I Know What You Saw Last Minute - The Chrome Browser Case

Ran Dubin[1], Amit Dvir[2] , Ofir Pele[2,3] , Ofer Hadar[1]

1. Department of Communication System Engineering, Ben-Gurion University of the Negev, Israel.
2. Center for Cyber Technologies, Department of Computer Science, Ariel University, Israel.
3. Department of Electrical and Electronics Engineering, Ariel University, Israel.

About Me

- Ph.D. candidate at Ben-Gurion University, Israel
  - Optimization of HTTP adaptive streaming
  - Encrypted network traffic classification problems

- Senior data scientist at Seculert
  - Seculert develop an automated breach analytics platform in the cloud.
  - Supervised machine learning for detection of malicious activity within the enterprise network

Agenda

- Motivation
- The scenario
- Our goal
- How can “I know what you saw”?
- Related works
- Proposed algorithm
- Results

Motivation

- Google encourages network privacy:
  - “77 percent of Google online traffic is encrypted"[1]
  - “Google started giving HTTPS pages a ranking boost”
- HTTPS keeps your data anonymous:
  - “No one will be able to snoop on the traffic — such as your ISP"[2]
  - Let’s try to break it!

[1] http://gadgets.ndtv.com/internet/news/google-reveals-77-percent-of-its-online-traffic-is-encrypted-814191

[2] http://www.makeuseof.com/tag/can-you-really-be-anonymous-online/

The Scenario

- Passive Sniffing:
  - Traffic control and optimization
  - *Open Source Intelligence Techniques (OSINT) vector*[3]
    - *Web searches, visited sites ..*

- YouTube is the world's leading social network video platform
  - YouTube is used also large protests and propaganda!
  - Protecting user privacy and viewing habits is important!

[3] ISPs Sell Your Data to Advertisers, But FCC has a Plan to Protect Privacy, http://thehackernews.com/2016/03/isp-sells-data-to-advertisers.html

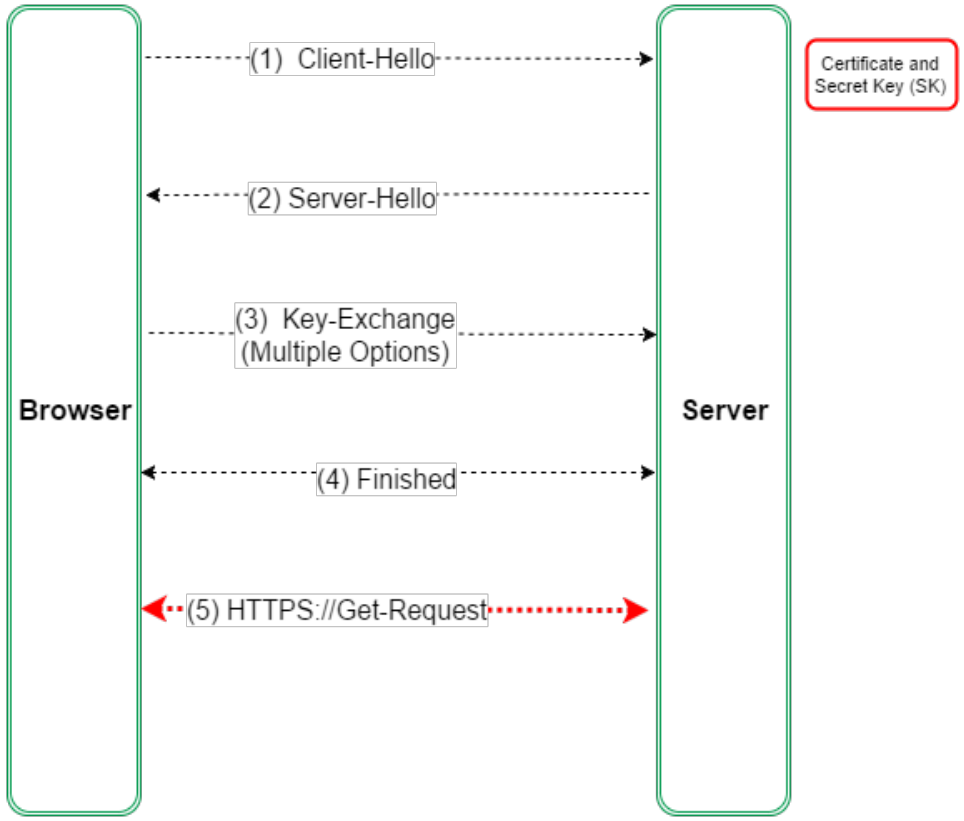
Our Goal

- **To show that HTTPS\2.0 is not enough in order to protect your viewing habits**

- **Contribution:**
  - **Dataset**
  - **Data crawler** - based on selenium
  - **New encrypted traffic feature and classification algorithm**

