# Classification of Applications in HTTP Tunnels

Gajen Piraisoody[1], Changcheng Huang[1], Biswajit Nandy[2], Nabil Seddigh[2]

[1]*Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Canada*
[1]*GajenPiraisoody@cmail.carleton.ca, huang@scel.carleton.ca*

*Solana Networks, Ottawa, Ontario, Canada*
[2] {bnandy, nseddigh}@solananetworks.com

*Abstract*— **Accurate traffic classification is an essential element of emergent cloud and datacenter architectures. Increasingly, however, different types of application traffic from the cloud are tunnelled over HTTP, thereby making accurate classification a challenge. Applications tunnelled over HTTP are wide in scope and diverse in nature, and include mapping, email, video, image, audio and file. This paper presents a novel approach for the accurate and effective classification of the dominant types of HTTP tunnelled applications, namely video, audio and file-transfer. The classification is carried out using information that is only available from flow-based protocols such as NetFlow v5. The proposed scheme is tested on live data traffic in a small enterprise network with a realistic mixture of regular HTTP and non-HTTP traffic. Outsourcing enterprise networks to cloud is a major cloud application. For the scenarios tested, the proposed algorithm accurately classifies at least 70% of the HTTP tunnelled traffic, and in some cases, up to 90%. In comparison to the results from approaches based on NaiveBayes algorithm and Support-Vector-Machine, the proposed scheme outperforms them by at least 10% as per performance measures.**

***Keywords***— **Traffic classification, cloud computing, NetFlow**

## I. INTRODUCTION

Service providers such as Microsoft, Amazon and Google are progressively building larger cloud datacenters to support cloud-based services like email, gaming, video, file-transfer and audio. With the emergence of cloud networking and services provided by Content-Delivery-Networks (CDN), HTTP tunnelled traffic has increased in size and usage. The rise in HTTP tunnelled traffic presents a challenge for accurate traffic classification – a requisite for monitoring traffic, enforcing policies, ensuring a high quality of service and establishing a secured service in a cloud network environment. In the lack of an effective classification process, HTTP traffic has become increasingly cluttered, rendering it difficult to maintain control over the different types of traffic.

This paper proposes an approach to identify the dominant applications tunnelled over HTTP, namely audio, video and file-transfer. In this approach, an effective classification scheme is presented using information that is available from flow-based protocols such as NetFlow v5. The scheme relies on a two-step process to classify tunnelled HTTP traffic. Firstly, it groups tunnelled applications based on traffic characteristics such as download rate, bytes-per-packet and bytes-per-flow. Secondly, concepts of 'occupancy' and 'flow-group' are applied to further classify the traffic more accurately.

This paper consists of five sections. Section II briefly describes the background and previous work relevant to the present study. Section III details the proposed technique and the related thresholds. Section IV discusses the experimental results and, lastly, Section V presents the conclusion.

## II. BACKGROUND AND RELATED WORK

### A. Background

Fig. 1 illustrates one taxonomy representing different approaches to network traffic classification. The four generalized traffic classification approaches are port-based, flow-based, payload-based and behaviour-based. These four general approaches are sometimes combined with additional techniques including heuristics and statistical analysis.

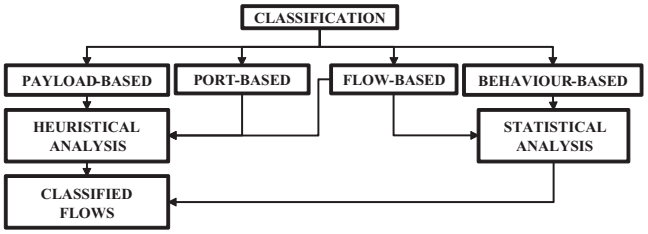



Fig. 1 Network Traffic Classification

Port-based classification identifies applications based on IANA (Internet-Assigned-Numbers-Authority) assigned port mappings. Port-based classification, however, is rather inaccurate for cloud-based services due to the prevalence of HTTP tunnelling and dynamic port assignment.

Flow-based classification utilizes network traffic statistics to identify applications. Typical flow classification approaches are based on a variety of flow statistics referred to as flow features. Of notable importance when classifying a flow are traffic statistics such as the average packet size, number of packets, start time, duration, protocol, protocol flags, port pairs and IP pairs. The drawback of the flow-based classification approach is that traffic statistics for a single flow may not be sufficient to clearly identify an application.

Payload-based classification is also referred to as deep-packet-inspection (DPI) . DPI methods analyze individual packet payloads to detect applications based on known patterns matched against a database. DPI approaches are attractive due to their accuracy. The drawback with DPI is that it requires faster CPUs or specialized hardware. In addition, when traffic is encrypted, signature-based techniques such as DPI are rendered inoperable.

