Graph-Based Deep Learning for Fraud Detection in ETH Transaction Networks

Stephen Gelinas · sgelinas@ucsd.edu · Kazuma Yamamoto · kayamamo@ucsd.edu · Ethan Zhou · ezhou@ucsd.edu

Background & Research Question

- According to the FTC, cryptocurrency scams have cost online users over $1 Billion since 2021.
- With access to Ethereum transaction networks, we can model and train phishing detection as a node classification problem.
- How do non-graph supervised learning algorithms compare to graph-based deep learning approaches for fraud detection?

Why Graph?

- Graph algorithms can numerically represent information that is inherent to a network.
- Graph neural networks take advantage of learning the structural information within a graph and embedding information about neighboring nodes in the network.
- This allows for graph models to heavily outperform traditional learning algorithms.

Data Source

The XBlock dataset containing transactions of 890 Ethereum accounts

1. Collect subgraphs by K-order sampling with K-in = 1, K-out = 3 for each of the 890 objective nodes
2. Splice into a large-scale network with 86,623 nodes and 106,083 edges

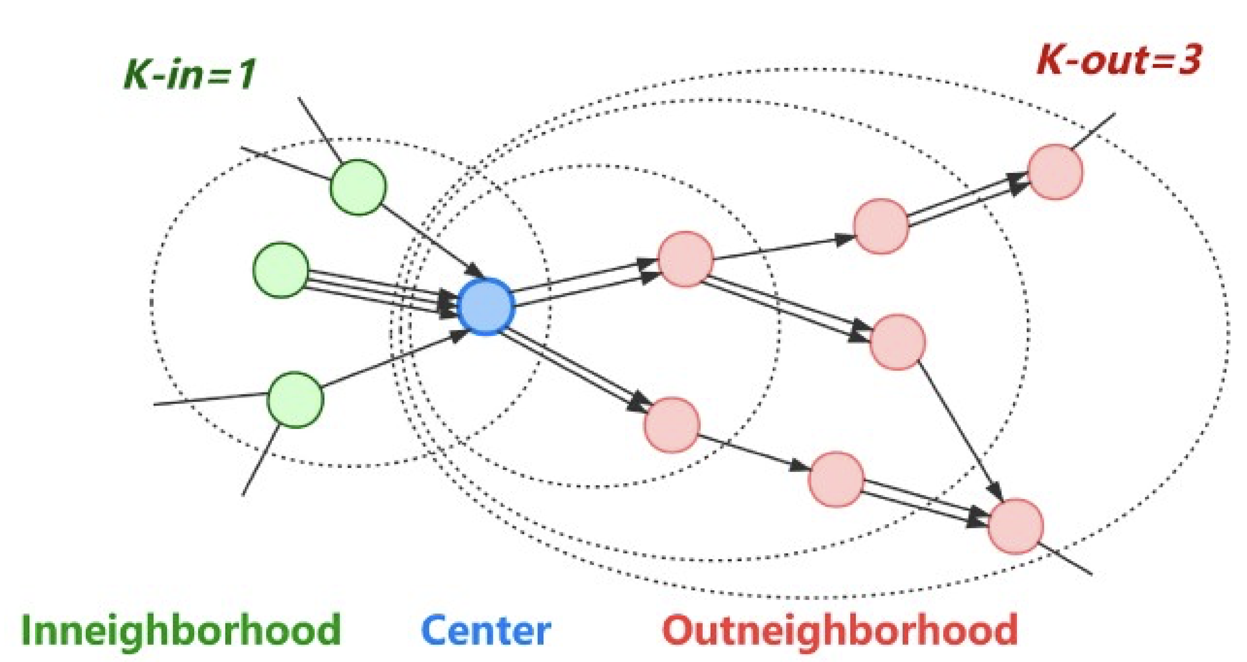

Figure 1: Schematic Illustration of a Directed K-Order Subgraph for Node Classification

On the figure above, based on the assumption that a typical money transfer flow is centered on a phishing node, the previous node of the phishing node may be a victim, and the next one to three nodes may be the bridge nodes with money laundering behaviors.