Brief Partial Overview of SSL/TLS

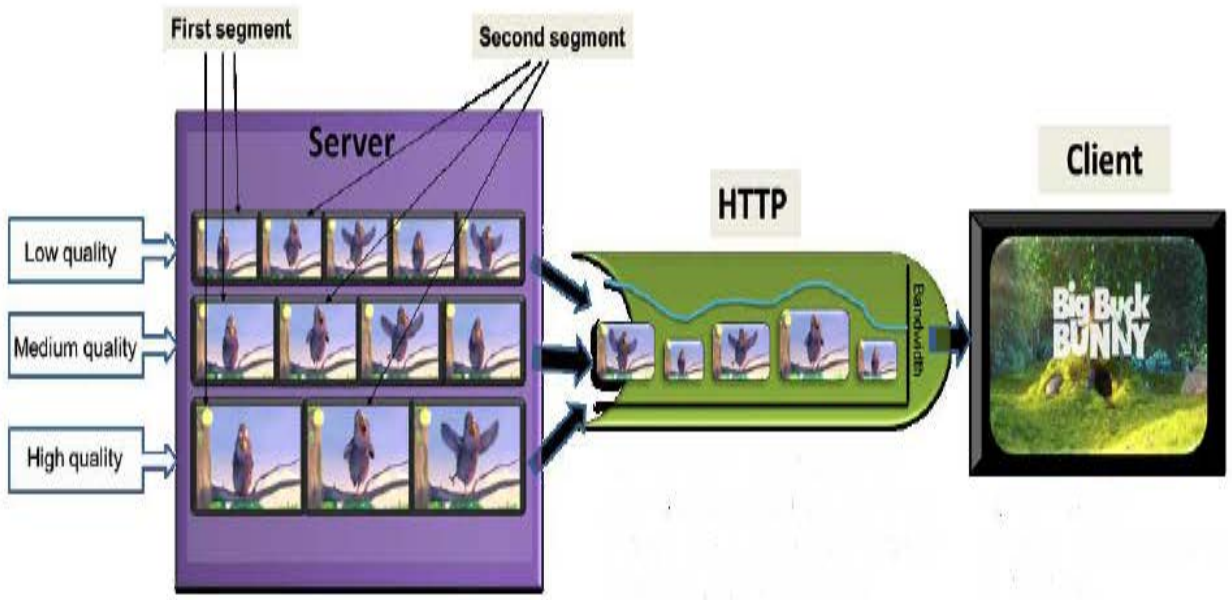
- Step (0): browse to: https://www.youtube.com/watch?v=_bP6aVG6L1w
- Step (1): use Service Name Indicator
- Step (5): content and header are fully encrypted
  - HTTPS request (URL) is not visible in the encrypted traffic
  - All HTTP headers are encrypted



How Can “I Know What You Saw”?

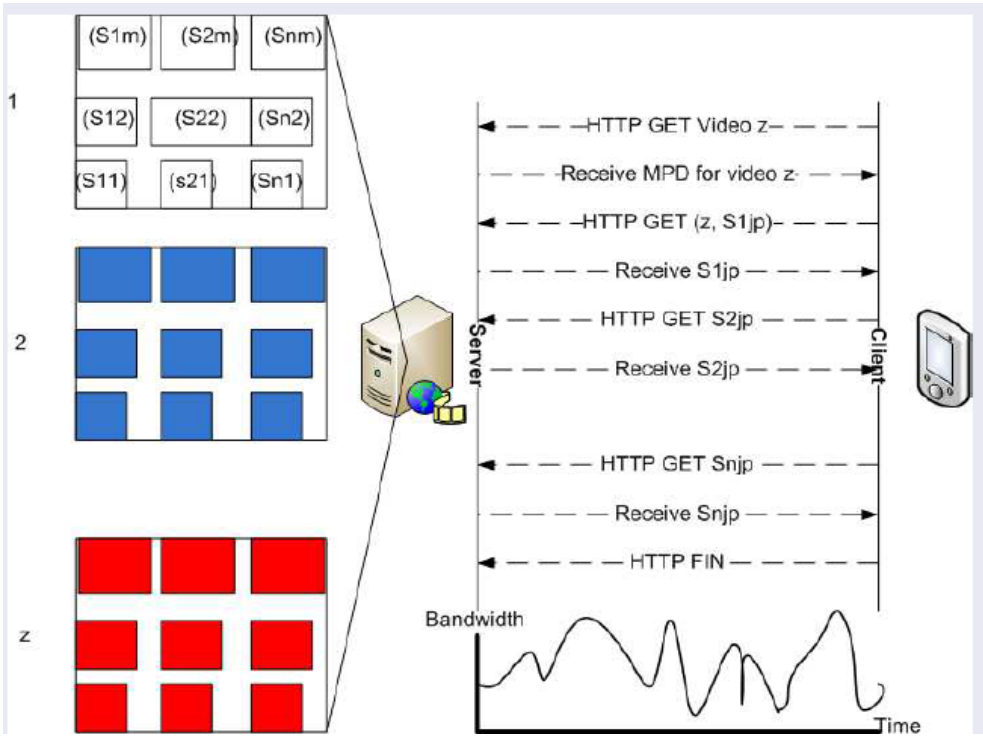
1. How are YouTube videos encoded?
2. How is the video downloaded?
3. What is the video download behavior in the network?
4. How to tie everything together for a classification?

Introduction To HTTP Adaptive Streaming (HAS)