978-1-4799-0568-3/13/$31.00 ©2013 IEEE

Behaviour-based traffic classification works by examining patterns based on social, functional and application characteristics. Host behavioural analysis, however, can fail to determine applications from single or few flows and is limited by its inability to classify tunnelled traffic.

As Szabo *et al.* point out, any single classification approach is prone to less accurate categorization. Thus, combining the aforementioned classification types with a heuristic-based or statistical based approach may yield higher accuracy. Variables like rate, packet size, flow-group, data size and others play important roles in heuristically analyzing traffic. The statistical approach is more inclined with the use of Machine Learning models.

This paper reports upon a new form of traffic classification that combines port-based, flow-based and behaviour-based techniques.

### B. Related work

This sub-section discusses prior work in the area of HTTP traffic classification.

In identifying twenty different applications tunneled over HTTP, Brice *et al.* employed a combination of behavior (pattern) based classification and packet-based classification. The study was based on offline Tcpdump trace-files. The paper concludes by introducing a preliminary draft of an approach to classifying HTTP traffic. The most important considerations the paper suggests include the Dirac shape of music traffic, the continuous shape of radio traffic and the depiction of video/file-transfer traffic at high intensity levels. The literature also identifies other types of HTTP tunneled traffic such as mail, chat, maps and images.

Kei *et al.* introduce the concept of flow-group. A flow-group consists of a set of flows that occur within a few seconds and have the same IP destination address. HTTP video flow-groups are very large, ranging to include more than twenty flows for a video flow-group. Flows in a flow-group can be involved in various activities through a video player, style-sheets, advertisements and activities involving content server communications. The flow-groups of HTTP file transfers are estimated to be small and, in most cases, to be only one owing to the fact that a typical file transfer involves a direct download from a server.

Samruay *et al.* introduce a combination of DPI and statistical behaviour analysis based on the attributes of video and audio traffic. The authors propose the use of a statistical approach on flow traffic while conducting DPI on a few packets of aggregated flows. The heuristics derived in Samruay *et al.* were valuable in validating the heuristics used in the classification design of this paper.

Cheng *et al.* 's YouTube measurement study presents an in-depth look at the unique characteristics of Youtube.. Detailed information, such as the distribution of video length, distribution of video file size and video rates are observed and presented. The study considers the special characteristics of YouTube against that of traditional video. Important features such as meta-data block sending and traffic statistics for short video are also researched and discussed.

Skype traffic is an interactive form of traffic and does not conform to the multimedia streaming traffic types that are within the scope of this paper. Maciej et al. analyze the attributes of encrypted Skype traffic to carry out accurate classification. As per Maciej, distinguishing between Skype voice and Skype video contents is complex and challenging. Although this paper does not include Skype traffic in its scope, it does provide insight and an approach to classifying video and voice traffic which can be further modified and extended to Skype traffic

The approach presented in this paper differs from those of previous studies in the following ways:

- An enhanced version of the flow-group concept is introduced. The process specifically attempts to identify CDN servers and includes flow-groups based on CDN.
- The proposed scheme does not require a training dataset.
- The proposed scheme relies on information solely available in flow-based protocols such as NetFlow-V5.
- The proposed scheme has the ability to classify encrypted data.

## III. PROPOSED TECHNIQUE

Previous works in the traffic classification domain have utilized the concept of flow-groups to identify applications. The main contribution of this paper is to enhance existing flow-group concepts and augment it with new techniques which consider "occupancy" in order to classify HTTP mention tunnelled traffic such as audio, video and file-transfer. The subsections below discuss the flow definition used to analyze network traffic, illustrate the methodology used to classify HTTP tunnelled traffic, introduce a new attribute referred to as 'occupancy', explain the concept of an enhanced flow-group method and outline the statistical properties of audio, video and file-transfer traffic. Finally, algorithms for classifying audio and video/file-transfer are presented in this section.

### A. Flow Definition

Network 'flow' refers to the aggregation of information from different packets of a given protocol to a flow defined by n-tuple within a time period. The HTTP tunnelled classification algorithm proposed in this paper uses NetFlow-V5 developed by Cisco . A unique flow in NetFlow-V5 is defined in terms of a seven-tuple consisting of the same source-IP, destination-IP, source-port, destination-port, type of service, interface and protocol (TCP/UDP). In addition to the attributes of a seven-tuple, the NetFlow configuration of flow-idle time and flow-period also delineate a flow. In this study, the flow-idle time and maximum flow-period are set to 3 and 120 seconds respectively. Our classification approach is based on the information provided by NetFlow. No other information is required. This makes our approach readily applicable to all existing routers that support NetFlow or a similar flow-based protocol.

