Network Analysis for Streaming Traffic Analysis

The paper talks about how machine learning based network analytics can be useful on traffic flows, both for network operations and forensic purposes. Systems that can assist human experts capable of analyzing “unknown background” traffic can be designed and developed.

The paper in its sections discusses the methodologies that can be used offline for streaming, the datasets and the evaluations that can be employed. The paper concludes by discussing the results.

For streaming the analytics and machine learning models such as stream GP and random forest can be used. For offline streaming analysis random forest along with Apache spark framework which enables offline trained random forest model to be applied for streaming.

The dataset used by the authors is the mixture of CTU-13 botnet dataset. The dataset contains all the different botnet attacks and is a combination of thirteen datasets. The random

Gini impurity index is used as the metric for measuring the performance of the random forest. 50 trees is set as the random forest parameter whereas for the streaming GP model the gap is set to 20. In regards to tools, Apache spark and microsoft power BI are used to make the analysis. The Stream GP dynamically changes the classifier and the random forest models are trained on the known behaviors.

Through the Stream GP based traffic analysis we can get the information about the unknown traffic and learn the ground truth. In a realistic scenario the Stream GP method is most helpful.