EPiC Series in Computing

Volume 89, 2022, Pages 1–10

Proceedings of 35th International Conference on Computer Applications in Industry and Engineering

A Bucket-Based Data Pre-Processing Method for Encrypted Video Detection*

Waleed Afandi[1]†, Syed M. A. H. Bukhari[1]‡, Muhammad U. S. Khan[1]§, Tahir Maqsood[1]¶, and Samee U. Khan[2]‖

[1] Department of Computer Science, COMSATS University Islamabad, Abbottabad, Pakistan
waleedafandi,ammar,ushahid,tmaqsood@cuiatd.edu.pk

[2] Electrical and Computer Engineering, Mississippi State University, USA
skhan@ece.msstate.edu

Abstract

As the number of video streaming platforms is growing, the risk factor associated with illegal and inappropriate content streaming is increasing exponentially. Therefore, monitoring such content is essential. Many researches have been conducted on classifying encrypted videos. However, most existing techniques only pass raw traffic data into classification models, which is an ineffective way of training a model. This research proposes a bucket-based data pre-processing technique for a video identification in network traffic. The bucketed traffic is then incorporated with a fine-tuned word2vec-based neural network to produce an effective encrypted video classifier. Experiments are carried out with different numbers and sizes of buckets to determine the best configuration. Furthermore, previous research has overlooked the phenomenon of concept drift, which reduces the effectiveness of a model. This paper also compares the severity of concept drift on the proposed and previous technique. The results indicate that the model can predict new samples of videos with an overall accuracy of 81% even after 20 days of training.

1 Introduction

With the rapid increase in internet users and ease of accessibility of mobile devices, the mobile data traffic has increased exponentially. According to the CISCO Visual Networking Index (VNI) report, by 2023, 71% of the world's population will be mobile users and the mobile connections per capita will increase 33% [1]. Moreover, this huge increase also result in an increase of video traffic on the internet. Concurrently, the identification of the video in the

*This material is based on work supported by National Center of Cyber Security (NCCS), Pakistan and the Higher Education Commission (HEC) under grant RF-NCCS-023.

†Designed methodology and performed tests

‡Performed tests

§Designed methodology and administered the project

¶Supervised the project

‖Supervised the project

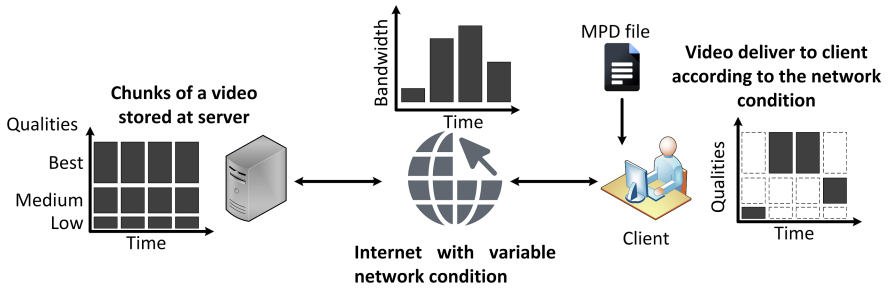

Figure 1: Dynamic Adaptive Streaming over HTTP overview

internet traffic to track down the criminal activities also plays an important role. Similarly, intelligence agencies also want to track down militants by analyzing their video-watch history.

Video streaming websites, such as YouTube and Dailymotion, have adopted Dynamic Adaptive Streaming over HTTP (DASH). DASH is a streaming technology widely used by video streaming service providers to improve user experience (QoE). In DASH, each video is sliced into 6 to 10 seconds chunks, sometimes called segments. These segments are then delivered to the client based on the network conditions for smooth streaming. The information about the segments is present in a Media Presentation Description (MPD) file. When a client requests for a video, the MPD file is first downloaded to the client's device, which contains all the links of the video segments. The video player requests the video segments according to the client's network condition. A generic overview of DASH streaming is presented in Figure 1.

In DASH, each quality segment is encoded with Variable Bitrate (VBR) encoding. The VBR is generally used in audio and video compression to achieve improved video quality. The VBR creates abnormalities in the network, making it difficult for researchers to identify the video in the Internet traffic. However, to handle these abnormalities, a stable fingerprint method [2] can be utilized to minimize the VBR effect. Furthermore, a Markov probability fingerprint [3] can also be used to identify the VBR encoded video in the internet traffic.