### B. Methodology Used to Classify HTTP Tunnelled Traffic

As shown in Fig. 2, the basic approach in our proposed classification algorithm is to firstly identify long duration HTTP flows because video, audio and file-transfer are long-flow by nature. Long-flow traffic can be identified by the large bytes-per-flow of a flow. Secondly, potential audio traffic, including both radio and music, is identified based on parameters like download rates, bytes-per-flow size, bytes-per-packet and 'occupancy'. Thirdly, potential video and file-transfer traffic is identified based on parameters like download rates, bytes-per-flow size, bytes-per-packet and flow-group. The order presented in Fig. 2 will provide optimal results for the proposed scheme.

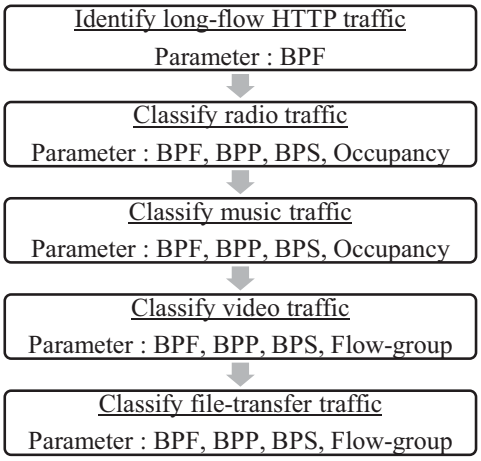



Fig. 2 Methodology of HTTP tunnelled traffic classification using occupancy, flow-group, bytes-per-packet (BPP), bytes-per-flow (BPF), and bytes-per-second (BPS)

### C. Occupancy

This section describes the notion of the occupancy attribute as a key factor in the classification algorithm. The definition of occupancy was formulated after observing the different traffic patterns of audio and video applications, coupled with our understanding of client-server interaction for HTTP audio and HTTP video.

Occupancy is a ratio of the flow duration (defined by the Netflow record for the multimedia phase) over the entire duration of a chunk of time. A multimedia phase includes a set of one or more related flows between the same multimedia client and the HTTP server. A discontinuity in interaction for a specific duration of time in multimedia is indicative of two separate multimedia phases. In order to calculate occupancy, time is divided into chunks. Intuitively, if a multimedia flow or group of related multimedia flows transfer data continuously over the entire chunk of defined time, the behaviour is termed as 'high occupancy'. If the flow (or group of related flows) transfers all of its data in a short period of the time chunk, the flow, or the group of flows, is termed 'low occupancy'.

In this paper, multimedia refers to streamed HTTP video and audio. Online radio and online music are two distinct forms of streamed HTTP audio. This paper excludes from its scope the interactive form of voice/video transmission, such as Skype.

Fig. 3 illustrates the behaviour of HTTP streamed music, radio and video by using data from three popular multimedia sites. Music applications exhibit a spike-like 'Dirac' pattern with medium download rates. Fig. 3 illustrates the two different phases in data transmission when music traffic is transmitted from Grooveshark. The two separate phases indicate two separate events of music download by the client from Grooveshark. By contrast, radio applications exhibit a continuous pattern as illustrated by the traffic pattern produced from Hdradio. Lastly, video content from CTV exhibit shaped or variable patterns consisting of four flows with distinctively large flow-size .

The occupancy metrics of the content download are indicative of music, radio and video traffic patterns. A high occupancy value indicates that the multimedia phase exhibits a continuous pattern (i.e. radio) and a low occupancy value indicates a Dirac pattern (i.e. music).

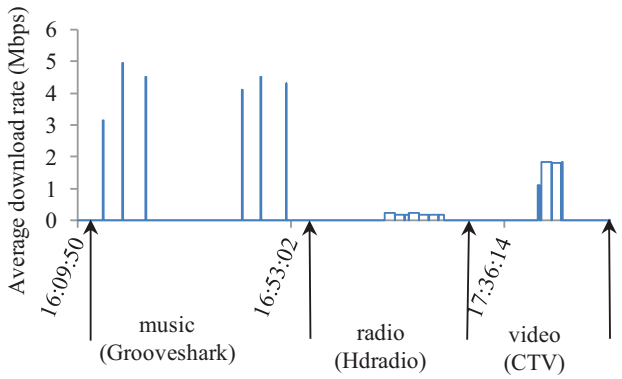



Fig. 3 Multimedia Traffic Patterns

### D. Flow-group

Kei *et al.* introduced the concept of flow-group. In , a flow-group is defined as the set of flows that occur within a few seconds and have the same destination-IP address. The article suggests first picking a potential video or file-transfer flow based on its large flow size and then inspecting other flows within one round-trip-time (couple of seconds) of the video/file-transfer flow. File-transfer flows mostly originate from one server. Video flows, by contrast, utilize Content Delivery Network (CDN) technology, and are consequently subjected to traffic from multiple servers and therefore accumulating large flow-groups. The ensuing multiple flows are not only video-flows from multiple CDN servers, but also include meta-data, advertisements, style-sheet, video-player and other such video-player-related contents from multiple CDN servers.

