Project Proposal
Examining the Correlations between Price Changes of Different Stocks

Deliar Mohammadi, Soroosh Esmaeilian

1. Nodes and Links

The nodes in our graph will be stocks; these stocks are mostly US and Canadian companies listed on major stock exchanges. The links of our graph will represent the correlation (similarity in percentage change) between the price of any two stocks. We plan to have a minimum threshold for the correlation percentage. This way, we can avoid having **N(N-1) / 2** links and keep only the meaningful links.

2. Data

The data contains roughly 3000 nodes (stocks) and these were obtained from the yahoo Finance API. Scripts were generated and run by ourselves to collect the monthly price data for the ~3000 stock symbols. Mining and cleaning the data took 2 weeks from start to finish, one thing to note is that all of this data is completely free and open source. Additionally, the raw data contains the prices of the stocks themselves, in order for us to build out our network we need to generate correlation values from each stock to every other stock (excluding itself).

3. Expected size of the network

We expect our network to contain anywhere from 2900-3500 nodes; For a stock to qualify as a node, it must have information for every single month from January of 2010 to December of 2021.

For each node, we will be calculating its correlation to every other node in terms of the percentage change. So technically, we will have the maximum number of links, N(N-1)/2, and we will then subtract nodes that don't meet the threshold requirements.

4. Questions we plan to ask and why we care

The network we plan to build can give us profound insight into the relationships between many different stocks. Moreover, it can serve as a benchmark for predicting future prices of stocks based on the change in the price of those stocks that have a strong correlation (link) to it.

Based on what we know about markets and pricing, we have several predictions about what we expect to see in the graph, one of which is a very strong correlation between oil stocks and all other goods that require shipping. Because oil and gas are used in transporting almost everything through cargo ships or diesel-powered trains, it would make sense that it has a strong negative correlation with other goods. Logically, if the stock price of oil & gas goes up, the price of shipping increases, and in turn, the price of goods are increased to maintain profit margins, which drives away sales and results in a downturn in stock performance. Additionally we expect to see strong correlation between competitors that are reaching out to the same market, such as Uber and Lyft.

We will be using these predictions to evaluate the results that we get. Since we have a few expectations and trends that we should see theoretically, we can compare this to our real outcome and use it as a reference point for the network's success.

We have several questions regarding the relationship between the price of goods, and hopefully the graph can give us some insight on these:

1. *How correlated are oil stock prices and other products?*
    a. **Visually**, we can check for this by looking at the network and seeing how many connections are made to the oil & gas sector.
    b. This can also be measured by getting the **weighted degree** of oil & gas stock nodes, which is expected to be higher than the average of the network.

2. Are there any nodes in the graph that have a large effect on other nodes? Put another way, Are there any stocks that have a large reach in terms of affecting the price of other stocks?
    a. Again, this can be checked visually and by calculating weighted degrees of all nodes and comparing them to the mean, looking for nodes who lie above the mean.
3. Are there any nodes that have very small linkage from/to other nodes?
    a. This can be checked in a similar fashion to Number (2), however this time we are looking for nodes who lie far below the mean weighted degree
4. How strong are the intra-sector relationships of the nodes?
    a. This can be checked visually by looking at the **communities** in our network and ensuring the links between members of the community have large weights.
    b. We hope to separate these communities by the sector and by using a built-in community detection algorithm, to find groups of like nodes.

We will be defining correlation as the similarity in percentage change of a stocks price, over the entire duration of the data. There are two general formulas we would like to use to calculate correlation they are as follows (in general form for x, y, where x and y are percent changes of some stock):

$$corr = min(abs(x - y)/x \, , abs(y - x)/y)$$

$$corr = min(abs(x /y), \, abs(y/x))$$

We would like to use both these calculations, and see which is a better match for our data.

Adding onto this, we would like to calculate correlation using different time periods, for instance the 1 year correlation and 5 year correlation could yield different values. Doing this would allow us to show the changes of the graph weights over time by animating through several time slices in which the correlations vary.