Real-time Traffic Classification & Graph Analytics for SCinet

David Ediger Dan Campbell Jason Poovey Trevor Goodyear
Georgia Tech Research Institute
Atlanta, Georgia

***Abstract*—The Innovative Computing group at Georgia Tech Research Institute (GTRI) develops algorithms and designs high performance software tools for real-time graph analytics, clustering, and classification. We propose to use SCinet at SC14 to evaluate and demonstrate our capabilities for network monitoring and streaming data analytics. The Real-time Classifier (RTC) uses non-negative matrix factorization to cluster and classify IP connections based on addresses, flags, protocols, data length, and content. STINGER, a streaming graph database, uses advanced algorithms for community identification to cluster groups of endpoints based on their interactions and behavior.**

I. Traffic Classification

Real-time classification of network connections and packet data is an important building block to anomaly and intrusion detection. The Real-time Classifier (RTC) is a new application of several hierarchical clustering and learning algorithms that can rapidly and efficiently classify data points into multiple classes. The RTC has been successfully applied to text document corpuses at the rate of 150,000 samples per second.

At SC14, we propose to evaluate the classifier on IP flows within SCinet. With IP connection data, we classify endpoints into groups by their online behavior. With advanced clustering, we may be able to identify groups of IP hosts, web sites, organizations, or technical research areas. The user will be presented with a sampled training set of current data that will be clustered to reveal prominent features. Interactively, the user will refine the clustering to produce a model. This model can be executed as a classifier against the real-time stream of traffic within SCinet.

GTRI will provide all necessary compute hardware and will work with SCinet to get relevent data, which may include Netflow data. Conference attendees will be able to view the real-time results of the analytics at the Georgia Tech booth.

II. Graph Analysis

Classical graph algorithms have proven useful in solving a variety of network-related problems. We have developed an open source in-memory graph database (STINGER) to analyze high-speed streams of network data using massively parallel systems. STINGER is capable of ingesting and analyzing over 3 million new graph edges per second. STINGER calculates structural features such as shortest paths, spanning trees, and clustering coefficients, as well as measures of influence such as betweenness centrality and eigenvector centrality.

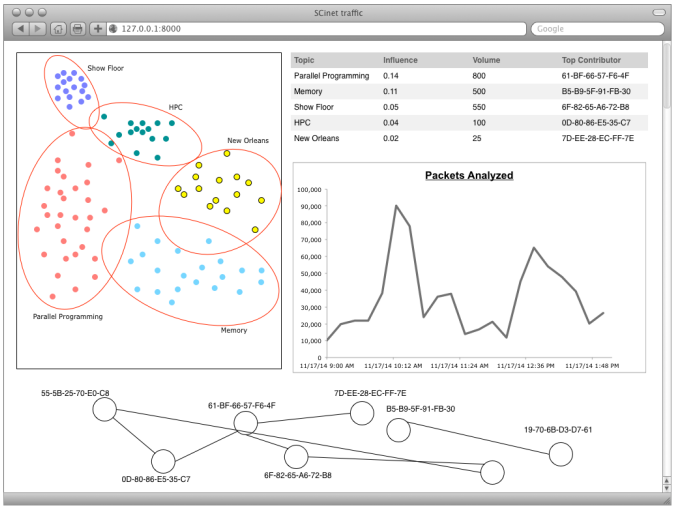

Fig. 1. Mockup of live dashboard

From Netflow data, we build a graph of IP address to IP address with feature vectors including time, protocol, and bytes transferred. We will run our suite of analysis algorithms updating at regular intervals of 5 to 30 seconds. At each moment, we calculate the most influential IP addresses, predict new connections, and identify hosts with unusual features. With community detection, we will try to cluster hosts by their connection behavior.

GTRI will provide all necessary compute hardware and will work with SCinet to ingest relevant data. Conference attendees will be able to view the real-time results of the graph analytics at the Georgia Tech booth.

III. Related Activities

In addition to our activities related to SCinet, GTRI will be using SC14 to demonstrate its text classification and graph analytics tools on social media related to the conference. We will be analyzing Tweets geo-tagged at the conference center or using the hastag #sc14 from the Twitter 1% sample stream. We will build a graph of mentions and re-tweets and analyze its structure in real time to identify influencers.