Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Simulation of Large-scale NetFlow Data with Botnet Activity

CHALLENGE

Researchers in academia and government need large NetFlow graphs to use in testing graph-analytic algorithms (e.g., for malicious node detection) because small to medium NetFlow graphs no longer represent today's web-scale graphs. However, a lack of such data sets in the public domain makes it more difficult for security researchers in academia and government to collaborate. Thus, to test new detection approaches on large-scale data, larger NetFlow data sets that contain labeled malicious activity are needed.

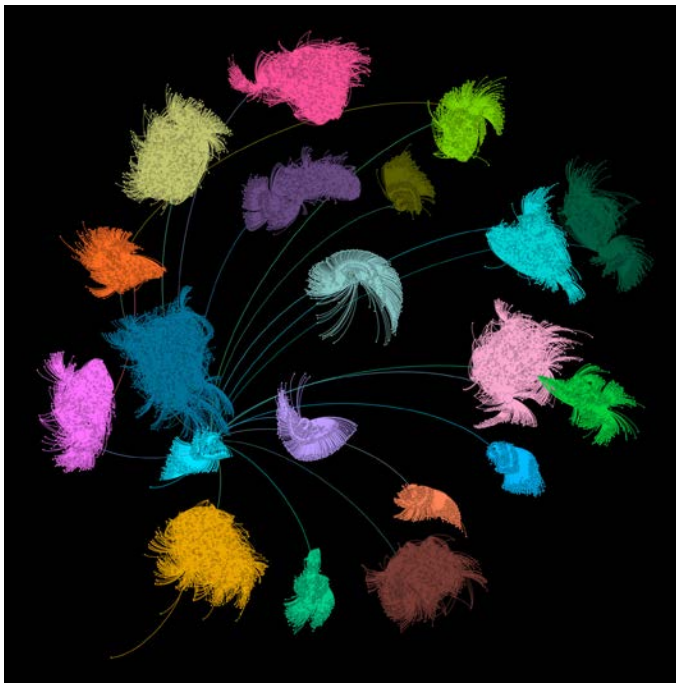
CURRENT PRACTICE

There is a lack of tools that can simulate extra-large NetFlow data, and there are even fewer tools that can simulate graph properties (e.g., degree distribution and clustering coefficient) in a large graph. In addition, only a small number of NetFlow data sets contain labeled malicious activity (e.g., botnets). We are aware of only one that includes botnets: the CTU-13 data set. Unfortunately, this data set has about 100,000 nodes, showcasing why larger NetFlow data sets are needed.

TECHNICAL APPROACH

In this task, we are developing a high-performance computing (HPC)-based simulation tool that can generate large-scale NetFlow data sets (~10^{10} nodes and ~10^{14} edges) that contain labeled malicious activity. This simulation tool will model NetFlow attributes (e.g., start and end time of flows), as well as topological

Develop an HPC-based simulation tool that can generate large-scale NetFlow data sets containing labeled malicious activity to test graph-analytic algorithms.



Large graph formed using an enterprise connection algorithm with the CTU-13 data set as the seed.

attributes (node degree, clustering coefficient), and will be implemented in Python using the mpi4py and NetworkX packages. In addition, we will develop a model for the spread of malicious activity (e.g., botnets within the graph).

We will simulate the extra-large NetFlow data based on a set of smaller real-world or simulated sample graphs. We will employ a multi-enterprise approach, where sample NetFlows are simulated at the same size with similar properties and added randomness. Then, these simulated NetFlows will be connected to form a larger NetFlow. Using this approach, we can ensure that characteristics of the sample NetFlows are preserved, while the NetFlow's size can be scaled up, limited only by the computing power.

This simulation tool is being developed on the Shadow II supercomputer at Mississippi State University.

Expected Accomplishments

This project is expected to produce a tool for generating large NetFlow data sets that are based on real data and contain labeled malicious activity. This tool will be expanded from previous work, adding other graph features, such as subgraph mixing rate, and betweenness centrality, as well as making it more customizable and user friendly. We will construct an intuitive user interface that will allow users to select the sample NetFlow data sets, the graph properties that need to be preserved, the resulting NetFlow data size, and the inter-enterprise connection topology.

IMPACT

The results of this project will facilitate expanded testing of graph algorithms among security researchers, and increase academic-governmental collaboration.

Contacts

Hugh Medal
Principal Investigator
(662)-325-3923
hmedal@ise.msstate.edu

David Dampier
Director, Distributed Analytics and Security Institute
(662) 325-0779
dampier@dasi.msstate.edu

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

MISSISSIPPI STATE UNIVERSITY