Fraud Analytics (CS6890) : Detection of circular trade

Soumen Sarkar, Rushikesh K. Vaishnav, Priyabrata Jena, Venkatesh S. Mangale

CS22MTECH11011 *CS22MTECH11001* *CS22MTECH11016* *CS22MTECH14006*

Indian Institute of Technology, Hyderabad

Submitted: 4 May 2023

Abstract

In this report we have explained our approach to detect traders who are involved in circular trading. We have tried to solve this problem using a graph based approach. The graph we are using is a weighted directed graph. Where each nodes are traders and the edge between them is a real value which denotes the suspicion parameter. We are calculating the suspicion parameter based on a custom defined formula which take into account various different things such as the trading behaviour and amount of money transferred. After the weight assignment we obtain different cluster and by examining the cluster we come up with a list of traders who are highly suspicious.

Keywords : Circular trading, Graph based algorithm

1 Introduction

Circular trading is still one of the biggest concerns of investors and regulators in the capital markets. Sometimes traders artificially inflate sales and revenue figures by trading within each other. This shows up as heavy cash-flow making people believe that the companies are actually doing well on the stock market.

Although many efforts have been made to reduce the effects of such malpractice on finance ecosystem, the problem still persists. The biggest challenge in detecting circular trade is that there are lots of false positive as well.

1.1 Dataset Description

The data set we are using here is a csv file which contains **Buyer ID**, **Seller ID**, **Value**. All the values are integers. Once we obtain the graph embeddings are found using node2vec, we extract clusters from it using **DBSCAN**. Then we analyse the traders who are a part of these clusters by tracing them to the original directed graph.

2 Finding

Graph structure is an expressive data structure model which is able to examine relationships among objects in applications such as transport networks, biological networks, computer and telecommunication networks and social networks. Recently, graph based approaches and algorithms have been considered as a powerful tool to analyze financial networks like the network formed by investors' transactions (see Lillo and Valdés (2016)).

Moreover, it is possible to categorize all nodes to three groups such as fraudster, accomplice and normal (see Chau et al. (2006)[1]) and as fraudsters and their accomplices probably form a bipartite graph, a Markov random field model could be built from the transaction history amongst all traders, then the belief propagation algorithm is applied to calculate the probabilities of fraudster, accomplice and normal user for each node in the resulted graphs.

3 Related work

- Recently, graph based approaches and algorithms have been considered as a powerful tool to analyze financial networks like the network formed by investors' transactions (see Lillo and Valdés (2016)[2])

4 Algorithm

1. Create a weighted directed graph from the given dataset.
2. Merge all the multi-edges.
3. Using a ***custom defined function*** assign weights to the directed graph.
4. Generate embeddings for each node of the undirected graph using node2vec algorithm.
5. Apply DBSCAN algorithm to find clusters of nodes that are densely connected together.
6. Once the clusters are found iterate over each cluster and find nodes involved in it.

Github Repository

5 Custom Formula

$weight(a,b) = \sqrt{2^{C_2} \times 3^{C_3}} \times e^{cov(a,b)} \times zscore(a,b)$

- C_2 = Number of 2-cycle in which the nodes a and b are involved.
- C_3 = Number of 3-cycle in which the nodes a and b are involved.
- $con(a,b)$ = Max coefficient of variation
- $zscore(a,b)$ = Normalise maximum weight of the cycles in which a and b are involved

5.1 Explanation of $\sqrt{2^{C_2} \times 3^{C_3}}$

Number of 2-cycles and 3-cycles are important criteria as more number of such edges indicate a suspicious behaviour.

In order to create injective mapping we are using powers of prime $2^{C2}, 3^{C3}$ this creates an injective mapping, which is essential as we do not want to arrive at same value.

As compared to 2-cycles we are giving more weightage to 3-cycles, hence 3-cycles are raised to the power of 3.

Since the values raised to powers can be big we are trying to scale them down using square roor. The square root also ensures that the final value is greater than 0.

Even a single change in the count of 2 or 3 cycles will increase the final weight by a relative amount, which is desired.

5.2 Explanation of $cov(a,b)$

When the money is rotating in a circle we want to verify how closely they are related to each other.

