Project Proposal
Examining the Correlations between Price Changes of Different Commodities

Deliar Mohammadi, Soroosh Esmaeilian

1. Nodes and Links

The nodes in our graph will be commodities; these come from various sectors, including oil & gas, the automotive industry, grocery products, and so on.
The links of our graph will represent the correlation (similarity in percentage change) between any two commodities. 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

Our dataset contains the prices of a multitude of commodities, from everyday grocery items to cars, plane tickets, and many more. These prices date back to January 2010 and span until January 2022. To gather the data, we are downloading multiple spreadsheets from many sources, such as the government of Canada, BMW, WestJet, Toyota, H&M, Shell, and many other sources that have historical price information on commodities. This data will then be normalized into a master spreadsheet; this will take the longest time because almost all the data available is in wildly different formats. Apart from formatting, we also need to ensure all prices are in $CAD, so in some instances, we may have to convert the currency based on historical conversion rates. Another part of the normalization stage is categorizing all the nodes (commodities) that we collect into sectors. Some example sectors are automotive, grocery, oil & gas, aviation, and fashion. The purpose of splitting these commodities into sectors is so that we can group them by sector when building our network. This way, we can identify intra-sector relationships and also inter-sector relationships. Given the complexity of obtaining and normalizing the data, we expect most of our time to be spent on this phase.

3. Expected size of the network

We expect our network to contain anywhere from 1500-2500 nodes; the reason we won't be able to expand beyond this is simply lack of data. There are many products/commodities that have price data only for 1-3 months; this is simply not enough as our data spans ~144 months. For data to qualify as a node, it must have a significant amount of information, which we specified as **more than half of the months (>72).**

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 nodes 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 goods. Moreover, it can serve as a benchmark for predicting future prices of a good based on the change in the price of those goods 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 prices and all other goods. Because oil and gas are used in transporting almost everything through cargo ships or diesel-powered trains, it would make sense that it strongly correlates with other goods. Because if the price of oil & gas goes up, so does the price of shipping goods, and in turn, the price of the goods is increased to maintain profit margins.

We will also 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 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 nodes, which is expected to be much 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 commodities that have a large reach in terms of affecting the price of other commodities?
    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 goods that have very little to no effect from 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 commodities 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 commodity):

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

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

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.