Identifying clusters using Node2Vec Embedding, Spectral, and GCN embeddings

Fraud Analytics Assignment 1

Sanket Deone
CS23MTECH11034

Sanket Rathod
CS23MTECH11033

Akshay Kumar
CS23MTECH11022

Abhishree Khangar
CS23MTECH11021

Arif Khan
CS23MTECH11024

Objective:

The objective is to apply various embedding and clustering techniques to effectively identify clusters and utilizing embedding and clustering techniques for fraud detection within a networked system. Extracting meaningful representations of nodes in the network that capture their behavioural patterns and interactions. The clustering algorithms help identify groups of nodes based on their embeddings, enabling the detection of anomalous clusters that may indicate fraudulent activity.

Problem Statement:

Given a Payment dataset with nodes and edges, the task is to identify clusters or communities within the network. This involves detecting groups of nodes that are densely connected internally while having fewer connections between groups.

Use embedding techniques Node2-Vec Embedding, Spectral, and GCN.

Describing The Dataset:

For the implementation and evaluation of all embedding techniques we have payments dataset.

Payments.csv:

- **Sender:** This column records the identifier (e.g., user ID or account number) of the party sending the money.
- **Receiver**: This column captures the identifier of the party receiving the money.
- **Amount**: This column lists the monetary value of each transaction.

With 130,536 rows, this dataset provides a comprehensive view of payment activities over a certain period, offering a rich source for analysing transaction patterns and relationships. Each row in the dataset represents a single transaction, indicating the sender, receiver, and the amount transferred. See details on fig.1.

Sender	Receiver	Amount
1309	1011	123051
1309	1011	118406
1309	1011	112456
1309	1011	120593
1309	1011	166396
1309	1011	177817
1309	1011	169351
1309	1011	181038
1309	1011	133216
1309	1011	133254

Fig 1. Payment's dataset

Algorithms Used and Results:

1. Node2Vec Embedding:

Node2Vec is a technique for learning continuous feature representations for nodes in networks, allowing for efficient exploration of the graph structure. It involves training a neural network to generate embeddings for nodes based on the network's topology.

- Generate random walks on the graph.
- Train a neural network (such as Word2Vec) to learn node embeddings from these random walks.
- Use these embeddings for downstream tasks like clustering.

In fig 2. Elbow Method is applied to identify number of clusters to use and fig 3. Shows result obtained by Node2vec Embedding.

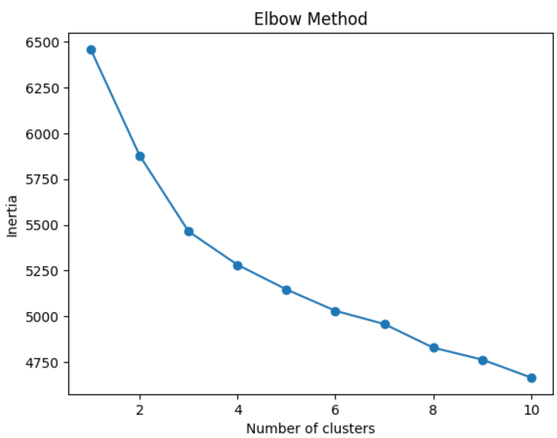

Fig 2. Elbow method for Node2Vec.

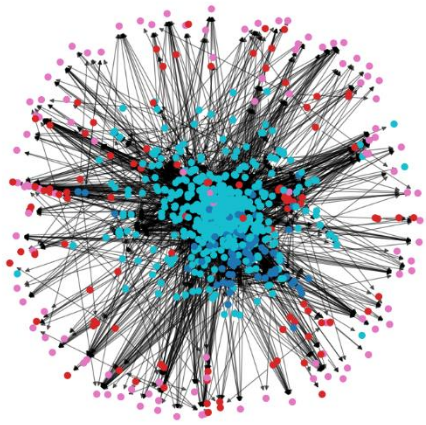

Fig 3. Clustering for Node2vec Embedding

2. Spectral Embedding:

Spectral clustering is a method for clustering data points by leveraging the eigenvalues of a similarity matrix derived from the graph Laplacian. It operates on the eigenvectors of the graph Laplacian, which capture the underlying structure of the graph.