Multi variable bit-rate streaming

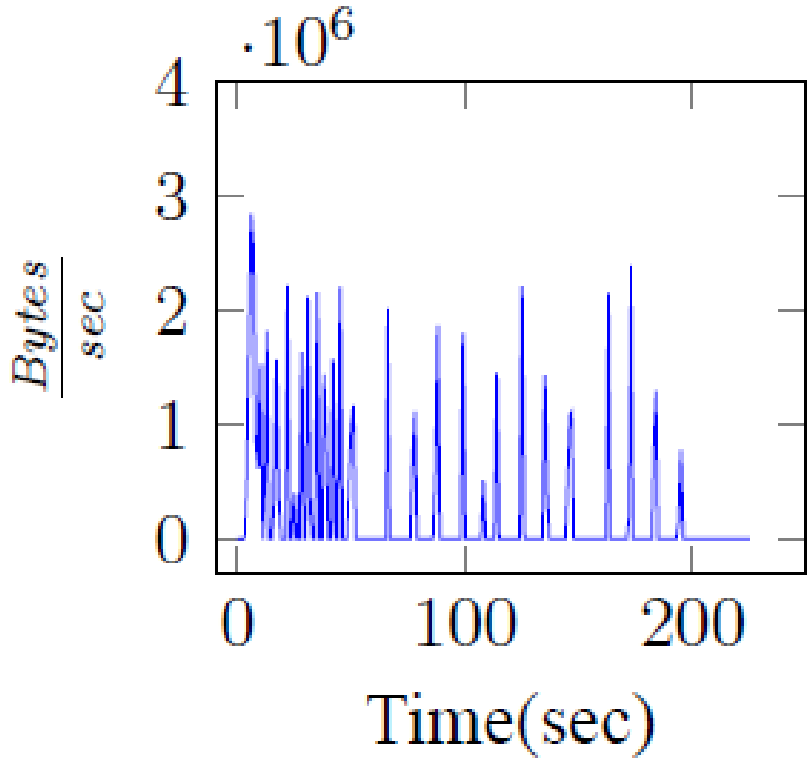
HAS Diagram



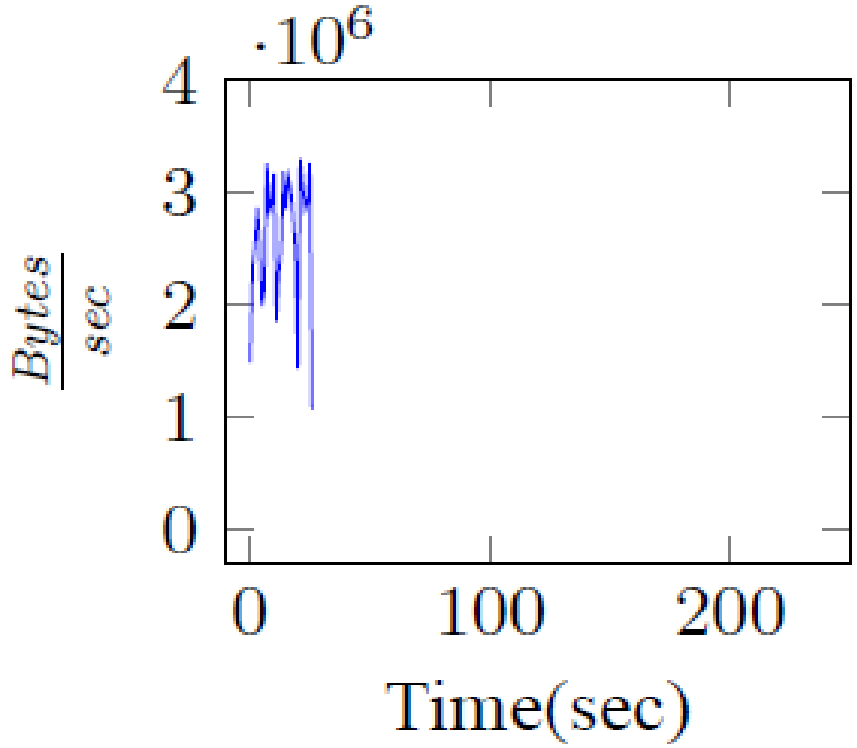
HAS Download Diagram

* How YouTube Works: https://www.youtube.com/watch?v=UklDSMG9ffU

YouTube Encrypted Network Traffic

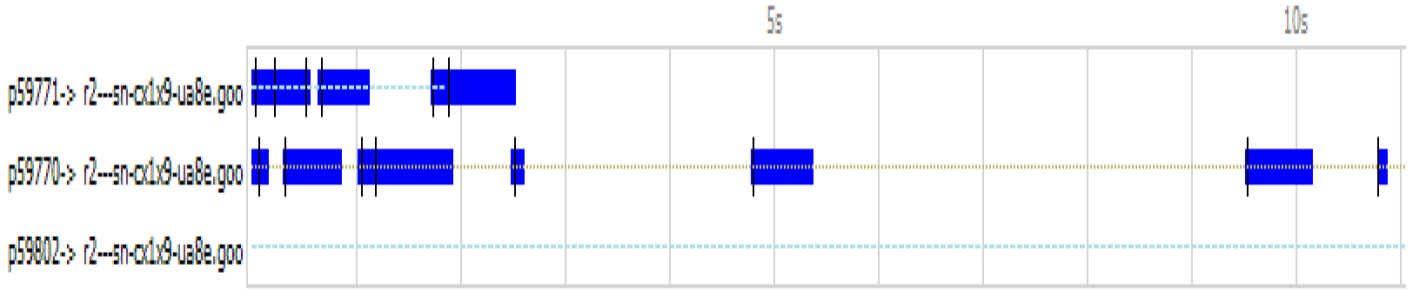


Chrome Automatic Mode

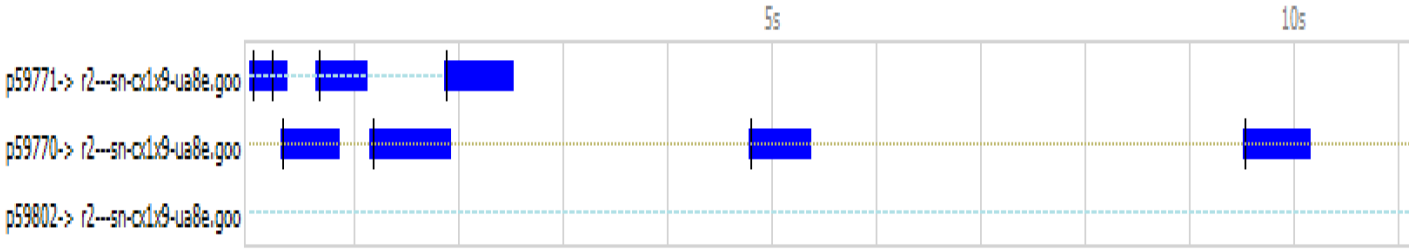


Chrome Fixed Mode

YouTube Flow Patterns – The Web Proxy Perspective



First 10 seconds of downloading audio + video

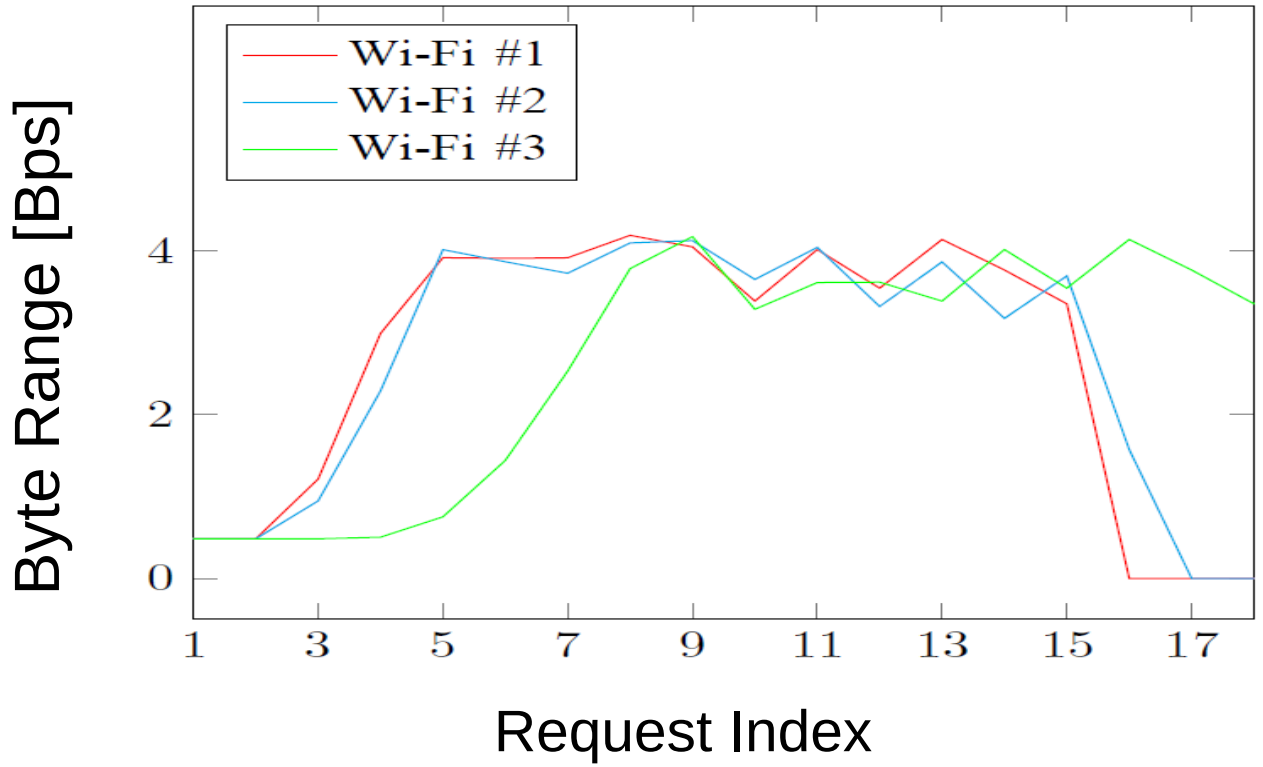


First 10 seconds of downloading only video

- Mixture of audio/video in a single flow
- HTTP2 - multiplexed application layer protocol
- Multi-Bit-Rate Video Encoding

YouTube HTTP Byte Range

Fiddler (Video) Stream Request Vs Byte Range