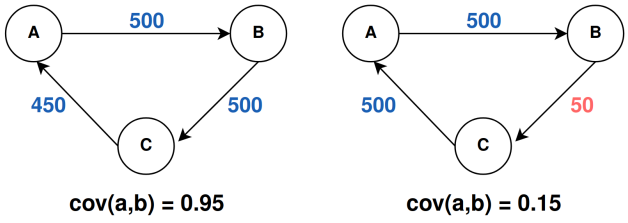

Figure 1: This figure explais that for any edge (a,b) the cov will give us a real value between 1-0 expressing the correllation between values.

Finding correlation is necessary step as it is an indicator of how suspicious the cycle is.

This is an important step as we want to remove false positive as much as we can. This part of the formula ensures even if the amount of trading is very high, the final weight will not grow too much as it will be reduced by the correlation factor.

Since all the correlation values are less than 1 we cannot directly multiply them in our formula.So we are scaling it up by raising it to the power of e which gives us a suitable multiplicative factor to be included in our formula.

This also works as a preventive measure for traders who might be unintentionally involved in circular trading.

5.3 Explanation of $zscore(a,b)$

Z-Score Normalization

Z-Score value is to understand how far the data point is from the mean. Technically, it measures the standard deviations below or above the mean.

$$zscore(value) = \frac{value - \mu}{\sigma}$$

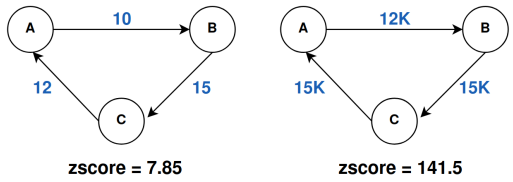

Figure 3: This diagram shows a measure of the maximum total money circulated in 3-cycle.

On our dataset we have already calculated the *mean* and *standard deviation* and we use the above formula to calculated the the scaled value.

5.4 Explanation of linear scale

We have created the formula in such a way that any minor change in any of the parameters of the formula will proportionally increase the weight by a smaller margin only. We have endured that all the scaling factors do not increase exponentially and after all the multiplicaiton the final weight should not be too much. During our testing on smaller custom graphs we have verified this.

6 Results

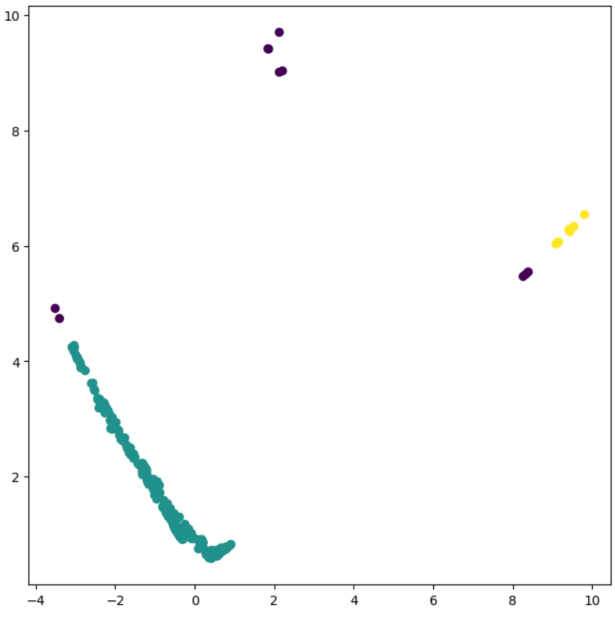

Figure 4: Clusters obtained after code execution.

7 Discussion and conclusion

In this assignment we have found circular traders who are highly likely to be involved in circular trade. Limited by the scope of the course we are performing analysis on a static dataset, where all the data is available to us beforehand.

Overall, graph-based approaches have the potential to be effective for detecting circular trading. By analyzing the structure of trade networks, these methods can identify suspicious activity and help prevent fraud.

As a future scope we would like to perform this analysis on a dynamic dataset where we would like the fraud detection algorithm to continually scan and provide alerts in realtime.

Overall by doing this assignment we gained a deep insight into the world of fraud analytic and new techniques that are being used to detect financial frauds.

References

[1] M. P. I. for Biological Cybernetics Spemannst, A tutorial on spectral clustering (2007).

[2] L. Anderloni and A. Tanda, Green energy companies: Stock performance and IPO returns, *Research in International Business and Finance* **39**, 546 (2017).