- Construct a similarity matrix from the graph data.
- Compute the graph Laplacian matrix.
- Compute the eigenvectors corresponding to the smallest eigenvalues of the Laplacian.
- Use these eigenvectors to perform clustering using standard techniques like k-means.

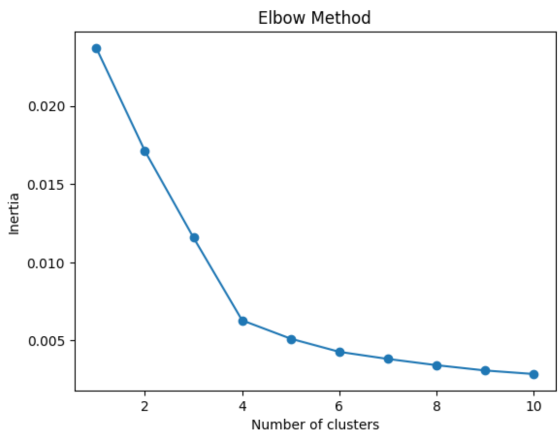

Fig 4. Elbow method for Spectral.

In fig 4. Elbow Method is applied to identify number of clusters to use and fig 5. Shows result obtained by Spectral Embedding.

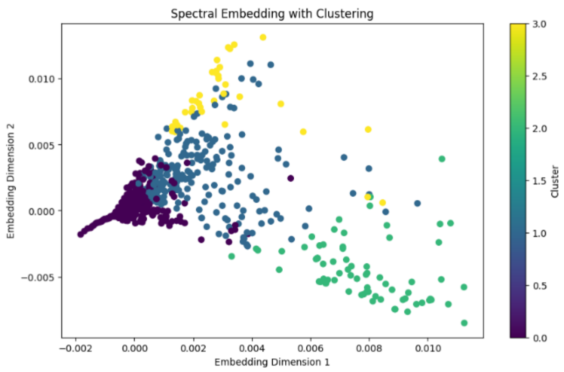

Fig 5. Clustering for Spectral Embedding(2D)

Visualization of Spectral Embedding clusters in different angles is given in fig 6. And fig 7.

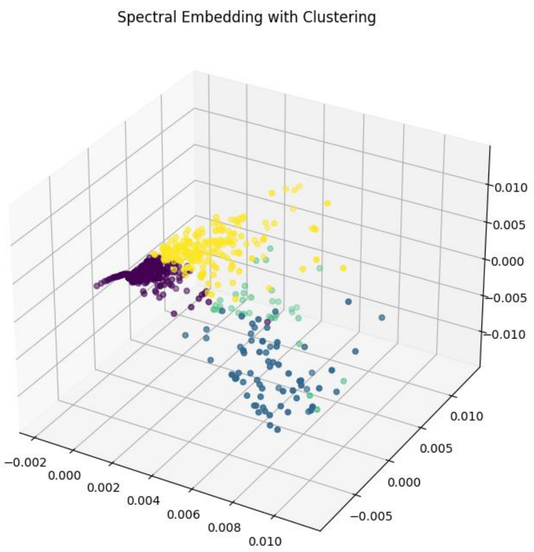

Fig 6. Clustering for Spectral Embedding(3D)

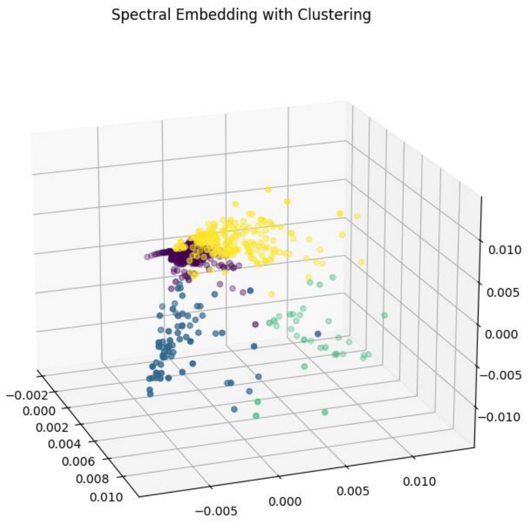

Fig 7. Clustering for Spectral Embedding(3D)

3. GCN Embedding:

GCNs are a class of neural networks designed to operate on graph-structured data. They leverage the graph's topology to perform convolutional operations on node features, allowing for effective feature learning and node classification.