In recent years, the Convolutional Neural Network (CNN) has shown an advantage over traditional machine learning algorithms. Instead of providing the hand-crafted features, the CNN automatically extract them from the input data. It is actively used in many machine learning domains, such as Pattern Recognition, Speech Recognition, and Image Classification. Many researchers have also utilized the CNN in the video identification in the internet traffic. The CNN is a probabilistic classifier that not only predicts the class label, but also calculates the probability of the classes.

In a real-time prediction problem, the data on which a pre-trained model tries to predict the output variable change over time. This phenomenon is known as *Concept Drift* in machine learning, predictive analysis, and data mining [4, 5]. In video identification, the pattern of transmission of the same video differs due to the adaptation of DASH and the client's network conditions. The same video, if played at different times on the client's system, shows a different pattern of the data received, as shown in Figure 2. This contingency affects the prediction accuracy of the trained model, affecting the overall efficiency of the model performance.

In this paper, we present a video identification method using the bin technique presented in [3]. We extracted the bytes per seconds (BPS) from the captured video traffic and allocate the BPS to the buckets according to the equation to reduce the effect of inconsistency. Further details are discussed in Section 3. The main contributions are as follows:

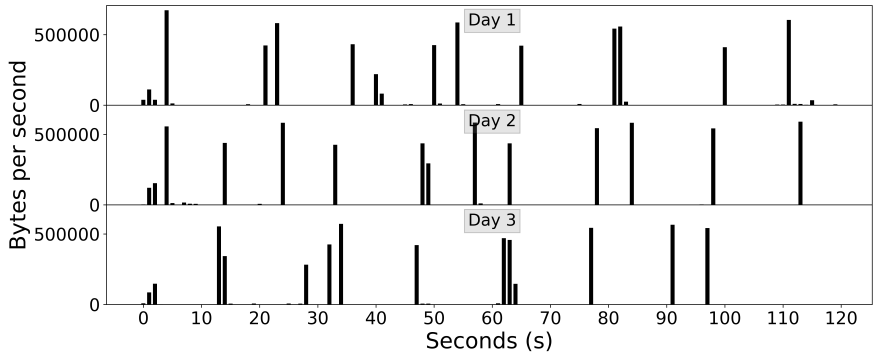

Figure 2: Streaming pattern of a video when capture on three different days

- A bucket-based fingerprint method for video identification in internet traffic
- We analyze the effect of size and number of buckets on the accuracy and conclude that 150 buckets, each of 5000 bytes are optimal for YouTube video identification.
- We perform various experiments on datasets of different days to show the effectiveness of our technique and observed the proposed technique to retain prediction accuracy of more than 85% even after 20 days of model training.
- We utilize the proposed technique to handle the concept drift phenomenon and demonstrate an overall difference of 20% model accuracy when compared to other techniques.

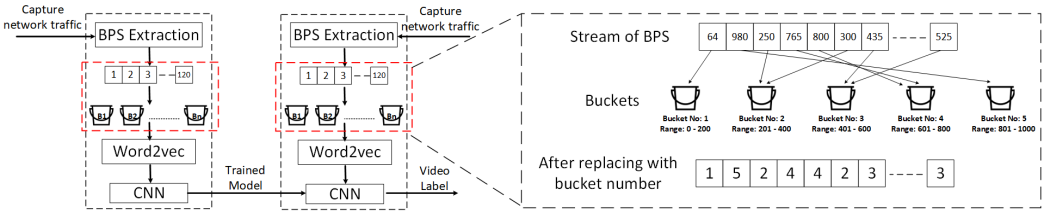

Figure 3: Video Identification steps using buckets

The rest of the paper is organized as follows. The literature review is presented in Section 2. The background and structure about the video identification using buckets are presented in Section 3. Section 4 presents the steps involved in dataset preparation and details about the CNN, and Section 5 detailed the performance evaluation of our proposed framework. Section 6 concludes the paper.

