Network Analysis of the Stock Correlations in the S&P500 Index

Hayden Church
University of Calgary
UCID: 30133957
CPSC 572

Khushi Bhatt
University of Calgary
UCID: 30068359
CPSC 572

Abstract

The S&P500 market index is a major indicator for the state of the economy, consisting of the approximately 500 large market capitalization stocks listed on US stock exchanges. Understanding the relationships between stock prices and their movements is a key part of predicting future financial trends and for effective portfolio management.
In this study, we aimed to study the interconnections and evolving dynamics between both stocks and sectors in the index through a network science lens, focusing on the pre-, during, and post-COVID-19 pandemic timeframes.
Our main goal is to improve portfolio recommendations by uncovering new insights around these network structures. The financial network we are modeling in this study revolves around representing individual stocks as nodes and constructing links based on the Spearman correlation values between pairs of nodes.
By modeling these networks over a period of time, we are able to capture the evolving relationships among these stocks, taking into account market shocks such as the pandemic.
Through applying various network science metrics and creating visualizations, we are able to quantify these changing relationships and analyze their implications for portfolio allocation. Our analysis reveals the following points:(1) the structure of the S&P500 is vulnerable to both smaller and larger market changes, but has a tendency to revert back to its preferred state (2) an investor can create a relatively risk-free portfolio by studying partitions formed by sectors, and (3) that the Industrials sector has the most influence over the network as a whole.
Future research could explore incorporating machine learning algorithms to enhance the predictive strength of the models and involve causality analysis as well for a more comprehensive examination. Overall, this study emphasizes the benefits of and improvements made by analyzing complex financial structures through the application of network science.

1 Research Questions

Our study endeavored to answer the following research questions, which we refocused to questions that could be answered with our current methodology of studying the stock price correlations within the S&P500 index:

1. How do stock price correlations (between and within sectors) evolve over the pre, during, and post-COVID-19 pandemic periods, and what are the underlying structural changes within the S&P500 network?

2. How can an investor create a diversified, relatively risk-free portfolio as best as possible by looking at the correlation trends between sectors?

3. How can we predict the influence of nodes and sectors, and in turn, how does that help with predicting future performance for investing purposes?

We aimed to answer these questions through the application of network science methodologies and an

empirical analysis. To address question one, we created a series of static networks to represent the varying time periods we are studying, and applied basic network science methods to study the similarities and differences over time. For question two, we studied the GICS sectors[9] in the S&P500 at an individual level, created partitions based on the Louvain algorithm[2] and studied how a stock investor could create a relatively risk-free portfolio by smartly diversifying. Finally, for question three, we applied the Katz centrality[12] method in order to identify which stocks are consistently the most influential in the network, from which we can garner insights into which investment sectors and stocks should be avoided by a risk-averse investor.

2 Introduction

The convergence of network science and the study of financial systems is not a novel concept, especially in the pursuit of garnering fresh insights on predicting future performance and portfolio management. This project builds upon existing literature on the subject with a unique focus on analyzing the time period before, during, and up to three years after the onset of the COVID-19 pandemic, with an emphasis on sector relationships in order to capture the more nuanced relationships among a broad range of asset classes. We have also integrated network science metrics such as Katz centrality in order to assess and rank stocks by influence/importance, and the Louvain Method for community detection.

Previous studies have explored various aspects of the S&P500's dynamics, from different perspectives, and with varying methodologies. For example, Kim & Sayama (2017)[5] employed a statistical model (ARIMA) on their network, which is a common traditional way for financial analysts to predict future performance, but not a network science specific methodology. Additional studies, such as Miskiewicz & Bonarska-Kujawa (2021)[8], also explored the COVID-19 angle, albeit only during the different stages of the pandemic which would not allow for long-term analysis of the effects of that crisis on the overall network's structure. Other studies, like Arai, Yoshikawa, & Iyetomi (2015)[13], also focused on the correlation aspect of the analysis, and applied methods such as Complex Principal Component Analysis and Random Matrix Theory, but did not delve into the future implications of their findings. On the other hand, Durcheva & Tsankov (2021)[4] chose to focus on a time span of 15 years, up until a few months into the pandemic, focusing on Granger causality and studying how the topology of the network transformed over time. The paper by Lee & Woo (2019)[6] uses a very similar method to ours, but with a focus on comparing different global market indexes such as the S&P500 and the Korean stock market index. They focused on analyzing cohesion due to correlation and Granger causality in bearish vs. bullish markets, and creating a portfolio recommendation system based on this information.

In contrast to existing literature, our study aims to bridge some of the gaps listed by adopting a more holistic approach to our analysis. In constructing a series of time-varying static networks, as well as the networks over the entire time period, applying correlation analysis, and studying the implications of our insights on future performance and portfolio management strategies.
We aim to provide a more comprehensive understanding of how the dynamics of the S&P500 network are affected by major crises over a period of several years. While existing literature provided us valuable insights and inspiration for our project, we have contributed further by approaching the analysis from a different angle to gain new insights. We also provided a series of portfolio recommendations to demonstrate the predictive power of our approach and its applicability to real-world investment scenarios. Thus, our study not only builds upon existing knowledge, but also offers a unique synthesis and analysis.

