भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Collusion Set Detection using Graph Clustering

Fraud Analytics: Assignment 5

Jatin Sharma
CS17BTECH11020

Anjani Kumar
CS17BTECH11002

Dhananjay Raut
CS17BTECH11014

Vijay Tadikamalla
CS17BTECH11040

(Team size of 4 with special permission from sir)

Objective

The objective of this assignment is to identify mal-practices in stock market trading – e.g., circular trading and price manipulation by an algorithm based on graph clustering on the given dataset. The algorithm is able to identify collusion, i.e it is able to detect the set of traders "heavy trading" among themselves, as compared to their trading with others.

Introduction

Collusion is a very commonly occurring practice that happens every day in the real world. Especially in the area which is stock market trading. Some of the malpractices which occur are circular trading which is making money out of a cycle of trades, and price manipulation, when some trading is done for the specific purpose of manipulating the price artificially and non organically in order to change prices for self profit. This usually results in a loss to other well-meaning and honest traders. These malpractices use the modus operandi of collusion.

Many different types of mal-practices may happen in stock market trading. In this assignment, we ignore the malpractices related to payment (e.g., payment default), delivery of shares (e.g., delivery default), etc. We also ignore some trading-related mal-practices such as insider trading, takeover bids, market cornering, etc. Instead, we focus on specific trading-related mal-practices such as circular trading, price manipulation, price hammering, price propping, etc.

In price manipulation, a group of individual traders tries to act together to artificially attempt to increase the price of a security. To achieve this, the traders in the group circulate a fixed number of shares among themselves in a large number of trades; they keep increasing the price in these trades, thereby forcing an increasing trend in the price as well as interesting other traders. When the overall trading price rises sufficiently, the traders in the group "exit" by selling their original shares. Since the price rise was not tenable, the price crashes back to its original level or below.

All these malpractices are generally referred to as collusion-based malpractices because all these mal-practices involve a group of traders acting and trading together to achieve the specific effect on the price or volume of target security.

In collusion-based malpractices, the individual trading transactions are usually superficially legal; the mal-practices are visible when the transactions are properly grouped together. Evidence for such malpractices is often hidden deep inside trading databases. Hence, surveillance of stock market databases for detecting collusion-based mal-practices is an important, knowledge-intensive & complex, task.

In this report, we are using graph clustering techniques for efficient detection of candidate collusion sets. Apart from efficiency, this approach has a major advantage that the user does not supply much knowledge; e.g., there is no need to specify the number of candidate collusion sets nor is there any need to quantify the notion of heavy trading.

Dataset

The dataset we got for the assignment consists of trade statistics containing three columns, Seller ID, Buyer ID, and amount. The dataset contains details of approximately 24000 transactions between 18000 unique traders.

	sid	bid	amount
0	6941	707	84
1	17371	707	27
2	18216	707	5
3	76646	707	29
4	78095	707	56

Figure 1: Sample data points from the dataset

We use the dataset to construct a Stock Flow Graph and apply an unsupervised graph clustering technique to make clusters for identifying traders practicing illegal methods like circular trading, etc.

Due to the high time complexity of the algorithm and huge dataset size, we sampled a small part of the dataset for testing our algorithm correctness.

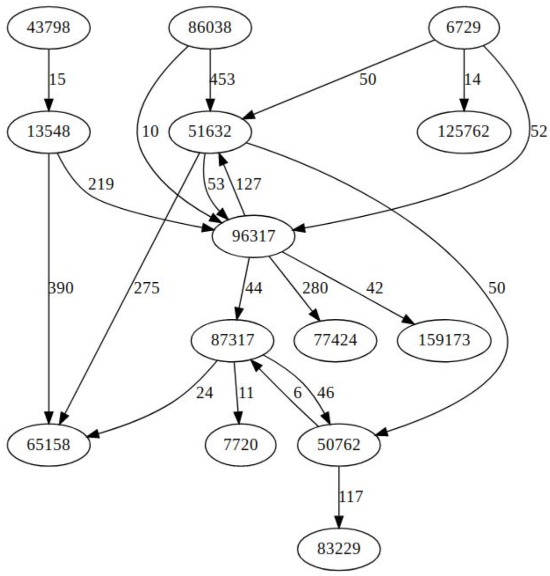

Figure 2: Small sample of the dataset.

Collusion Clustering

Collusion Clustering is an unsupervised algorithm, specifically tailored for the problem of detection of candidate collusion set. The algorithm has both advantages and disadvantages.

Advantages:

- No need to specify the number of candidate collusion sets
- No need to quantify the notion of heavy trading.

Disadvantages:

- The algorithm has a high time complexity $O(kn^3)$. Hence it is not very efficient at detecting malicious practices in huge datasets.
- The algorithm works only for transactions in a given time period. Hence, it cannot be used for real-time collusion detection.

Procedure:

In this algorithm, we make use of collusion index, $\phi(C)$ where C is the set of traders.

Definition: $\phi(C)$ for a set C of traders is defined as the ratio of the amount of internal trading to the amount of external trading.

Hence, a higher collusion index implies higher probability of traders involved in malpractices.

Results

The Collusion Clustering algorithm has three hyperparameters k, m, h and as the algorithm is very sensitive to the value of m, we fix it to one.
So, we tested the algorithm for various values of k, h.

Table 1: Clusters formed on the sampled dataset with (m = 1).

k	h	Some Clusters
3	0.6	[87536, 113914, 218343], [40411, 129313, 2311711], [113728, 22059, 27700, 3042] [9310, 33449, 636911], [20378, 25315, 138414]
3	0.7	[87536, 113914, 218343], [113728, 22059, 27700, 3042] [10942, 11166, 16310], [11599, 20378, 25315, 138414], [9310, 33449, 636911],
4	0.6	[87536, 113914, 218343], [40411, 129313, 2311711], [9310, 22824, 33449, 40115 636911], [4000,11599, 20378, 25315, 138414],
4	0.7	[87536, 113914, 218343], [40411, 129313, 2311711], [136850, 2309051, 2466551], [9310, 22824, 33449, 40115 636911]

Since traders like [87536, 113914, 218343] and [9310, 33449, 636911] always occur in the same cluster for different values of k and h, we suggest that they are a candidate collusion set.

Conclusion

In this report, we applied the graph clustering algorithm for collusion set detection. The problem of detecting colluding traders is important because many mal-practices in stock market trading – e.g., circular trading and price manipulation – use the modus operandi of collusion.
The most important advantage of this algorithm is that the users do not have to specify any background knowledge, training examples, etc. (e.g., there is no need to specify what constitutes heavy trading). The algorithm is adaptive in the sense that they can be used without change for different securities having vastly different prices and trading ranges. The reasons are in the basic philosophy of these algorithms: which is to rank the vertices (or traders) rather than use numerical distance in terms of trading volume etc.

References

- G.K. Palshikar, M.M. Apte Collusion set detection using graph clustering