Jacob DeRosa: 50464566

1 Project Proposal

For my project, I want to study the structure and behavior of markets. Essentially, I aim to analyze the influence items within a market upon one another and determine the underlying structure or (most influential) parts. I also aim to identify and analyze groups of items that are heavily dependent on one another. The market data is inherently temporal because as time changes demand changes and environmental factors influence the market. I plan to analyze a single timeframe of about 1 year, for the sake of simplicity, but the analysis applied over multiple windows would yield more interesting conclusions. The network is defined where nodes are items within the market and edges are defined by some pairwise similarity metric I will describe, but only nodes with similarity above the median will have an edge between them. The most accessible data is through the stock market, which is what I will be using, but we could easily swap it out for currencies, resources such as gas, electricity, other internationally traded goods or even the Steam community market owned by Valve. I do not plan on using every single stock, but I think I can find something worth studying with roughly 1000 different stocks, these can be found with a simple google search so I'm not including sources for it. Most likely I will aggregate several Kaggle databases for my study that I will cite.

To define similarity, I am going to apply a pairwise Pearson Correlation which is basically a normalized version of covariance to the items. This will allow me to capture linear dependencies between items. It will not capture non-linear relationships which are more nuanced and complicated, but this is fine for my study. The 1-year window of data will be sectioned off and Pearson correlation will be applied to each pair of items for each section creating a series of similarity plots. I can dynamically threshold the items with the highest similarity by picking only relationships stronger than the median. The series of plots is chosen over a simple numerical value because over time some items will vary in their dependence on one another, and this will allow for the older stronger relationships to become less valuable if the items grow to be less dependent in recent sections.

For my analysis, I want to group the items together to expose market sectors that are highly dependent and visualize the dependence between market sectors. I also wish to define the most influential items within the whole market as well as items that are the most influential within a market sector. One thing we did not cover in class that I want to do is find the minimum spanning tree of the market so that we can see the (skeleton) that is underlying the market where these items have the strongest relationships and will reveal conclusions about the robustness and

resilience of the network. One thing that also really interests me is to study the time delay between market sectors. For example, if one sector crashes how long does it take its close neighbor to feel the crash too. This is highly valuable for making predictions for the future. This would involve doing more analysis across time frames rather than examining a single one.

The software I plan to use will be Python as it's my favorite and most familiar method of practicing data science. It offers many extremely useful libraries. I plan to use numpy, pandas, matplotlib and networkx to perform my analysis.

Here is a sample of the network I am proposing to study which features roughly 30 items. The actual project will have about 30 times as many nodes.

