Project Title: Node Embeddings for Botnet Detection

Joseph Brinton: jcbrinton@gmail.com
Christopher Harker: chris.harker77@gmail.com
Jack Zhao: u1129777@utah.edu

Background

Botnets are a significant source of network attacks. They consist of several compromised computers capable of performing coordinated attacks, including DDoS attacks, click-fraud scams, spamming, and identity theft. A key component in preventing attacks from botnets requires methods to detect them.

Modern botnet detection methods still primarily rely on heuristics built upon the manual analysis of traffic patterns, malware code, etc. Machine learning approaches help automate the process, relieving some of the manual labor placed on analysts while showing promise in their ability to detect botnets. However, most machine learning algorithms require inputs defined in the euclidean domain. Therefore, when used on network data, machine learning algorithms fail to take into account crucial relational information between nodes in the network.

Proposed Solution

A new subfield of machine learning called Geometric Deep Learning attempts to address this concern by studying machine learning algorithms that can take non-euclidean data, such as graphs, as inputs. Geometric deep learning algorithms are usually divided into two categories, namely, graph embedding algorithms and graph neural networks. Graph embedding algorithms attempt to learn latent representations of the nodes in a graph, encoding relational information in a continuous vector space that can be used in traditional machine learning algorithms. Graph neural networks attempt to perform machine learning using the graph data structure itself as input.

Our project involves studying the node embeddings produced by some of the most commonly used graph embedding algorithms, such as DeepWalk and GraphSage. Botnets generally have different structural properties than background networks. For example, centralized botnets have a strong hierarchical structure, while decentralized botnets are designed for efficient diffusion of information. We hypothesize that since botnets might have different structural properties than other nodes in the network, their node embeddings should be different than nodes elsewhere in the network. If so, node embeddings can be a valuable tool used by analysts to quickly identify botnets with much less of the effort required by modern methods.

Implementation and Evaluation

Testing our hypothesis will consist of a few key steps: data processing, training, and evaluation. The data processing step involves writing scripts to process networks in CTU-13 dataset, preparing them for input into each of the graph embedding algorithms. The training phase involves writing scripts that implement and train the DeepWalk and GraphSage algorithms. Finally, the evaluation step consists of producing similarity metrics that will be used to compare embeddings of nodes inside and outside of the botnet, as well as visualizations of the embeddings that will assist us in our analysis.