3 Dataset

Our raw data was retrieved from multiple sources. The stock data over time including prices, ticker sym-

bols, and dates were retrieved from Yahoo Finance historical stock data pages, which is a reliable and reputable source for financial data. The list of GICS sectors and company names currently listed on the index was retrieved from Wikipedia[3], with multiple sources on that webpage to verify correctness. The data from Yahoo Finance was retrieved via an open–source API called yfinance[1]. The data from Wikipedia was obtained through web scraping via Pandas in Python.

Prior to getting the data in a network format, it was essential to clean and merge the different data sets, and get them in a single CSV format that we required. The required columns include the ticker symbol, company name, GICS sector[9], date, and close price (daily), over the required period of time in chronological order (Jan 1, 2019 - Dec 31st, 2023). We also found it necessary to handle missing values for certain dates in the timeframe. This manipulation was also done via Pandas, and ultimately outputted a full CSV file of all our data, which we used for further manipulation.

Our basic network construction involves representing individual stocks as nodes, and edges as relationships between stocks, defined by the (inverse of the) value of the Spearman correlation coefficient[7], with a threshold of 70% (0.70) or higher. The edge weights represent the strength of the correlation between two pairs of stocks.

We chose a higher correlation threshold in order to filter out weaker correlative relationships and focus on the more significant ones, as well as to not end up with dense and meaningless networks. We created one network to represent the entire system as a whole, over the entire time frame of five years. We also created smaller sub-networks representing the relationships within and between select sectors, which can be viewed within the Visualizations section.

In order to represent the temporal aspect of the system, we constructed various static networks to represent different fiscal years and periods of time within our larger dataset, and compared them visually and statistically.

4 Network Visualizations, Statistics & Null Models

In this section, we outline the results of the network metrics/statistics that were calculated on our full networks and sub-networks, along with all network visualizations.