Related Works

1. Most discuss application type classification and not content classification
2. HTTPS classification was found to achieve low accuracy
3. Wright et al. exploit the VBR codec characteristics of encrypted Voice Over Internet Protocol (VOIP) for language identification
4. Liu et al. and Saponas et al. presented methods for video title classification of RTP/UDP and TCP internet traffic (not MBR)
5. **Changes in video traffic over the Internet:**
   - **HTTP byte range selection over HTTP**
   - **MBR adaptive streaming**
   - **HTTP version 2**

Proposed Machine Learning Solution

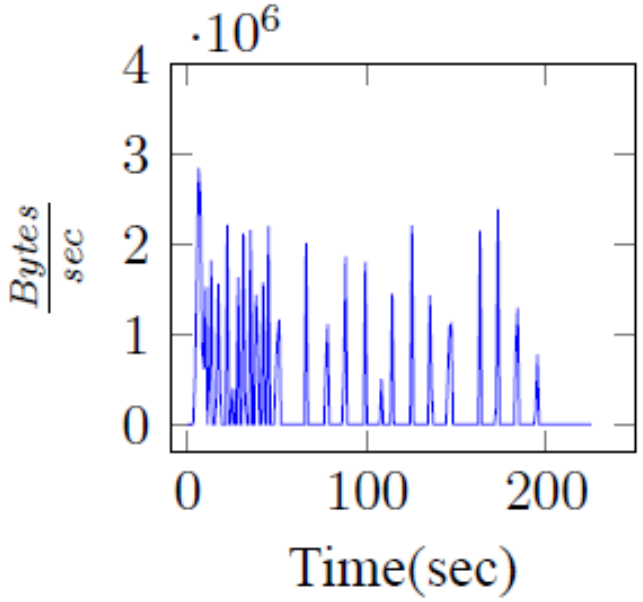
1. Traffic Analysis
2. Traffic Features
3. Traffic Preprocessing
4. Machine Learning Algorithms

