Report

Network traffic analytics has always been an important field for cybersecurity and network operations. The analysis could be done on the stream of data (network traffic) In an online mode for real-time network operations or on the captured traffic data in an off-line mode for forensic analysis.
Network traffic analysis is a fundamental part of network security and operations. The more we know about the traffic we are dealing with, the more we could provide better services and protect the network from malicious activities.
In real-world scenarios, not all network traffic is known to the network expert who labels the traffic. Its goal is to detect the behaviors of the CTU-13 datasets i.e normal, botnet, C&C, in the background traffic and differentiate those from the “unknown unknowns”.

Tools: 1) Apache Spark: Apache Spark is as clustering framework suitable for working on big data analysis and computation. Spark MLlib is a distributed ML library on top of Spark core.
2) Microsoft Power BI: Microsoft power BI is a business analytics solution for visualization of data.

Applications: Its main application is, Analyse network traffic: Network traffic analysis (NTA) is a method of monitoring network availability and activity to identify anomalies, including security and operational issues. Common use cases for NTA include Collecting a real-time and historical record of what's happening on your network. Detecting malware such as ransomware activity.

Results: a) Network Topology: The local network topology for CTU-13 datasets. Normal and botnet systems along their IP addressed and demonstrated. b) Traffic Analysis: The CUT-13 mixed background is investigated by two means: Stream-GP and Random Forest.

Conclusion: The background traffic in CTU-13 botnet dataset is investigated by means of a streaming ML classifier, Stream-GP, and offline Classifier, Random Forest. The outcome of the two different models are models are analyzed and compared. Stream GP-based traffic analysis not only sheds light on unknown traffic but also seems to generalize what it learns given the ground truth traffic better. It could conclude that the human analyst under a more realistic scenario. Future work will explore such unknown traffic analysis on other datasets for more benchmarking efforts.