2 Related Work

Throughout the course of modern computing, encrypting and decrypting data is one of the major challenges in computer science. Encryption is a fundamental requirement to ensure user's privacy and security. After the widespread adoption of Hypertext Transfer Protocol Secure

(HTTPS) due to its secure nature, many researchers started to contribute to decipher encrypted traffic. Shen *et al.* [6] propose a method using second-order Markov chains to classify encrypted traffic. Fu *et al.* [7] use different characteristics of traffic data, such as packet length and time delay, and passed them to a Hidden Markov Model (HMM) for the classification of encrypted Internet traffic. Likewise, Liu *et al.* [8] use Multi-attribute Markov Probability fingerprints to classify encrypted traffic. Chen *et al.* [9] show that side-channel attacks are possible with modern encryption techniques. Even hardware-oriented attacks are viable, as demonstrated by Han *et al.* [10]. User activities are suspected to be leaked [11–14]. If the adversary is present inside LAN, she can sniff packets and gain vulnerable information [15]. Similarly, sniffing of Wi-Fi signals can also put user privacy at risk [12]. Location-based applications are also susceptible to leaking users' private data [16–19].

Video traffic data however, is much different from website and application traffic data. While website and traffic remains much the same whereas video traffic may differ depending upon the network condition and the suitable video quality for the network. As mentioned before, DASH is a popular video streaming protocol responsible for optimizing video quality according to the network conditions. Khan *et al.* [20, 21] have demonstrated the use of bytes per second (BPS) as a feature to identify video labels through convolutional neural networks. However, their technique does not take variable bit rate of videos into consideration. Meanwhile, Gu *et al.* [2] create segments of received bytes to account for the inconsistent variable bit rate. Yang *et al.* [3] use Markov chains to detect encrypted videos' titles. Dubin *et al.* [22] use KNN and SVM for adaptive video streaming title classification. According to Schuster *et al.* [23], most videos can be identified by their unique burst patterns. Ameigeiras *et al.* and Ravattu *et al.* [24, 25] also leveraged these burst patterns to identify video titles. The On-Off period in-between two consecutive bursts is also used as an effective feature to identify videos in [26, 27]. In addition to identifying video titles, Dubin *et al.* [28] also contributed in detecting the video quality using machine learning.

While considering the DASH protocol, this research aims to handle VBR based streaming through the concept of data bucketing. The optimal values for the total number and size of buckets are identified through rigorous testing. Furthermore, the a vocabulary of the bucketed values is formed using *word2vec* and the generated vocabulary is passed to a CNN explicitly designed for the bucketed network data. To our knowledge, this paper is the first to incorporate data bucket methodology with word2vec in-order to create an encrypted video title predictor. Furthermore, the previous works do not address the phenomenon of *Concept Drift*, in which old training data become ineffective and require retraining with the most recent generated dataset. This paper aims to investigate the severity of a concept drift by training the model over a selected range of days, and its prediction accuracy is analyzed on a future-generated dataset. Moreover, this also provides information on future proofing and long-term effectiveness of the model.

3 Video Identification using Buckets

This section presents the framework that can be used to identify videos on the Internet, as shown in Figure 3. In the proposed framework, the traffic is captured and the BPS streams from the traffic are extracted. The individual bytes from the BPS stream are allocated to different bucket numbers, depending upon the size of buckets and total number of buckets. We perform various experiments to choose the suitable number and bucket and their sizes. The following equation generates i, which is the bucket number allocated to each bytes in the stream:

Table 1: Hyper-parameter summary

	Parameter	**Value**
Embedding	Embedding Dimension	256
Conv1D (1st Layer)	Filters	512
	Kernel Size	5
	Activation fuction	tanh
Conv1D (2nd Layer)	Filters	256
	Kernel Size	4
	Activation fuction	tanh
Conv1D (3rd Layer)	Filters	128
	Kernel Size	3
	Activation fuction	tanh
Max Pooling 1D 1st, 2nd, and 3rd Layer	Pool size	4, 3, 2
Droupout	Rate	0.4
Dense Layer	Activation fuction	Relu and Softmax
Model	Optimizer	Adam
	Batch size	10
	Epochs	100

$$i = \begin{cases} \lfloor \frac{bytes}{size} \rfloor, & 0 \leq bytes < size \times total\ buckets \\ \\ total\ buckets - 1, & bytes \geq size \times total\ buckets \end{cases} \tag{1}$$

After allocating each byte a bucket number, all the bytes in the BPS stream are replaced with their respective bucket number. This bucket stream is then converted into a vector using the word2vec technique. The word2vec technique is used to combine the words with their related occurrences. It converts the text into numerical format that is understandable to neural networks. The output of word2vec is a vocabulary from which neural network can detect the similarity of the words based on the context. It combines the vector of similar occurring words in a vector space. This technique is effectively being used in sentiment analysis, natural language processing, and recommendation systems.