Feature Extraction

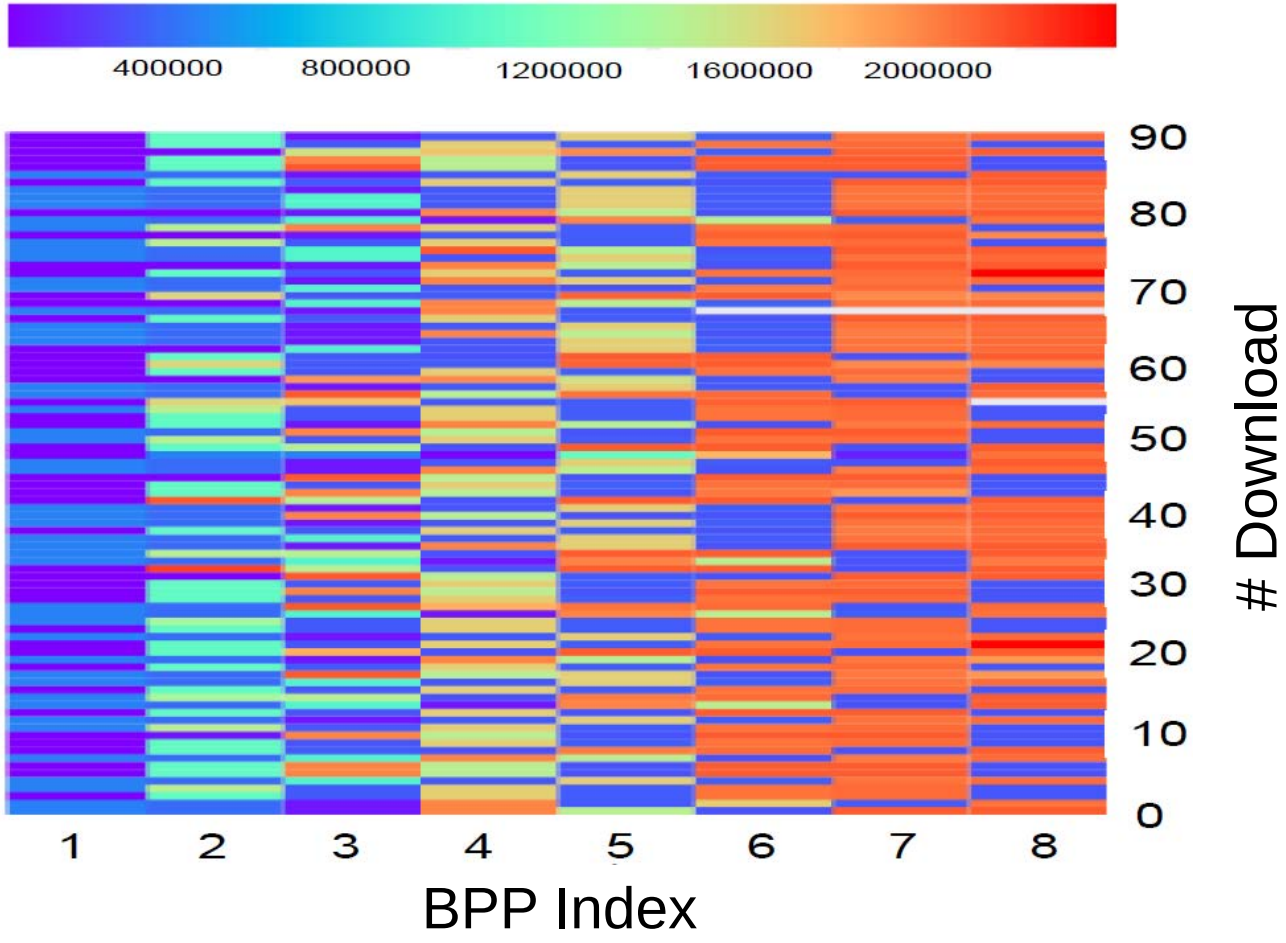
1. Many features:
   Number of packets in a session, payload size, information bit rate,
   Round-Trip Time (RTT), packet time differences

2. Bit Per Peak (BPP): Sum of bytes in each peak after TCP ACK mechanism

3. Why BPP?
   - Represent the traffic On/Off behavior
   - Real time classification constraints
   - Compact feature representation
   - Robust to packet loss and delays



