Project Proposal - COSC 421
Team: Chinmay Arvind, Jerry Fan, Dmitry Kostyukov, Rylan Millar

Members of group & division of work

Chinmay Arvind - Data cleaning and preparation, PCA Analysis, Similarity Metric creation, Network Metric Interpretation, Network Graph Creation, Research Questions 1 and 2, Report writing
Jerry Fan - Data cleaning and preparation, PCA Analysis, Similarity Metric creation, Metric Interpretation, Research Question 3, Report Writing
Dmitry Kostyukov - Data cleaning and preparation, PCA Analysis, Network Graph Creation, Similarity Metric creation, Research Question 4
Rylan Millar - Report writing, Similarity Metric Creation, Research Questions 2 and 4

Problem(s) being solved in this project/what we plan to achieve

Our project aims to provide a solution for bank account fraud detection that will find outliers in the dataset to separate non-fraudulent and fraudulent bank account applications. We plan to answer research questions that will be utilizing various network science metrics taught in class, such as Eigenvector centrality, network density, degree centrality, Katz centrality, closeness centrality, local clustering coefficient, cosine similarity, and Jaccard coefficient, and concepts such as strongly connected components and cores. The idea is to use the above-mentioned metrics along with the other columns in the dataset to classify nodes as either having a higher or lower risk of committing bank account fraud. The network science metrics interpreted in the context of bank fraud detection will provide a better idea of which nodes could be linked and commit bank fraud, potentially identifying criminal organizations that are participating in bank fraud. The other columns in the dataset will allow for creating edges between the nodes based on the similarity of their application details alongside the network metrics, allowing for a complex multigraph to be formed between nodes, providing a bigger picture of different facets of bank account applications to be considered when identifying fraudulent applications.

Research Questions our project will answer

Our project aims to answer the following 4 research questions:

1. Which were the key fraudulent players within the network?
2. Were there any specific fraudulent groups within the network that could be collaborating to defraud the bank? And if so, what were their characteristics?
3. What was the average profile of a fraudulent customer?
4. What differences exist between fraudulent account applications and non-fraudulent account applications?

Data Source

Our project will be using data from the following Kaggle dataset: https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022/data. There are 32 columns in the dataset, and 6 files of data on which we plan to perform Principal Component Analysis (PCA) on to determine the most important predictors of bank account fraud from the dataset, and use the identified most important columns along with computed network metrics as our final set of predictors to solve the above research questions, and attempt to provide an approximate method of bank account fraud estimation given a dataset with relevant predictors.

Definition of Nodes and Edges

The nodes in our network will represent the customers who make applications for accounts with a bank. The edges are connections between nodes (customers) based on the similarity of an attribute, such as: the time of the application, credit risk score, etc. Our graph will be a multigraph connecting nodes based on the similarity of attributes as mentioned above (the similarity score will be computed separately and will have to account for different types of columns in the dataset). We plan on coding the edges to hold information on what type of attribute 2 nodes are linked to each other based on (or color them differently and provide a legend for each different type of relationship based on the attribute) and weighting the edges based on the similarity score we compute for the 2 nodes. This will form a sufficiently complex network (which isn't disconnected into isolated pieces based on different attributes which would otherwise not have allowed for in-depth network analysis).

Metrics we plan to use

We plan to use the following metrics to give us a better idea of which customers are fraudulent, an average fraudulent customer's profile, comparisons between a non-fraudulent and fraudulent customer, and to determine any groups of bank account holders committing bank fraud in organizations:

1. Eigenvector centrality - to find pivotal fraudulent players in the network
2. Network density - to determine specific fraudulent groups within the network that could be collaborating to defraud the bank
3. Degree centrality - to find pivotal fraudulent players in the network
4. Katz centrality - to find pivotal fraudulent players in the network
5. Closeness centrality - to find pivotal fraudulent players in the network
6. PageRank centrality - to find pivotal fraudulent players in the network
7. Degree distribution - to find pivotal fraudulent players in the network
8. Local clustering coefficient - to determine specific fraudulent groups within the network
9. Cosine similarity - to determine the average profile of a fraudulent customer and find differences that exist between fraudulent account applications and non-fraudulent account applications
10. Jaccard coefficient - to determine the average profile of a fraudulent customer and find differences that exist between fraudulent account applications and non-fraudulent account applications

The above listed metrics are just a few we plan to use, we plan on adding a few more to the list of metrics that we will use as we continue development.

Analysis we plan to conduct

As mentioned above, we plan to conduct a thorough network analysis by computing the metrics mentioned above and using them to answer the respective research questions mentioned beside the metrics. The different centrality measures will provide us with a good idea of key players in the network, which we will then use to possibly find links between the players and the strength of their influence in the network. The clustering coefficient and network density will be used to find links between key players and possibly discover criminal organization structures. Cosine similarity and Jaccard coefficients will be used alongside descriptive statistics to determine average profiles of fraudulent customers. Centrality metrics, along with finding fraudulent groups determine key players with the highest centrality as the most influential players in the fraudulent applications subgraph. The concepts of cores, graph laplacian, modularity, and many other metrics will be used along with the clustering coefficient and network density to further solidify/falsify any hypotheses about criminals working together to defraud the bank, or to answer any of the remaining research questions above.