The vectors generated from the word2vec technique are then used as input to train the Convolutional Neural Network (CNN). The details of CNN is provided in Section 4. The trained model is then used for video identification in network traffic. For video identification, the same methodology is followed; i.e., the BPS are extracted from a captured stream and assigned to the buckets, then the BPS stream is replaced by the number of buckets. This stream of bucket number is converted to vectors and passed to the model for video prediction.

4 Experimental Setup

We performed experiments on a Dell Precision 5820 Tower Workstation, equipped with an Intel Xeon W-2223 CPU @ 3.60GHz with 8 cores. The system has a RAM of 32 GB with a storage capacity of 4.5 TB. The system has NVIDIA RTX 3060Ti with 4864 CUDA cores and the Ubuntu 20.04 LTS operating system.

4.1 Dataset Preparation

In this work, we prepared two datasets. For this purpose, we set up a dummy client using selenium, mimicking the behavior of a real user playing YouTube videos. We randomly select 43 videos from different YouTube channels. The video quality is set to *Auto* mode to simulate a real world streaming scenario whereas in previous research by Khan *et al.* [20,21], the quality is fixed i.e., 360p. For traffic capturing, we use *Tshark*, a command-line interface of Wireshark, and saved the captured traffic in Packet Capture (PCAP) files.

We capture the dataset at two different times. At first, we captured the dataset for a month containing 155 streams of each video, making a total of 6665 streams. We call this a month-wise dataset and trained the model on this dataset. Also, we captured five streams of each video for 20 days for testing purposes and call it a day-wise dataset.

4.2 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is used to train on the stream of buckets for video identification. We fine-tuned the CNN through several experiments by varying the different hyper-parameters. The CNN model comprises of one embedding layer and three 1-dimensional convolutional layers, separated by three 1-dimensional Max-Pooling layers. The first convolutional layers have 512 filters with kernel size equal to 5. The second and third convolutional layers have a number of filters equal to 256 and 128 whereas kernel sizes in these layers are set to 4 and 3, respectively. All convolutional layers have the ***tanh*** activation function. After convolutional and max-pooling layers, there is a drop-out layer with a rate of 0.4, followed by a flatten layer. The model contains a dense layer with 64 neurons and the activation function ***ReLU***. The output layer has the activation function softmax, and the number of neurons equals the number of classes. Table 1 presents the summary of the final CNN used in this study during the experiments.

5 Performance Evaluation

In this section, we present the video identification results of our proposed technique. Initially, we perform various tests to identify the most suitable combination of the number and size of buckets. After the selection, we analyze the effectiveness of our technique. For this purpose, we check our model's accuracy on the day-wise dataset to see whether our model can handle the network abnormalities during the video identification process. Moreover, we compare our results of the stream of buckets with BPS to show the performance of our proposed technique. The performance measure used in this study is accuracy which is defined as

$$Accuracy = \frac{Number\,of\,correct\,predictions}{Total\,number\,of\,predictions} \tag{2}$$

5.1 Size and Number of Buckets

The bucket's number and size of the buckets depend directly on the available resources. The higher number of buckets and the larger size of the buckets require more extensive resources for processing. For this purpose, we perform experiments with different combinations of number and size of buckets. We select 100, 150, and 200 buckets for the number of buckets. We tried each number with 2500, 5000, and 7500 size. For example, in the scenario of 100 buckets with 2500 bucket size, there are total 100 buckets where the size of each bucket is 2500 bytes. The

difference between consecutive buckets is 2500. For example, Bucket1 holds the BPS from 0 to 2500 bytes. Similarly, Bucket2 holds BPS from 2501 to 5000.