In this paper, the definition of flow-group formulated by Kei et al. is expanded to include flows from different CDN nodes and flows greater than one round-trip-time. The different CDN nodes are aggregated to the same source server

by inspecting IP addresses in the same subnet of the source-IP. Thus, the expanded flow-group also includes flows that occur within an accepted duration that have the same subnet masked source-IP address and the same destination-IP address.

### *E. Statistical Properties of Audio, Video & File-transfer*

A detailed study of different kinds of traffic generated by a variety of applications indicates that the majority of potential audio, video and file-transfer traffic can be grouped in terms of the parameters as shown in Fig. **4**. The Z-axis represents multimedia player rates.

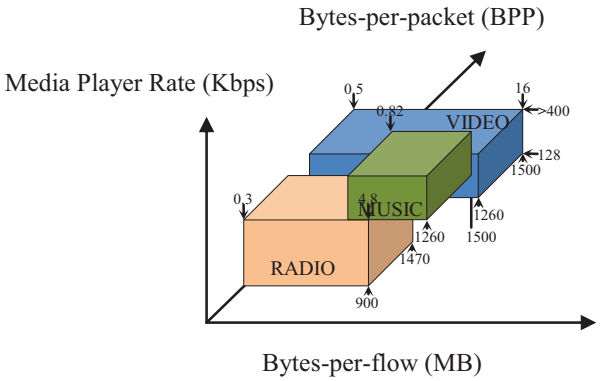



Fig. 4 Potential radio, music, video and file-transfer boundaries

Based on 3683 online radio rates offered, 99.4% of the radio player's rates are between 20Kbps and 320Kbps. According to 52 popular online music sites (like Grooveshark, Soundcloud and others), 98% of the music player rates are within 64Kbps and 320Kbps. Cheng *et al.*'s comprehensive study of video characteristics are used in this paper. The bytes-per-packet thresholds in the y-axis are obtained from studies conducted by Samruay *et al.* . The bytes-per-flow thresholds in the x-axis are garnered through multiple factors such as song length, flow-period configuration of NetFlow-V5. Multimedia player rates are presented on the z-axis.

TABLE 1 shows the confidence interval of music occupancy and radio occupancy. The occupancy values are estimated from the three datasets described in TABLE 3. These occupancy values are used as a primary factor in characterizing audio traffic.

TABLE 1. OCCUPANCY

| | Average Value % | 95 % Confidence Interval |
|---|---|---|
| Music | 19 | 0,55 |
| Radio | 92 | 82,100 |

### *F. Proposed Audio Algorithm*

The identification of audio traffic is accomplished in two steps. In the first step, all radio contents are identified. In the next step, music contents are identified. The key characteristics of radio and music contents are presented in **TABLE 2**. The statistical properties of radio and music traffic are presented in **Fig. 4**. The respective distinguishing characteristics, together with statistical properties, are used to identify audio traffic.

TABLE 2. TWO DIFFERENT TYPES OF AUDIO

| **Differences between Radio & Music** | |
|---|---|
| Radio flows exhibit continuous pattern | Music flows exhibit Dirac pattern |
| The max/min size of a radio flow is dependent upon the maximum flow-period configuration and the offered radio player rates | The max/min size of a music flow is dependent upon the max/min song duration and offered online music rates |
| Very high occupancy value | Very low occupancy value |
| Audio-listener behaviours : Listens to at least 5 minutes of radio-phase | Audio-listener behaviours : Listens to at least 5 minutes of music-phase |
| Maximum radio-phase timeout is based on flow-period configurations | Maximum music-phase timeout is based on maximum song duration |

The first step in the algorithm groups potential audio flows per multimedia-phase (audio-phase) based on the parameters presented in Fig. 4. For example, an audio-phase for music is a selection of songs that are continuously played without interruption. The interruption is determined by phases that are a maximum of one song length apart. Likewise, an interruption in a radio audio-phase is at least a flow-period apart since radio traffic has very high occupancy rates.