Figure 2: An example of how our graph schema looks, at the basic level

Model	Avg. Testing Accuracy	Type
TA-GCN	82.2	Graph
GraphSage	81.9	Graph
XGBoost	81.6	Tree
GCN	79.6	Graph
GAT	78.5	Graph
Node2Vec	76.6	Graph
k-NN	74.6	Traditional
SVM	60.5	Traditional

Table 1: Performance of each model

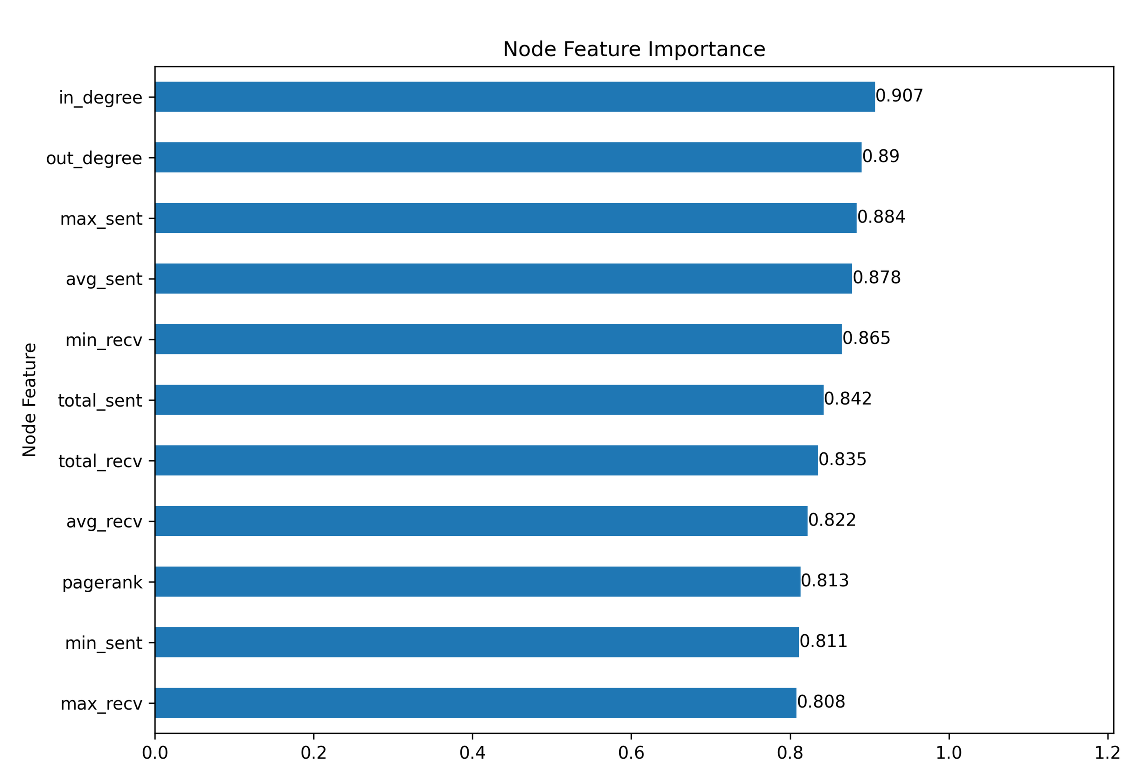

Figure 3: The relative importance of each feature to our models

Methods Overview

- The transaction network is represented as a directed graph, where each node represents a wallet in the network, and each directed edge between wallets represents the transfer of currency.
- Features are assigned to each node, including in-degree, out-degree, total transaction value, and more.
- The fraud detection task is ran as a node classification problem, and the performance of the models will be evaluated.

Conclusion and Summary of Findings

- Graph-based features improves overall model performance for both graph-based and non-graph-based models.
- Graph neural networks, specifically TA-GCN, performed upwards of 20% better in the fraud detection task, compared to traditional machine learning techniques.
- The most important features for predicting fraudulent wallets are in-degree and out-degree of the accounts.

Recent Advancements in the Field

- The field of graph data science is relatively new and very active, and key advancements happen very frequently
- For example, all of the models we used were developed in the last several years. GCN and N2V are from 2016, GAT, GraphSage, and TA-GCN are from 2017.

Future Works

- For further analysis we would like to explore more types of algorithms that require different data structures
- We would also like to try to combine some of the existing models to develop a model specialized for certain tasks, such as Ethereum
- For example, we could look deeper into why a tree-based model such as XGBoost is so adept at learning this data structure, and utilize it to improve accuracy.

QR Code for our github

QR Code for our Website