In the first experiment, we select 100 buckets and perform experiments with different bucket sizes. The results show that the accuracy is highest when the bucket size is 7500 bytes. The results of 100 buckets with different bucket sizes are summarized in Figure 4. Similarly, we have carried out experiments with 150 and 200 buckets. The results show that the buckets of 5000 bytes outperform the other bucket sizes when the number of buckets is 150 and 200. Although 150 and 200 buckets show similar results in accuracy when the size of the buckets is 5000, we select 150 buckets as higher number of buckets require more processing resources. The Figure 4 summarizes the results of a different combination of number and size of buckets.

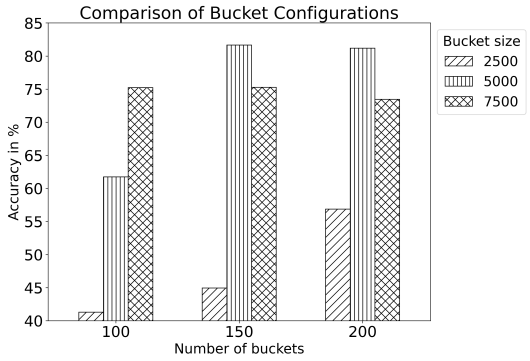

Figure 4: Comparison of different combination of size and number of buckets

5.2 Concept Drift Handling

After selecting the suitable combination of the number and size of buckets, we analyze how well our proposed technique handle the concept drift. For this purpose, we train the model on the month-wise dataset and test its accuracy on the day-wise dataset. We compare our proposed technique with the BPS technique presented in [20,21]. For the proposed technique, we prepare the dataset using the bucket procedure presented in the paper. However, for BPS, only bytes from the PCAP files are extracted and directly passed to the classifying model. The result shows that our proposed technique not only outperforms the BPS model but also efficiently handles the concept drift by showing uniformity in accuracy. The proposed technique performs 20% better than its counterpart in the prediction of video labels. Figure 5 shows the results of the day-wise accuracy test of our technique compared to BPS.

5.3 Model Comparison

This section compares the proposed model with the BPS model. The dataset prepared in this section follows the bucket technique presented in this paper. The extracted bytes of the captured video streams are assigned a bucket number for both BPS and Bucket model. These bucket streams are then converted into vectors using the word2vec embedding technique, and the model is trained on the input data. We follow the same procedure for training and testing purposes as discussed in the previous subsection, however for BPS model, we used the hyper-parameter settings presented in [20] whereas for Bucket model we used the hyper-parameter settings presented in Table 1. The results show that initially the Bucket model depicted accuracy

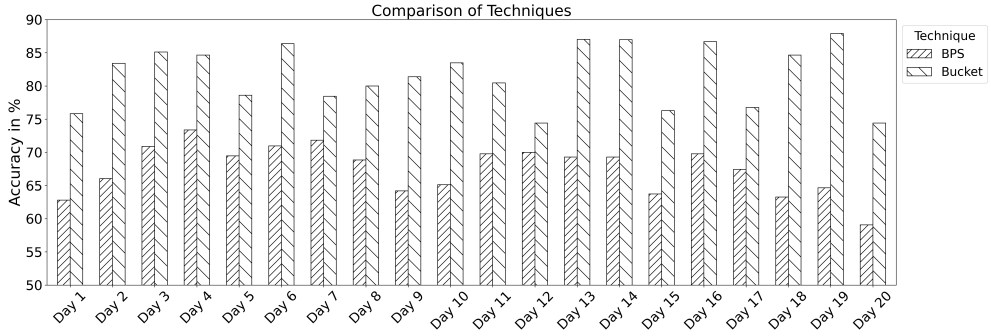

Figure 5: Concept drift handling

results similar to those of the BPS model. Moreover, on Day-9 and Day-10 datasets, the BPS model shows slightly better results than the Bucket model. However, on Day-11 and onwards datasets, the Bucket model outperforms the BPS model. The average accuracy of the BPS and Bucket model is 78% and 81%, respectively. The maximum accuracy achieved by Bucket model is 86.3%. The day-wise results of both of the models are presented in the Figure 6.