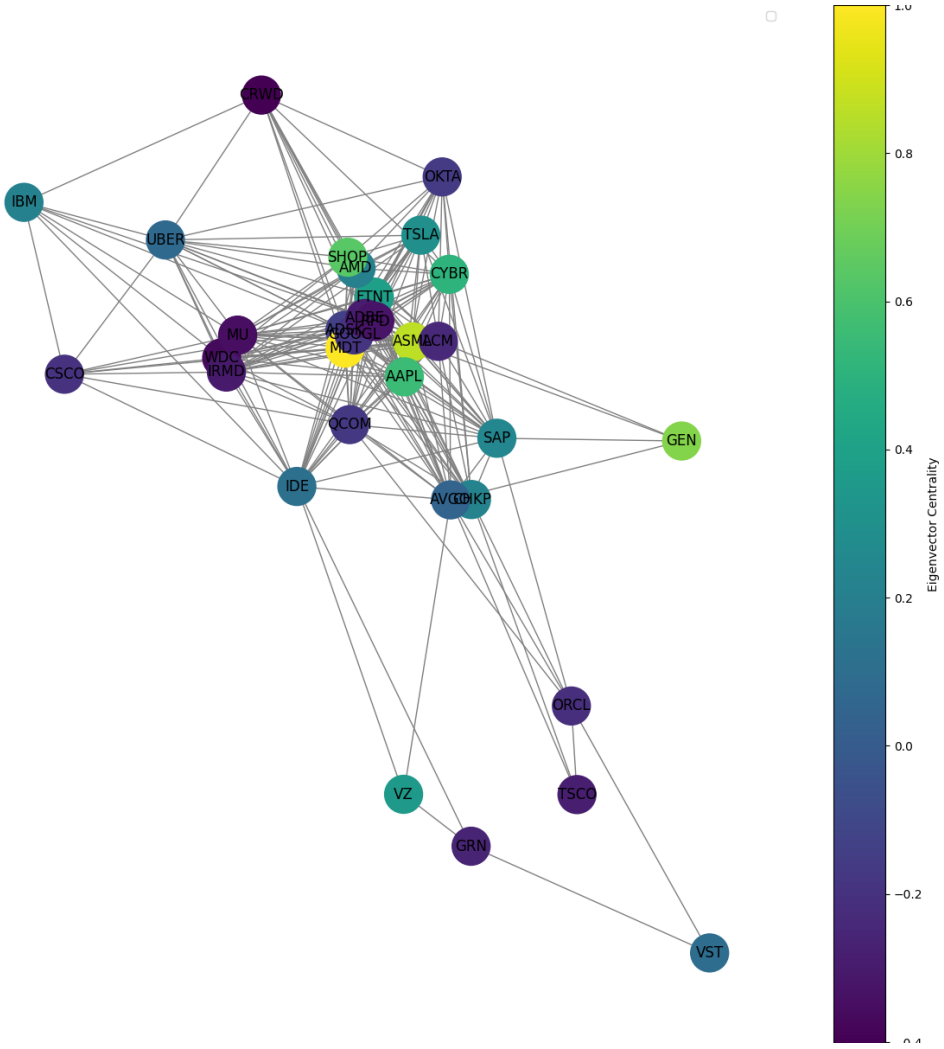

Figure 1: Eigenvector Centrality applied to a subset of the stock market (32 nodes)

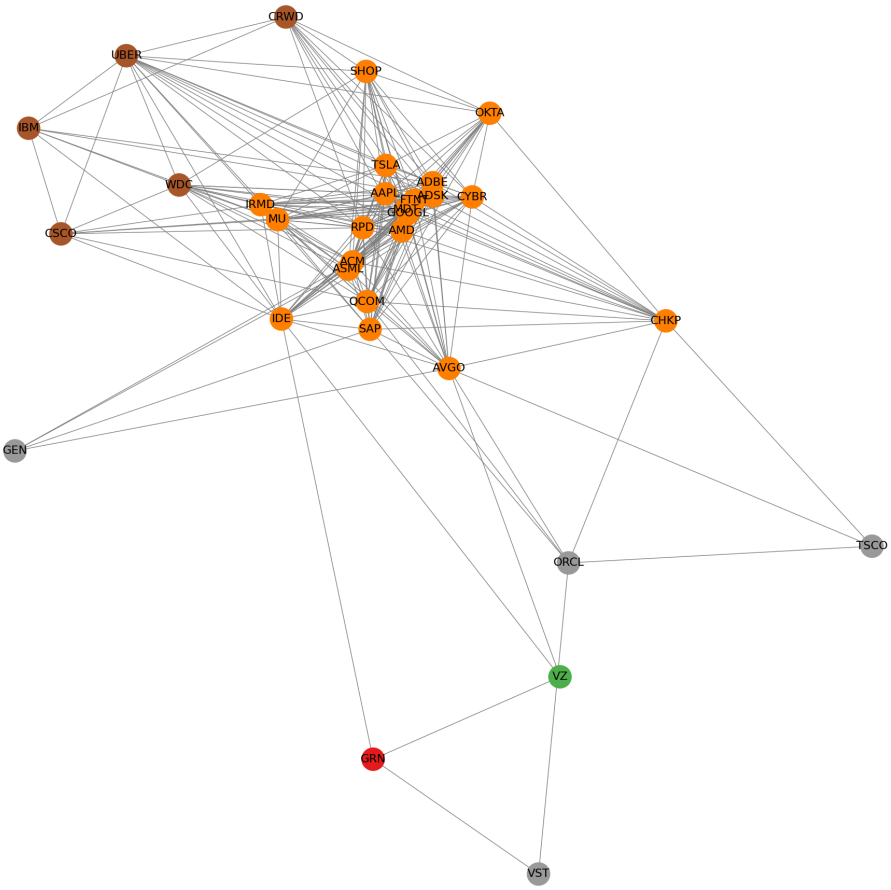

Figure 2: Spectral Clustering applied to a subset of the stock market (32 nodes)