Identifying Circular Trading in Transactional Data using Multi-stage Clustering

Nithin
CS16BTECH11005

Surya Pramod
ES16BTECH11015

Krishna Chaitanya
CS16BTECH11011

Overall Approach

Our approach uses a combination of several methods to identify small sized clusters that might be involved in circular trading. We represent the transaction as directed weighted graph with edges going from sellers to buyers and the transaction value representing the weight of the edges. We first preprocess this graph and then identify small clusters. Finally we check these clusters for cycles to identify circular trading.

Preprocessing

Benford Analysis

We use Benford analysis to see how likely the transactions captured in the given time frame are to be fraudulent. Depending on the value of Mean Absolute Deviation (MAD) from Benford's law we can check if the transactions captured in the corresponding time frame are illegitimate (i.e not naturally occurring).

Pruning the Graph

Before we start with clustering we prune the graph by only considering nodes that have both incoming and outgoing edges (i.e both buyers and sellers). This is because our goal is to identify circular trading and the nodes need to have both incoming and outgoing edges to be part of a cycle. Doing this step we can reduce the size of the transaction graph and our model can converge faster.

Graph Clustering

Clustering is done in two stages. First we cluster the graph using well known clustering algorithms. Since this might result in large clusters and the clusters that indicate circular trading tend to be of smaller size we need to further cluster them. To achieve these smaller clusters we use the method proposed in this paper by Klymko et al [1]

Clustering Stage 1

Shared NN or related methods (like Betweenness clustering, HCS clustering) will be used to form clusters in the weighted directed graph.

We plan to implement few of these clustering methods and then compare them to find the best method that fits our approach.

Clustering Stage 2

Since the sizes of clusters obtained from the above methods can greatly vary we might need to do further clustering. This clustering should be such that it considers small cycles to be important as they relate closely with presence of circular trading. To this end we use the method proposed in the paper (Using Triangles to Improve Community Detection in Directed Networks).

We decided to use this method on the already formed clusters instead of the original graph because this method does not take the weight of the edges in directed graph into account. But once we have the clusters we plan on ignoring the weights and just consider the edges. This converts our graph into an unweighted directed graph which is what the paper by Klymko et al uses to find clusters.

Evaluating the Clusters

Once we have the final clusters we plan on evaluating these clusters based on the size of the cycles present in them and their transactional cost.

Cycle Size

We find the length of the smallest cycle in each cluster. Since small cycles are good indicator of the presence of circular trading we can use the length of the smallest cycle in each cluster as a metric for evaluation.

Transactional Cost of Clusters

We also use the total cost of all the transactions within a cluster as an indicator to identify clusters involving heavy trading among themselves. Since circular trading involves multiple transactions within the same parties the transactional cost of the cluster is likely to be higher than non-fraudulent clusters.

[1] https://arxiv.org/pdf/1404.5874.pdf