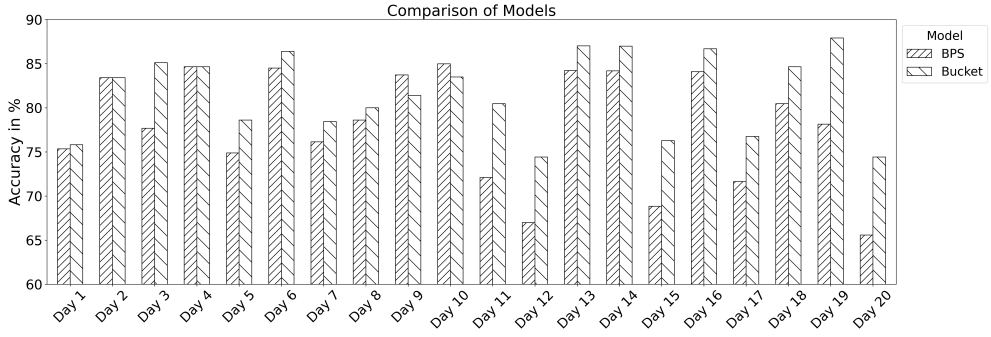

Figure 6: Comparison of different models

6 Conclusions

Research on the identification of encrypted video streams is of the utmost importance to monitor inappropriate and illegal videos. This paper presents a technique for identifying encrypted videos by pre-processing the traffic data through bucket technique. Buckets help to generalize the traffic data which is useful in minimizing the affect of network fluctuations. First the bytes per second are extracted from the generated PCAP file while the headers are removed from the packets. These BPS are passed through the bucket process, and the resultant data is a bucketed BPS. A bucketed BPS vocabulary is formed using word2vec and passed onto a CNN. As the proposed technique currently works for sniffing the network traffic upto 120 seconds, for future work, we aim to reduce the this length to 60 seconds.

References

[1] U Cisco. Cisco annual internet report (2018–2023) white paper. *Cisco: San Jose, CA, USA*, 2020.

[2] Jiaxi Gu, Jiliang Wang, Zhiwen Yu, and Kele Shen. Walls have ears: Traffic-based side-channel attack in video streaming. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1538–1546. IEEE, 2018.

[3] Luming Yang, Shaojing Fu, Yuchuan Luo, and Jiangyong Shi. Markov probability fingerprints: A method for identifying encrypted video traffic. In *2020 16th International Conference on Mobility, Sensing and Networking (MSN)*, pages 283–290. IEEE, 2020.

[4] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[5] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

[6] Meng Shen, Mingwei Wei, Liehuang Zhu, and Mingzhong Wang. Classification of encrypted traffic with second-order markov chains and application attribute bigrams. *IEEE Transactions on Information Forensics and Security*, 12(8):1830–1843, 2017.

[7] Yanjie Fu, Hui Xiong, Xinjiang Lu, Jin Yang, and Can Chen. Service usage classification with encrypted internet traffic in mobile messaging apps. *IEEE Transactions on Mobile Computing*, 15(11):2851–2864, 2016.

[8] Chang Liu, Zigang Cao, Gang Xiong, Gaopeng Gou, Siu-Ming Yiu, and Longtao He. Mampf: Encrypted traffic classification based on multi-attribute markov probability fingerprints. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2018.

[9] Shuo Chen, Rui Wang, XiaoFeng Wang, and Kehuan Zhang. Side-channel leaks in web applications: A reality today, a challenge tomorrow. In *2010 IEEE Symposium on Security and Privacy*, pages 191–206. IEEE, 2010.

[10] Jinsong Han, Chen Qian, Panlong Yang, Dan Ma, Zhiping Jiang, Wei Xi, and Jizhong Zhao. Geneprint: Generic and accurate physical-layer identification for uhf rfid tags. *IEEE/ACM Transactions on Networking*, 24(2):846–858, 2015.

[11] Muhammad US Khan, Assad Abbas, Mazhar Ali, Muhammad Jawad, and Samee U Khan. Convolutional neural networks as means to identify apposite sensor combination for human activity recognition. In *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 45–50. IEEE, 2018.

