MIT LINCOLN LABORATORY

The Bulletin

News, events, and notices transmitted weekly

March 20, 2020

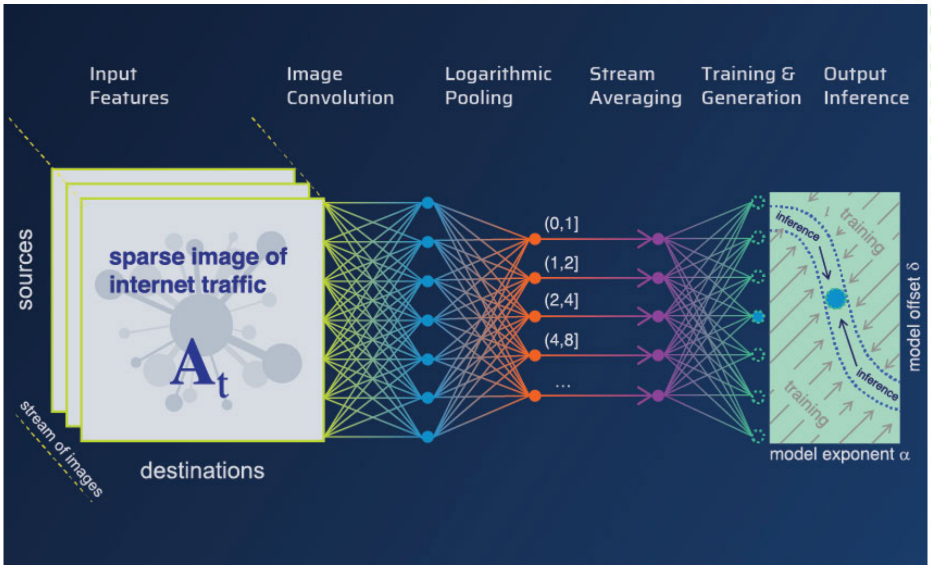
Cybersecurity Phenomenology Exploration and Reasoning

Technology Office

Most cyber defense tools identify potential malicious events based on known signatures. Methods that are not signature-based are used to identify new, unknown attack vectors. These methods typically include establishing models to baseline normal traffic behavior and then using those models to find anomalous traffic patterns. The accuracy of these models is dependent on the amount of data used to baseline the network activity. The analysis pipeline to construct these models requires the support of high-performance computing in order to run computationally demanding analytics over large quantities of network packet data.

As part of a Technology Office Seedling project, Laboratory researchers developed a neural network pipeline and utilized the Lincoln Laboratory Supercomputing Center to analyze large, publicly available internet traffic datasets. These data sets comprise 50 billion data packets collected at different locations and at different time points over a period of several years. As part of this work, a low-dimension distribution model was found to accurately fit a range of network characteristics. The parameters of the distribution model were able to delineate different network traffic topologies, which therefore provided a method for characterizing normal traffic behaviors.

The Line-funded Cybersecurity Phenomenology Exploration and Reasoning (CyPhER) project will expand on this prior work by further optimizing the data processing and analysis pipeline, which will enable using larger network traffic matrices to build models. This effort will include integrating GraphBlas, which is a suite of tools for analyzing sparse matrices. The low-dimension distribution models examined as part of the Seedling project will also be further validated using data from the Lincoln Research Network Operations Center. These data will be used to examine the stability of the distribution models over time. With validation, the network traffic baseline models will be utilized to develop anomaly detection algorithms that are not signature-based. The techniques and tools developed as part of this project will improve the efficiency of analyzing traffic matrices and the characterization of network traffic at a large scale.



The approach used in Cybersecurity Phenomenology Exploration and Reasoning for analyzing sparse traffic matrices is depicted in this figure, from streaming snapshots of network traffic to model fitting.

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.