pp. 1–18, 2018.

[39] Y. Fu, H. Xiong, X. Lu, J. Yang, and C. Chen, "Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps," IEEE Trans. Mob. Comput., vol. 15, no. 11, pp. 2851–2864, 2016.

[40] M. Shen, M. Wei, L. Zhu, and M. Wang, "Classification of Encrypted Traffic with Second-Order Markov Chains and Application Attribute Bigrams," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 8, pp. 1830–1843, 2017.

[41] G. Sun, T. Chen, Y. Su, and C. Li, "Internet Traffic Classification Based on Incremental Support Vector Machines," Mob. Networks Appl., pp. 1–8, 2018.

[42] J. Kampeas, A. Cohen, and O. Gurewitz, "Traffic Classification Based on Zero-Length Packets," IEEE Trans. Netw. Serv. Manag., vol. 4537, no. c, pp. 1–14, 2018.

[43] A. Almubayed, A. Hadi, and J. Atoum, "A Model for Detecting Tor Encrypted Traffic using Supervised Machine Learning," Int. J. Comput. Netw. Inf. Secur., vol. 7, no. 7, pp. 10–23, 2015.

[44] A. Cuzzocrea, F. Martinelli, F. Mercaldo, and G. Vercelli, "Tor Traffic Analysis and Detection via Machine Learning Techniques," IEEE Int. Conf. Big Data, pp. 4392–4398, 2017.

[45] Tor Project, Tor: Pluggable Transports, 2018. Retrieved 15 April 2018, from https://www.torproject.org/docs/pluggable-transports.html.en

[46] ONF, Software-Defined Networking (SDN) Definition - Open Networking Foundation, 2018. Retrieved 15 April 2018, from https://www.opennetworking.org/sdn-definition/

[47] S. Omar, A. Ngadi, and H. H. Jebur, "Machine Learning Techniques for Anomaly Detection: An Overview," Int. J. Comput. Appl., vol. 79, no. 2, pp. 975–8887, 2013.

[48] Z. Chen, L. Ruan, J. Cao, Y. Yu, and X. Jiang, "TIFAflow: Enhancing traffic archiving system with flow granularity for forensic analysis in network security," Tsinghua Sci. Technol., vol. 18, no. 4, pp. 406–417, 2013.

[49] Y. Wang, Y. Xiang, J. Zhang, W. Zhou, G. Wei, and L. T. Yang, "Internet traffic classification using constrained clustering," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 11, pp. 2932–2943, 2014.

[50] S. S. L. Pereira, J. L. De Castro E Silva, and J. E. B. Maia, "NTCS: A real time flow-based network traffic classification system," Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 368–371, 2015.

[51] N. Goernitz, M. M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection," J. Artif. Intell. Res., vol. 46, pp. 235–262, 2014.

[52] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," IEEE Trans. Knowl. Data Eng., vol. 17, no. 11, pp. 1529–1541, Nov. 2005.