Subsequently, as per the definition of the radio or music phase type, occupancy thresholds are used to identify radio traffic or music traffic.

### *G. Proposed Video & File-transfer Algorithm*

This paper proposes to first group potential video and file-transfer traffic using the parameters presented in Fig. 4. Then, the notion of an enhanced flow-group methodology is used to improve the accuracy of the classification. The proposed technique is illustrated via the flow chart in Fig. 5. With the use of flow attributes and video parameters presented in **Fig. 4**, all traffic is grouped around video-phases. First, the flow of a video-phase is identified for analysis. Based on the flow, other related flows are inspected to identify the associated flow-group based on the definition of flow-group by Kei *et al.* The expanded definition of flow-group is then applied to further identify flow-groups based on CDNs from the same source servers. A flow-group counter is used to increment a counter per flow-group found within each video-phase. Video-phase with at least a flow-group is labelled video whereas those without a flow-group are labelled file-transfer traffic.

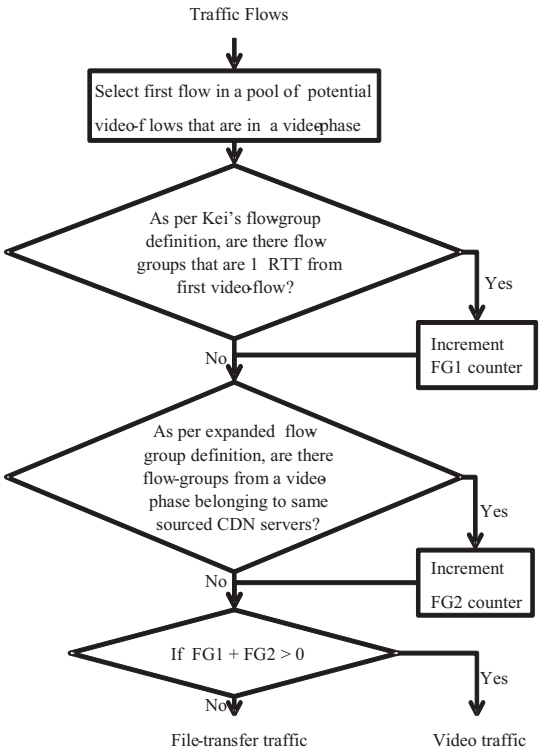



Fig. 5 Flow chart of flow-group technique to classify video & file-transfer

## IV. RESULTS

This section presents datasets, algorithm evaluation metrics and a comparison of the algorithms, results and discussions.

### A. Dataset

TABLE 3. DATASETS

| | SME-6 | SME-7 | SME-8 |
|---|---|---|---|
| Date | 01/07/2013 | 1/22/2013 | 1/23/2013 |
| Duration(s) | 24723 | 28207 | 13628 |
| Start-time (GMT-5) | 10:18:04 | 10:29:04 | 10:56:20 |
| Flows | 249822 | 287616 | 198409 |
| Packets | 13376109 | 15351639 | 10170693 |
| Bytes | 11158181285 | 13589511746 | 8728052938 |
| HTTP Flows | 75485 | 87181 | 63951 |
| HTTP Packets | 7346663 | 8814438 | 5628558 |
| HTTP Bytes | 10456335955 | 12545720613 | 7982629610 |

Eight different datasets were collected in a small and medium enterprise (SME) with controlled real world experiments. The focus was placed on traffic in a SME because one of the major cloud applications is to outsource enterprise networks to cloud networks. To this end, the datasets collected can be used to estimate the performance of the proposed approach under future cloud environments. The SME includes activity from 15 to 30 hosts. The first five datasets are used as training data. This paper describes and analyzes the proposed scheme against the last three datasets named: SME-6, SME-7 and SME-8. Further details on the datasets are provided in TABLE 3. The datasets used include online radio traffic, online music traffic, online video traffic and file-transfer traffic.

For the purpose of labelling raw data, the flow collector probe itself had the ability to employ deep-packet-inspection (DPI), detect specific applications and label them appropriately. However, only specific applications such as YouTube (video), 'lastfm.com' (radio) and Grooveshark (music) were labelled in this manner (DPI). Netflix (video), 'Centos' file and other multimedia contents were manually inspected using Wireshark and labelled accordingly. Traffic upon which classification could not be completed for a variety of reasons was labelled 'Others'. Any incorrect prediction by the proposed scheme only increases false alarms and thus does not positively increase the performance measure

The intention is to apply this proposed algorithm in a service provider network and prove an accurate classification process. Typically, a SME may not have a high volume of HTTP audio and HTTP video traffic. To emulate public network characteristics, certain measures were enacted. In particular, a high percentage of video and a relatively high percentage of audio traffic were injected into a controlled SME environment. In addition, typical file-transfer contents were also tested. Traffic was also collected during peak hours to assess classification during high network activity.

