Finding similar stocks by detecting cliques in market graphs

Sudhashree Sayenju
Department of Statistics and Analytics
Advisor: Dr. Joseph DeMaio

Welcome!

Please stand by. This event will start at 3:45 PM.

Please note the following :

- Please mute your microphone once entered into the meeting.
- Once the presenter starts to talk, please “pin” the presenter so they are the main focus of your screen

24th Annual
SYMPOSIUM OF STUDENT SCHOLARS
VIRTUAL EDITION
APRIL 16, 2020

Recognizing excellence in student scholarship and creative activity

KENNESAW STATE UNIVERSITY

Finding similar stocks by detecting cliques in market graphs

KENNESAW STATE UNIVERSITY

Sudhashree Sayenju

Advisor: Dr. Joseph DeMaio

Department of Statistics and Analytics

ABSTRACT

The stock market provides an abundant source of data. However, when the amount raw data becomes overwhelming it grows increasingly difficult to know how the stocks interact with each other. Stock data visualization as a market graph serves as one of the most popular way of summarizing important information. When modeling the data as a graph, vertices correspond to stocks and edges correspond to strong correlation in their pricing in a certain period of time. This project presents a technique to find stocks that behave very similarly. Such information helps investors make decisions on which stocks to purchase next. The investors can utilize this information to select a valuable portfolio of stocks showing an increasing price trend. On the other hand, it can also help stock owners to make decisions on whether or not they should sell their stocks.

```
sudhashree@sudhashree-sayenju:~/Documents/Graph_theory/graph_theory_project$ python3 graph_theory_project.py
The largest cliques are :  [('HBANP', 'MYLNP', 'CME'), ('HBANP', 'ATRI', 'PCLN')]
The clique number omega(G) =  3
The edge weights in maximum cliques :  [[1.3, 1.2, 1.2], [1.4, 1.5, 1.3]]
The variance of edge weights on each clique is :  [0.0022222222222222266, 0.006666666666666664]
```

Figure 1: Console output after the program has finished running.

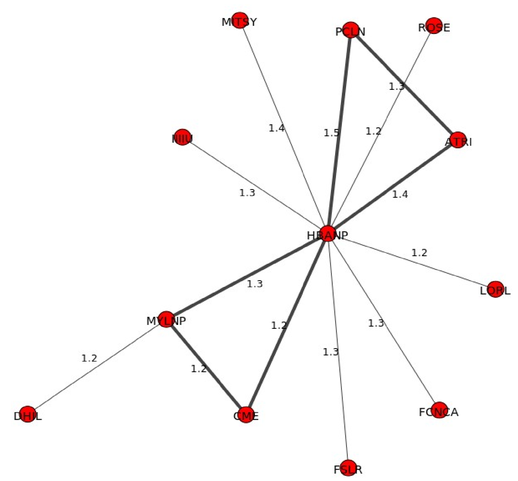

Figure 2: The graph image (left) of stocks with cliques marked with dark edges on them after the program has finished running.

PROBLEM STATEMENT

1) Historical market data is a graph (G = (V, E)). For each pair of vertices (u, v ∈ V) we first calculate the correlation coefficient r(u, v).Vertices represent stocks and edges appear between the vertices only if the correlation is larger than a specified threshold c ∈ [−1, 1].
∀u, v ∈ V, r(u, v) > c ⇒ {u, v} ∈ E

2) In G include weights for each edge. We take the correlation coefficient r(u, v) to be the weight of each edge {u, v}. $\mathbf{w_{uv} = r(u, v)}$

3) Find all possible maximum cliques (complete subgraphs) and clique number ω(G) in the graph. The cliques will give us groups of stocks that behave most similarly.

4)Find the variance of edge weights in each maximum clique.

METHOD

At first we read the given text file with stock market information as a graph. In the data there is information about various price fluctuations of each stock like opening price, closing price, highest price and lowest price. Each stock represents a vertex v ∈ V in the graph G = (V, E). The implementation of my program was done in Python language and the graph simulation was done with the help of *igraph* package.

To draw edges, we first need to know whether or not the stocks are strongly correlated to each other. For this, we calculate the pairwise correlation r(u, v) between all the pairs of stocks. We set a manual correlation threshold c ∈ [−1, 1]. We first filter the pairs of stocks which have higher correlation than the given threshold c. The graph will only contain those stocks and pairwise edges. Moreover, we also include edge weights which is the correlation between those stocks.

After converting the data into vertices and edges, the next task is to find the maximum cliques. With the help of some functions in *igraph* package we can determine which set of vertices form the maximal clique. With this we determine the clique number ω(G). The edges in the cliques are plotted dark in color in the graph. There might be multiple maximal cliques and each clique represents a groups of stocks that behave most similarly. To find which clique is more tightly correlated we can check which clique has lower variance for their edge weights.

CONCLUSION

This project presents a way to discover the most interesting information from raw stock market data like finding cliques with the strongest correlation. If an investor is looking to invest in a portfolio of stocks which show increasing trend, finding maximal cliques with increasing price trend would make a good choice. Furthermore, the findings of this project can also help stock owners on whether or not they should sell their stocks if they are in cliques showing decrease in price.

References: Eoddata. Historical price data. http://www.eoddata.com/products/historicaldata.aspx, Vladimir Boginski, Sergiy Butenko, Oleg Shirokikh, Svyatoslav Trukhanov, and Jaime Gil-Lafuente. A network-based data mining approach to portfolio selection via weighted clique relaxations. Annals OR, 216(1):23–34, 2014.