BPP Index Vs Download Copy



Pre-Processing

- With/without audio removal
- <400 Kbytes BPPs are considered as audio

Proposed algorithms

1. Support Vector Machines (SVM) with Radial Basis Function (RBF)
   - With a BPP feature vector

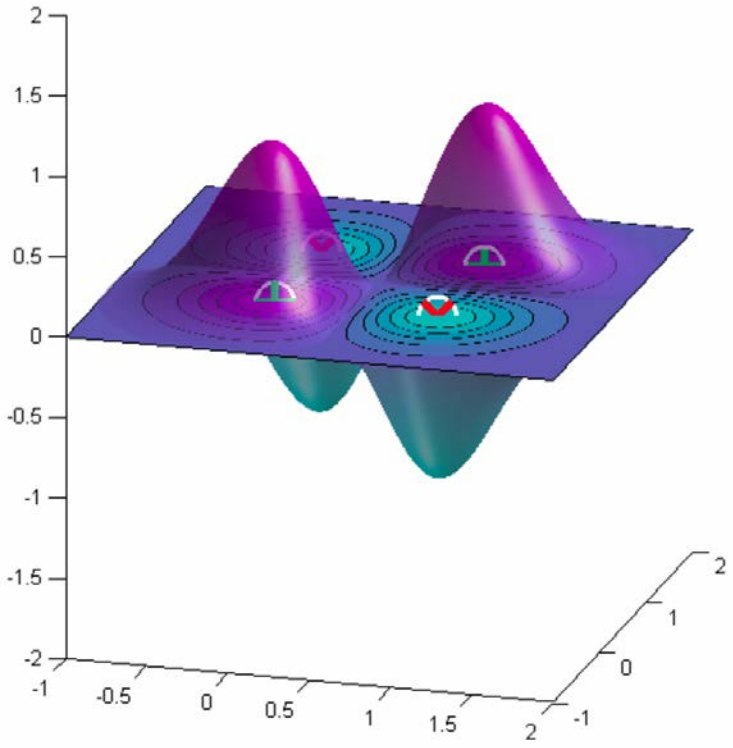
2. Nearest Neighbor Algorithm – NN
   - With a set of BPP features

SVM with Radial Basis Function (RBF) Kernel

- SVM RBF maps data to high dimensional space. The classifier:

$$f(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i y_i \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2\right) + b$$

- Ongoing work uses SVM with intersection similarities as features



BPP Set Feature

- S_{ij} is a set of Bit-Per-Peak (BPP) features (no duplicates)
  - *i* - video title index
  - *j* - stream index

- Note that each BPP-set may have different cardinality

NN Algorithm

- Similarity score between two BPP-sets is the cardinality of the intersection set:

$$\mathrm{sim}(S, S') = |S \cap S'|$$

- At test time, each video stream BPP-set, S_{test}, is classified as the video title i (class) that matches the maximum similarity score to class index. m_i is the number of streams per title i:

$$\forall \ 1 \le i \le n, \quad s_i = \max_{j=1}^{m_i} \mathrm{sim}(\mathcal{S}_{\mathrm{test}}, \mathcal{S}_{ij})$$

$$y(\mathcal{S}_{\mathrm{test}}) = \begin{cases} \operatorname{argmax}_{i=1}^{n} s_i & \text{if } \left(\max_{i=1}^{n} s_i\right) > \mathrm{Thr} \\ \text{unknown} & \text{otherwise} \end{cases}$$

Dataset

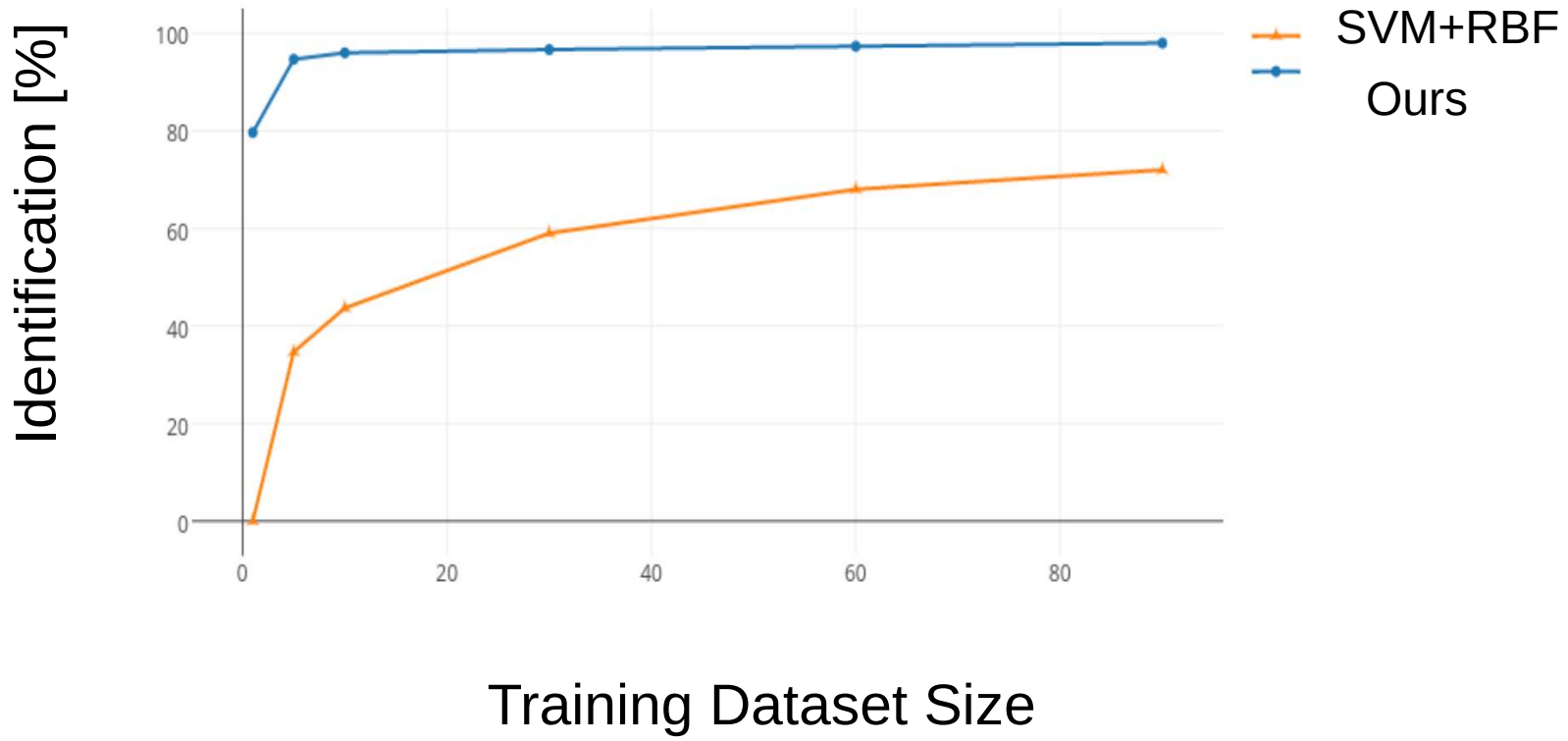
Train/ test: 30 different titles, each with 100 streams copies (Train- 90, Test -10)