### B. Traffic Accuracy Definitions

The accuracy of the proposed algorithm is assessed on a per-application basis. The performance of the algorithm is quantified and evaluated using metrics such as precision, recall, accuracy and F-measure. The metrics are calculated through false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP) as defined below .

- *False negatives (FN) are the percentage of raw traffic (video, audio and file transfer) that this classifier failed to accurately classify.*
- *False positives (FP) are the percentage of raw traffic incorrectly classified as video, audio and file-transfer.*
- *True negatives (TN) are the percentage of non-audio, non-video and non-file-transfer traffic which have been correctly classified.*
- *True positives (TP) are the percentage of actual video, audio and file-transfer that have been correctly classified.*
- *Precision is the system's correct predictions against all predicted values. That is precision = TP / (TP+FP).*
- *Recall is the system's correct predictions against all actual correct values. That is recall = TP / (TP + FN.)*
- *F-Measure is the harmonic mean of recall and precision. That is F-measure => 2 * Precision * Recall / (Precision + Recall)*
- *Accuracy = TP + TN / (TP + FP + FN + TN).*
- *Confusion matrix is a specific table layout useful for visualizing the performance of an algorithm.*

The measuring metrics (precision, recall, accuracy and F-measure) defined above are visualized by using confusion

matrix. An unbalanced dataset of network traffic, characterized as a dataset dominated by a class or classes that have significantly more data than the rest, can be accurately evaluated using F-measure.

### C. Machine Learning Algorithm used for Comparison

The utility of our proposed classification technique is compared against the results produced by machine learning algorithms; namely Support Vector Machine (SVM) and NaiveBayes algorithm. The Weka tool was configured to run the NaiveBayes algorithm and extended to support the SVM library. The SVM used in this paper is Version 3.17 provided by Chih-Chung Chang and Chih-Jen Lin . The training datasets use 9 flow-features: source-IP, destination-IP, source-port, destination-port, bytes, packets, start-time, end-time and protocols to classify dataset.

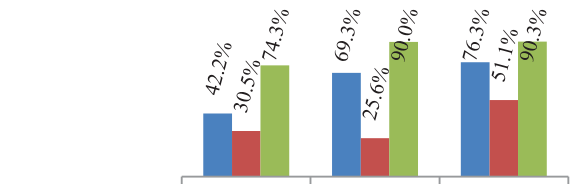
Kim *et al.* in his paper revealed 7 supervised ML methods. The paper reported that SVM had a better F-measure. As a result, SVM was considered for comparison. The NaiveBayes is a probabilistic classifier used by Kei *et al.* as one of the comparison algorithm to evaluate flow-group, and the same algorithm is used to evaluate the enhanced flow-group in this paper.

### D. Results

Three algorithms are tested on three different datasets: TABLE 4, TABLE 5 and TABLE 6 show the precision, recall, F-measure and accuracy results of audio, video and file-transfer traffic. The datasets also include an averaged performance of all types of applications-per-dataset measure. TABLE 4 and TABLE 5 represent the classification performance of two comparison algorithms. TABLE 6 presents the classification performance of the proposed algorithm. The proposed algorithm needs no training of dataset, but the comparison algorithm requires the training of dataset. In addition, TABLE 6 includes an additional row of the proposed algorithm performance per application type. The proposed algorithms, on average, can at least classify with 81% accuracy and 82% F-measure value.

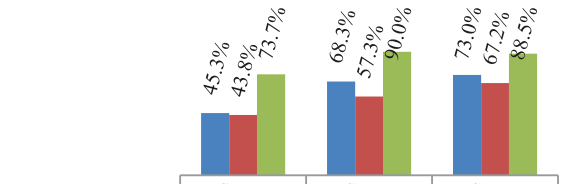
### E. Discussions

Kei *et al.* formulated the concept of a flow-group. Expanding on the work of Archibald *et al.*, Kei proposed and incorporated the flow-group into the machine learning algorithm. The use of the flow-group resulted in a 10% increase in accuracy and is classifiable above 80% F-measure mark. However, using similar algorithms from Kei *et al.*, the classification applied on SME-6, SME-7 and SME-8 indicated lower than 50% F-measure value. This difference may be attributable to the different datasets and the accuracy of labelling a raw dataset.




| | SME-6 | SME-7 | SME-8 |
|---|---|---|---|
| NaiveBayes | 42.2% | 69.3% | 76.3% |
| SVM | 30.5% | 25.6% | 51.1% |
| Proposed Scheme | 74.3% | 90.0% | 90.3% |

