Robust Streaming Video Traffic Classification
Jordan Ebel, December 2015
Computer Science Department, Stanford University

Background
Internet service providers use network monitoring tools to manage their networks and ensure quality of service for all subscribers. The growth of streaming video and encrypted traffic have made traditional management tools ineffective. This project aims to:

• Identify a top performing machine learning algorithm to classify streaming video traffic
• Build a real time system to classify traffic while maintaining performance, privacy, and video quality

Methods
• Datasets collected of streaming video and Internet browsing using the Wireshark packet analyzer
• Python script implemented to extract features and identify true packet classifications
• Logistic regression, Naïve Bayes, SVM, and K-Means Clustering algorithms prototyped and tuned in Matlab
• Top performing algorithm implemented in Python
• Real time traffic classification system written using trained algorithm and live feature extraction

Data
• Packets captured from the test computer for one minute, resulting in approximately 80,000 test and train packets
• Packets captures contain the timestamp and entire contents of each packet
• A visual observation of the training packet capture revealed interesting patterns. Streaming video traffic usually occurred in bursts, with a large number of very similar packets appearing at short intervals. Standard web traffic was more irregular and featured a wider variety in the features of the packets.

Features
To capture the patterns found in the data, features describing individual packets and groups of packets were extracted, with a focus on not including features that would diminish the robustness of the system. Figure 2 below details the features and their sources.

Supervised Learning Results
The SVM algorithm with a Gaussian kernel proved to be the top performing algorithm on the dataset. The SVM algorithm displayed test and training errors below 10% for a wide range of training set sizes, as shown in Figure 4.

Real Time System
The SVM algorithm was selected as the algorithm to implement in a real time classification system, shown below in Figure 7 operating on live traffic.

Unsupervised Learning Results
The K-Means Clustering algorithm failed to identify either traffic group accurately. Figure 6 below shows an inaccurate grouping of clusters to labels.

Figure 1: Packet number, Time, Destination IP, Source IP, Protocol, Length, and Description of Streaming Video Packets

Figure 2: Packet Features

Figure 3: Biplot of Training Dataset in Principle Component Space

Figure 4: Test and Train Errors

Figure 5: ROC Curves

Figure 6: Clustering of Data in Principle Component Space

Figure 7: Real Time Classifier In Operation