Videos outside of the dataset: 200 additional different video titles (titles not in the regular dataset used only in testing)

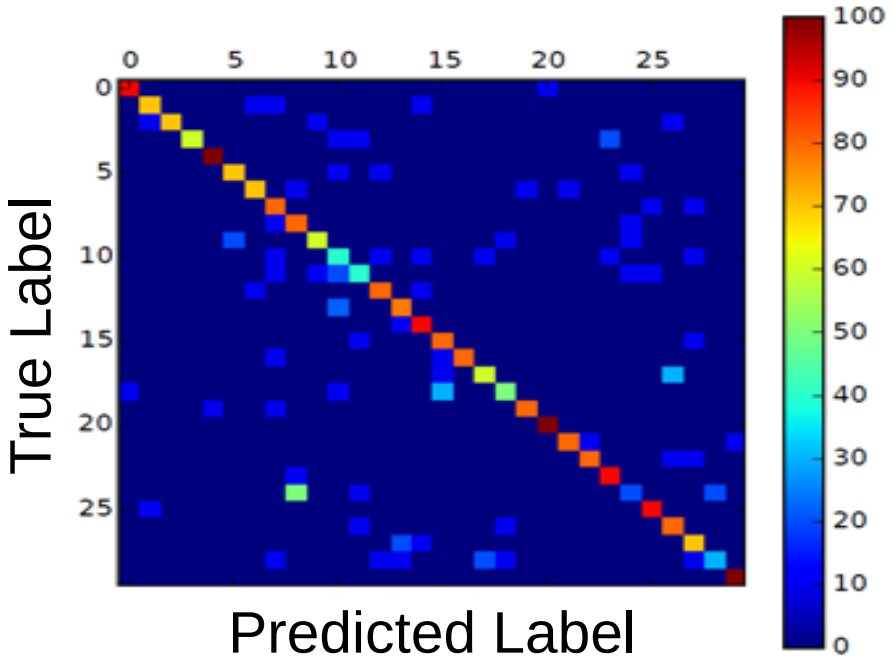
Added delay evaluation: 4 subsets with added delay of 100/300/600/900 ms. (10 titles with 10 different downloads)

Added packet loss evaluation: 4 subsets with added packet loss of 1/3/6/9 % (10 titles with 10 different downloads)

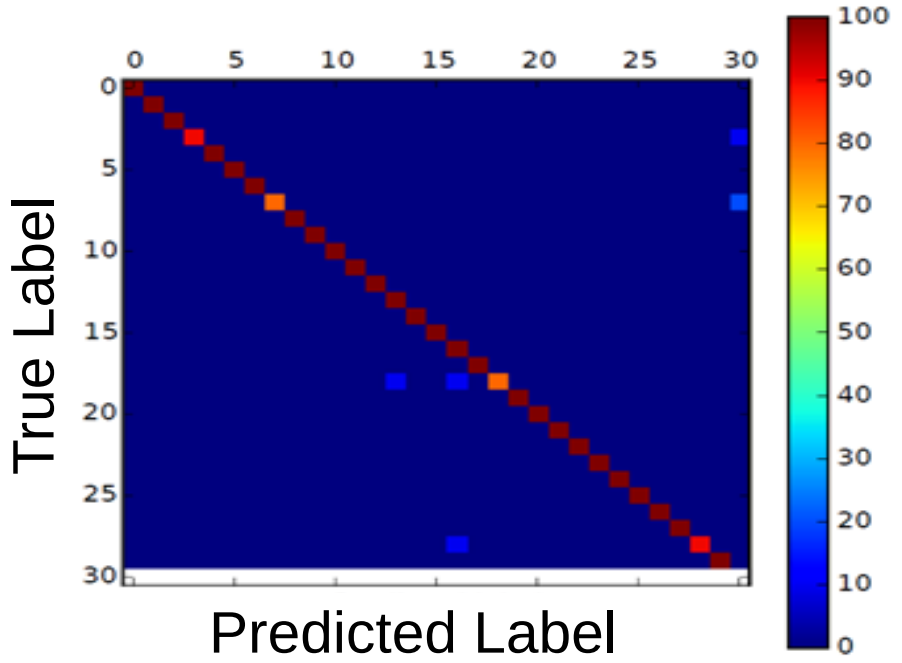
Classification Accuracy



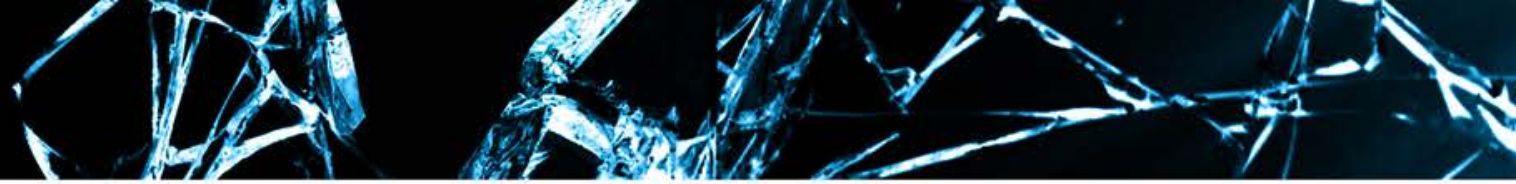
Confusion Matrices



SVM+RBF (accuracy 72%)