Fig. 6 F-Measure results of NaiveBayes, SVM and Proposed algorithms




| | SME-6 | SME-7 | SME-8 |
|---|---|---|---|
| NaiveBayes | 45.3% | 68.3% | 73.0% |
| SVM | 43.8% | 57.3% | 67.2% |
| Proposed Scheme | 73.7% | 90.0% | 88.5% |

Fig. 7 Accuracy results of NaiveBayes, SVM and Proposed algorithms

As shown in Fig. 6 and Fig. 7, the proposed scheme, on the other hand, shows more than 73% F-measure and accuracy performance in detecting video, file transfer and audio traffic per dataset. The audio traffic classification shows a very promising performance with over 83% classification F-measure. The proposed scheme classification accuracy and F-measure are at least 25% more than NaiveBayes and SVM.

The improved results from our classification scheme in comparison with other approaches can be attributed to the use of 'occupancy' and effective identification of flow-group from CDN.

## V. Conclusions

In conclusion, this paper proposes a novel classification technique to identify video, audio and file transfer traffic tunnelled over HTTP. The technique introduces the concept of 'occupancy' as a key element in the classification process and proposes an improved definition of 'flow-group' for the purpose of classifying audio, video and file-transfer traffic using NetFlow records. The classification scheme uses a combination of port-based, flow-based, and behavioural based approaches to classify HTTP tunnelled traffic.

In comparison to the results produced by NaiveBayes algorithm and SVM, the proposed scheme achieves between 13% and 50% higher accuracy rates. At least 70% of the HTTP tunnelled traffic, and in some cases 90% of the HTTP tunnelled traffic is classified accurately using this proposed scheme.

TABLE 1

CLASSIFICATION RESULTS OF AUDIO, VIDEO AND FILE-TRANSFER BY NAIVEBAYES

| | Audio | | | | | | | File-transfer | | | | | | | Video | | | | | | | Average | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| SME6 | 0.28 | 0.76 | 0.72 | 0.27 | 0.28 | 0.27 | 0.26 | 0.50 | 0.19 | 0.50 | 0.72 | 0.50 | 0.59 | 0.66 | 0.36 | 0.47 | 0.64 | 0.43 | 0.36 | 0.39 | 0.45 | 0.47 | 0.38 | 0.42 | 0.45 |
| SME7 | 0.56 | 0.43 | 0.44 | 0.57 | 0.56 | 0.56 | 0.57 | 0.75 | 0.14 | 0.25 | 0.85 | 0.75 | 0.80 | 0.81 | 0.78 | 0.42 | 0.22 | 0.65 | 0.78 | 0.71 | 0.68 | 0.69 | 0.70 | 0.69 | 0.68 |
| SME8 | 0.77 | 0.55 | 0.23 | 0.58 | 0.77 | 0.66 | 0.61 | 0.47 | 0.00 | 0.53 | 1.00 | 0.47 | 0.64 | 0.74 | 0.99 | 0.30 | 0.01 | 0.77 | 0.99 | 0.87 | 0.85 | 0.78 | 0.74 | 0.76 | 0.73 |

TABLE 2

CLASSIFICATION RESULTS OF AUDIO, VIDEO AND FILE-TRANSFER BY SVM

| | Audio | | | | | | | File-transfer | | | | | | | Video | | | | | | | Average | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| SME6 | 0.18 | 0.95 | 0.82 | 0.16 | 0.18 | 0.17 | 0.12 | 0.13 | 0.00 | 0.87 | 1.00 | 0.13 | 0.23 | 0.57 | 0.27 | 0.00 | 0.73 | 1.00 | 0.27 | 0.43 | 0.64 | 0.72 | 0.19 | 0.30 | 0.44 |
| SME7 | 0.12 | 0.00 | 0.88 | 1.00 | 0.12 | 0.21 | 0.56 | 0.07 | 0.00 | 0.93 | 1.00 | 0.07 | 0.13 | 0.54 | 0.25 | 0.00 | 0.75 | 1.00 | 0.25 | 0.40 | 0.63 | 1.00 | 0.15 | 0.26 | 0.57 |
| SME8 | 0.43 | 0.00 | 0.57 | 1.00 | 0.43 | 0.60 | 0.72 | 0.33 | 0.00 | 0.67 | 1.00 | 0.33 | 0.50 | 0.67 | 0.27 | 0.00 | 0.73 | 1.00 | 0.27 | 0.43 | 0.64 | 1.00 | 0.34 | 0.51 | 0.67 |