[12] Fan Zhang, Wenbo He, Xue Liu, and Patrick G Bridges. Inferring users' online activities through traffic analysis. In *Proceedings of the fourth ACM conference on Wireless network security*, pages 59–70, 2011.

[13] Mauro Conti, Luigi Vincenzo Mancini, Riccardo Spolaor, and Nino Vincenzo Verde. Analyzing android encrypted network traffic to identify user actions. *IEEE Transactions on Information Forensics and Security*, 11(1):114–125, 2015.

[14] Rizwana Irfan, Osman Khalid, Muhammad Usman Shahid Khan, Faisal Rehman, Atta Ur Rehman Khan, and Raheel Nawaz. Socialrec: A context-aware recommendation framework with explicit sentiment analysis. *IEEE Access*, 7:116295–116308, 2019.

[15] Mauro Conti, Nicola Dragoni, and Viktor Lesyk. A survey of man in the middle attacks. *IEEE Communications Surveys & Tutorials*, 18(3):2027–2051, 2016.

[16] Zimu Zhou, Zheng Yang, Chenshu Wu, Wei Sun, and Yunhao Liu. Lifi: Line-of-sight identification with wifi. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2688–2696. IEEE, 2014.

[17] Xi Chen, Xiaopei Wu, Xiang-Yang Li, Xiaoyu Ji, Yuan He, and Yunhao Liu. Privacy-aware high-quality map generation with participatory sensing. *IEEE Transactions on Mobile Computing*, 15(3):719–732, 2015.

[18] Yi Guo, Lei Yang, Bowen Li, Tianci Liu, and Yunhao Liu. Rollcaller: User-friendly indoor navigation system using human-item spatial relation. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2840–2848. IEEE, 2014.

[19] Qiang Ma, Shanfeng Zhang, Tong Zhu, Kebin Liu, Lan Zhang, Wenbo He, and Yunhao Liu. Plp: Protecting location privacy against correlation analyze attack in crowdsensing. *IEEE transactions on mobile computing*, 16(9):2588–2598, 2016.

[20] Muhammad US Khan, Syed MAH Bukhari, Tahir Maqsood, Muhammad AB Fayyaz, Darren Dancey, and Raheel Nawaz. Scnn-attack: A side-channel attack to identify youtube videos in a vpn and non-vpn network traffic. *Electronics*, 11(3):350, 2022.

[21] Muhammad US Khan, Syed MAH Bukhari, Shazir A Khan, and Tahir Maqsood. Isp can identify youtube videos that you just watched. In *2021 International Conference on Frontiers of Information Technology (FIT)*, pages 1–6. IEEE, 2021.

[22] Ran Dubin, Amit Dvir, Ofir Pele, and Ofer Hadar. I know what you saw last minute—encrypted http adaptive video streaming title classification. *IEEE transactions on information forensics and security*, 12(12):3039–3049, 2017.

[23] Roei Schuster, Vitaly Shmatikov, and Eran Tromer. Beauty and the burst: Remote identification of encrypted video streams. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1357–1374, 2017.

[24] Pablo Ameigeiras, Juan J Ramos-Munoz, Jorge Navarro-Ortiz, and Juan M Lopez-Soler. Analysis and modelling of youtube traffic. *Transactions on Emerging Telecommunications Technologies*, 23(4):360–377, 2012.

[25] Radha Ravattu and Prudhviraj Balasetty. Characterization of youtube video streaming traffic, 2013.

[26] Ashwin Rao, Arnaud Legout, Yeon-sup Lim, Don Towsley, Chadi Barakat, and Walid Dabbous. Network characteristics of video streaming traffic. In *Proceedings of the seventh conference on emerging networking experiments and technologies*, pages 1–12, 2011.

[27] Youting Liu, Shu Li, Chengwei Zhang, Chao Zheng, Yong Sun, and Qingyun Liu. Doom: a training-free, real-time video flow identification method for encrypted traffic. In *2020 27th International Conference on Telecommunications (ICT)*, pages 1–5. IEEE, 2020.

[28] Ran Dubin, Ofer Hadar, Amit Dvir, and Ofir Pele. Video quality representation classification of encrypted http adaptive video streaming. *KSII Transactions on Internet and Information Systems (TIIS)*, 12(8):3804–3819, 2018.