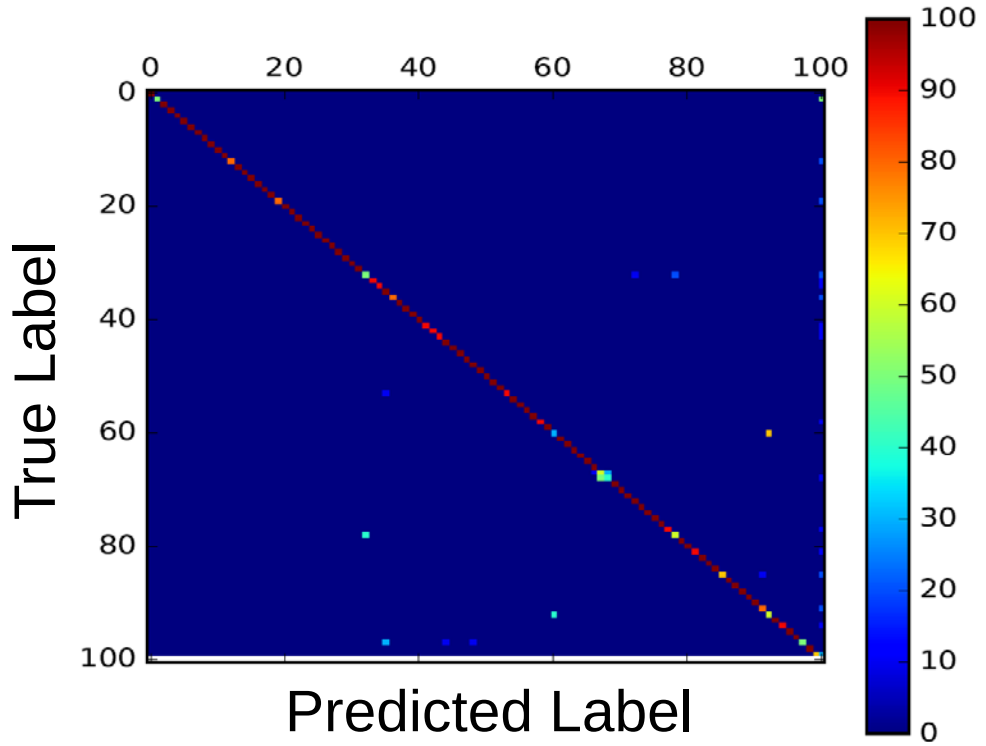
Ours (accuracy 98%)



Classification of Unknown Videos:

100% accuracy

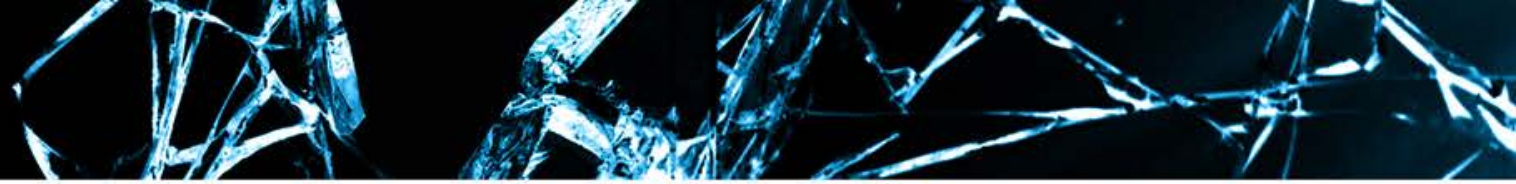
Ongoing Results



Accuracy: 93.6%

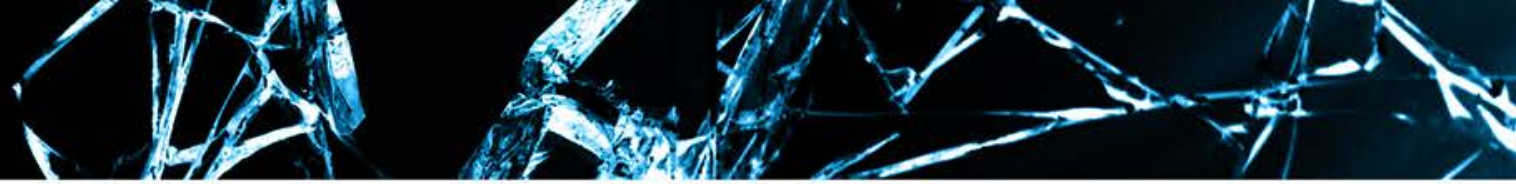
Conclusions

- Created an OSINT vector from YouTube video traffic
- We demonstrated that HTTP2.0 is not protecting your viewing habits.
- NN algorithm - 98% accuracy
- BPP feature - is robust to high network delays and packet loss
- Ongoing research – 10000 streams of 100 titles, similar results
- Contribution – crawler, dataset and algorithms



Thank you!☺

Questions?



Backup Slides

Different Network Conditions