TABLE 3

CLASSIFICATION RESULTS OF AUDIO, VIDEO AND FILE-TRANSFER BY PROPOSED SCHEME

| | Audio | | | | | | | File-transfer | | | | | | | Video | | | | | | | Average | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | TP | FP | FN | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| SME6 | 0.18 | 0.95 | 0.82 | 0.16 | 0.18 | 0.17 | 0.12 | 0.13 | 0.00 | 0.87 | 1.00 | 0.13 | 0.23 | 0.57 | 0.27 | 0.00 | 0.73 | 1.00 | 0.27 | 0.43 | 0.64 | 0.72 | 0.19 | 0.30 | 0.44 |
| SME7 | 0.12 | 0.00 | 0.88 | 1.00 | 0.12 | 0.21 | 0.56 | 0.07 | 0.00 | 0.93 | 1.00 | 0.07 | 0.13 | 0.54 | 0.25 | 0.00 | 0.75 | 1.00 | 0.25 | 0.40 | 0.63 | 1.00 | 0.15 | 0.26 | 0.57 |
| SME8 | 0.43 | 0.00 | 0.57 | 1.00 | 0.43 | 0.60 | 0.72 | 0.33 | 0.00 | 0.67 | 1.00 | 0.33 | 0.50 | 0.67 | 0.27 | 0.00 | 0.73 | 1.00 | 0.27 | 0.43 | 0.64 | 1.00 | 0.34 | 0.51 | 0.67 |

## REFERENCES


[1] D Moore, K Keys, R Koga, E Lagache, and K Claffy, "The CoralReef software suite as a tool for system and network administrators," in *Proceedings of the 15th USENIX conference on System administration*, San Diego, California, 2001.

[2] V. Carela-Espanol, P. Barlet-Ros, J. Sole-Pareta, *Traffic Classification with Sampled NetFlow,* Traffic, 2009.

[3] B.Park, Y.J.Won, M. S. Kim, J.W.Hong, "Towards Automated Application Signature," in *IEEE on Network Operations and Management Symposium(NOMS 2008), pp.160-167*, Salvador, 2008.

[4] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh and G. Varghese, "Network monitoring using traffic dispersion graphs," in *7th ACM Special Interest Group on Data Communication Conference on on Internet measurement (SIGCOMM 2007), pp. 315-320*, New York, 2007.

[5] G. Szabo, I. Szab, D. Orincsay, "Accurate Traffic Classification," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM 2007), pp.1-8*, Helsinki, 2007.

[6] B. Augustin, A. Mellouk, "On Traffic Patterns of HTTP Applications," in *IEEE on Global Telecommunications Conference (GLOBECOM 2011), pp.1-6*, Texas, 2011.

[7] K. Takeshita, T. Kurosawa, M. Tsujino, M. Iwashita, "Evaluation of HTTP Video Classification Method Using Flow Group Information," in *14th IEEE lTelecommunications Network Strategy and Planning Symposium (NETWORKS 2010)*, Warsaw, 2010.

[8] S. Kaoprakhon , V. Visoottiviseth, "Classification of Audio and Video Traffic over HTTP Protocol," in *9th IEEE International Symposium on Communications and Information Technology(ISCIT 2009), pp. 1534-1539* , Incheon, 2009.

[9] Xu Cheng, Jiangchuan Liu, and Cameron Dale, "Understanding the Characteristics of Internet Short Video Sharing: A YouTube-based Measurement," in *IEEE Transactions on Multimedia*, 2010.

[10] M. Korczynski and A. Duda, "Classifying service flows in the encrypted skype traffic," in *International Conference on Communications, pp.1064 - 1068*, Ottawa, 2012.

[11] *CISCO "Cisco IOS Flexible NetFlow Overview",* http://www.cisco.com/en/US/docs/ios/fnetflow/configuration/guide/fnetflow_overview.pdf.

[12] V. Labatut,H. Cherif, "Accuracy Measures for the Comparison of Classifiers," in *International Conference on Information Technology(ICIT 2011), pp.1-5*, Amman, 2011.

[13] C. Hsu, C. Chang, C. Lin, "A Practical Guide to Support Vector Classification," Department of Computer Science, National Taiwan University, Taipei, 2003.

[14] *Weka 3: Data Mining Software in Java,* http://www.cs.waikato.ac.nz/ml/weka/index.html.

[15] *A Library for Support Vector Machines - V3.16,* http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[16] H. Kim, k. claffy, M. Fomenkov, D. Barman, M. Faloutsos, K. Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices," in *6th ACM Special Interest Group on Data Communication on Conference on emerging Networking EXperiments and Technologies (SIGCOMM 2008), pp.11*, New York, 2008.