- Define the graph structure and node features.
- Propagate information through the graph using graph convolutional layers.
- Incorporate pooling layers to aggregate information from neighbouring nodes.
- Use the final node representations for tasks such as clustering or classification.

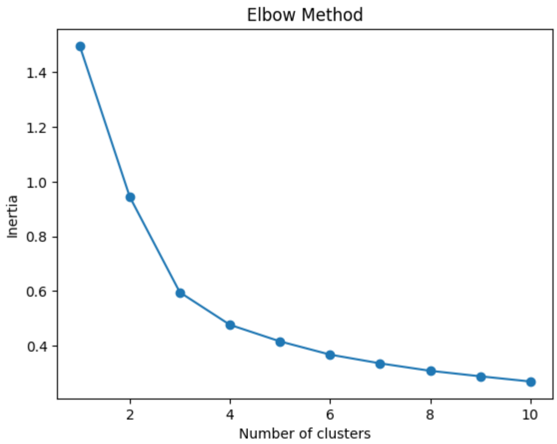

Fig 8. Elbow method for GCN.

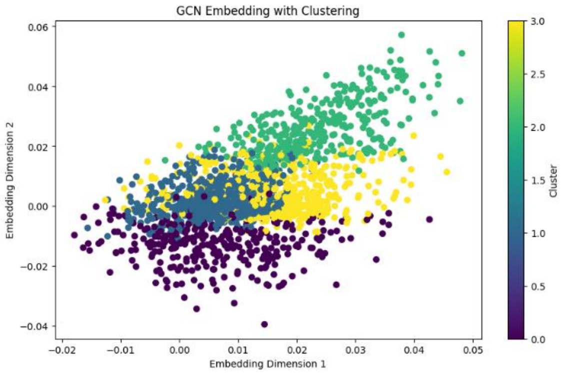

Fig 9. Elbow method for GCN.

fig 9. Shows result obtained by GCN Embedding in 2D space fig 10 and fig 11. Shows the result of GCN Embedding in 3D space with different angles of visualization.

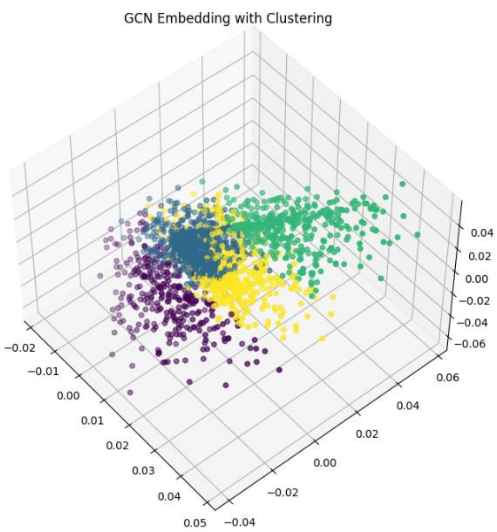

Fig 10. Clustering for GCN Embedding(3D)

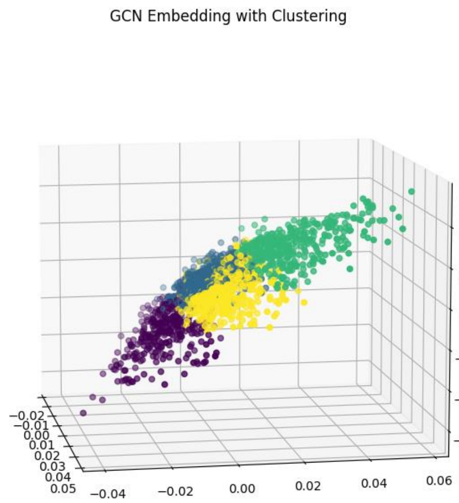

Fig 11. Clustering for GCN Embedding(3D)

Conclusion:

The Node2Vec embedding, Spectral clustering, and GCN embedding provides valuable insights into the financial interactions within the network, which can be leveraged for various analyses, including fraud detection, network modelling, and anomaly detection, ultimately contributing to a better understanding of the underlying dynamics and relationships.

Reference:

1) node2vec: Scalable Feature Learning for Networks. https://arxiv.org/pdf/1607.00653

2) Spectral embedding and the latent geometry of multipartite networks. https://arxiv.org/pdf/2202.03945

3) Knowledge Embedding Based Graph Convolutional Network. https://arxiv.org/pdf/2006.07331