Jacob Serfaty
Machine Learning for Computer Systems Project Proposal

Group Members: Jacob Serfaty (just me)

Project Title: Adding Packet Inter-Arrival Time Statistics to NetML STAT Features

Project Summary:
NetML is a very useful library for reading PCAP files, converting the packets to flows, and extracting features from the flows. However, the statistics features (returned when requesting "STAT" features) are missing lots of common features that can help train networks. Particularly, there are no statistics included about Inter-Arrival Time (IAT)! These features are particularly useful as they reveal information about the timing in which data was transferred, which can be tremendously useful for contexts in which timing can distinguish factors that packet size cannot. For example, when trying to identify streaming services based on network traffic, all streaming services will be transmitting similar amounts of data, but the timing of the data transmissions may differ between streaming services.

In this project, I will add IAT features to the NetML library, including mean, median, standard deviation, min, max, lower quartile, and upper quartile. I will then train a model to identify streaming services based on network traffic (the first 10 SYN packets for each video session) using the NetML STATS features. I will train the model both with and without these new features and compare the results.

Data: I will use the streaming video service dataset linked in the nPrint GitHub site. I believe I will have to convert this from a .pcapng file to a standard .pcap, though I will address this issue when I work on the project.

Machine Learning: I will train 2 Random Forest Classifier Models – One with only the default STATS features and one with the IAT stats also included. I will use the same training and testing data for both models (aside from the extra features for the second model).

Evaluation: I will determine whether the additional features improved the model by comparing the accuracy, precision and recall, ROC/AUC, and confusion matrices between both models.

Learning Objective: I am hoping to learn how the addition of new features can affect the quality of machine learning models. I am also hoping to learn how effective statistics about timing can be when making inferences based on network data.