4.1 Network Visualizations

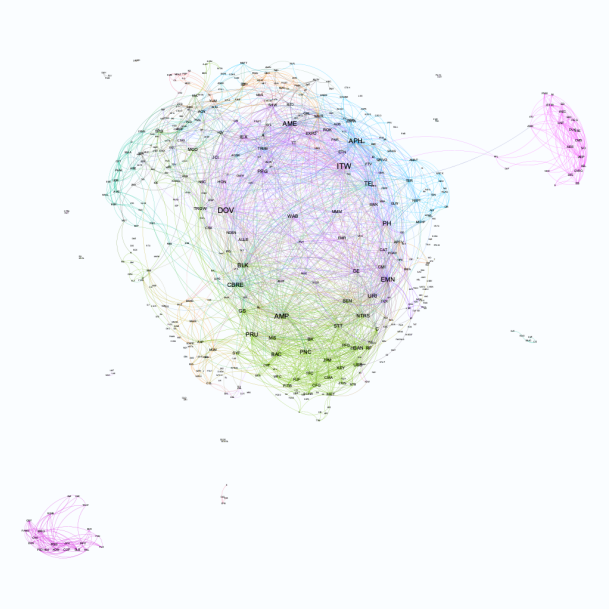

Figure 1: 2019-2023 Network - All Sectors - Monthly

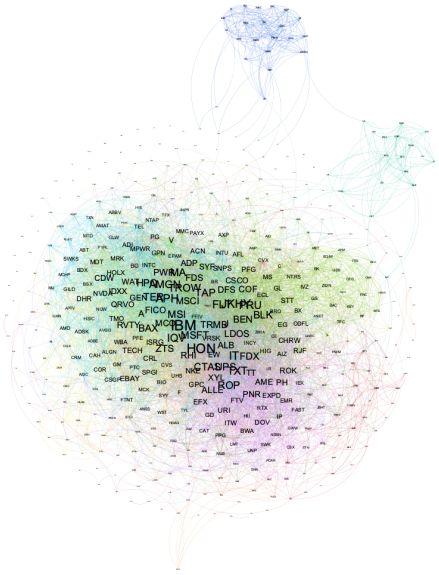

Figure 2: 2019 March Network - All Sectors - Daily

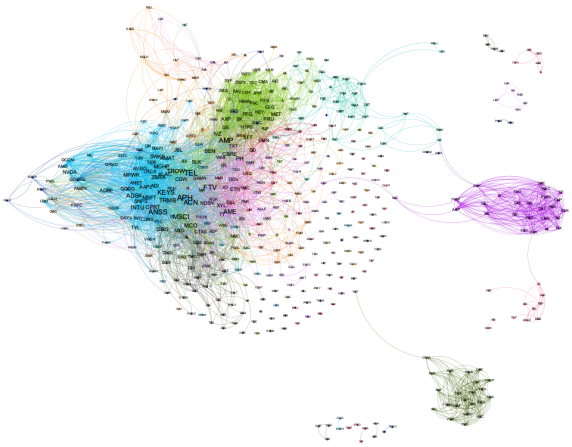

Figure 4: 2022 Network - All Sectors - Daily

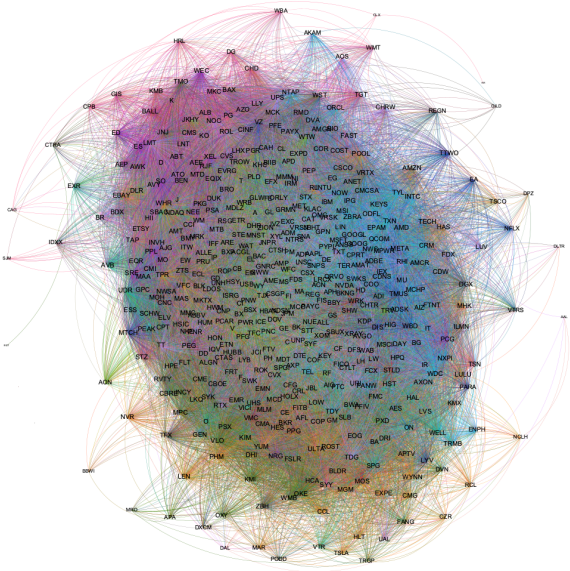

Figure 3: 2020 March Network - All Sectors - Daily

4.2 Statistics & Null Models

Table 1: 2019-2023 Network Statistics

Statistic	Value
Nodes	489
Links	2290
Avg. Degree	9.366
Connected Components	136
Modularity	0.500
Avg. Clustering Coefficient	0.358
Avg. Shortest Path	3.282

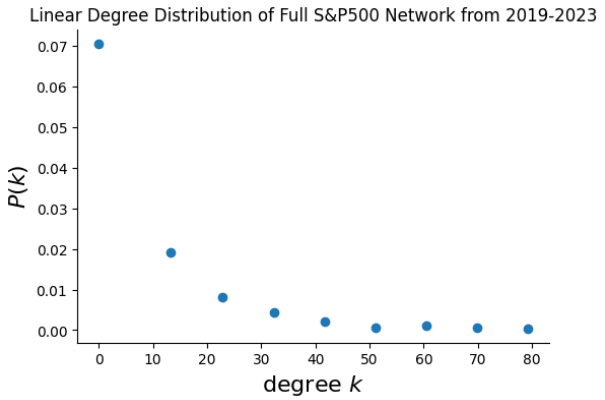

Figure 5: Linear plot of 2019-2023 Network Degree Distribution

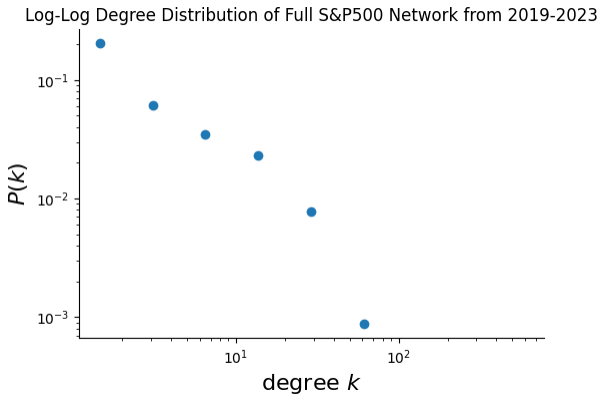

Figure 6: Log-Log plot of 2019-2023 Network Degree Distribution

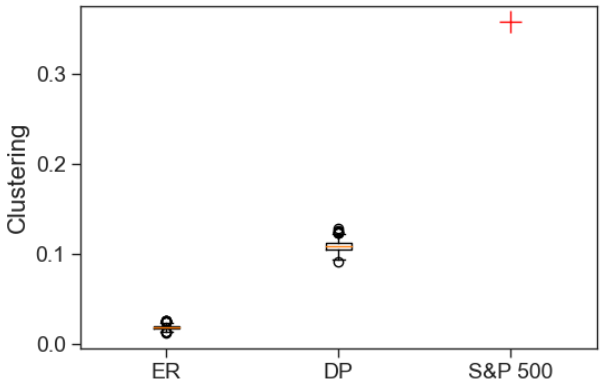

Figure 7: 2019-2023 Network Null Models Clustering Comparison

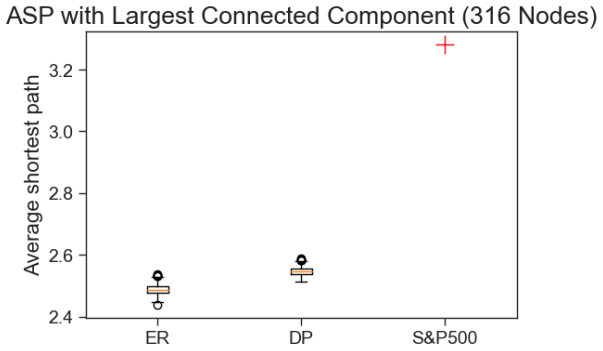

Figure 8: 2019-2023 Network Null Models Average Shortest Path Comparison

When looking at our 2019-2023 Network in Figure 1, we can visually see a community structure appear between sectors. The modularity is moderate at 0.500, however we will see that it is actually much higher than when we focus in on the other sub networks. We can see that we have an extremely high amount of connected components at 136 in relation to our amount of nodes at 489. This tells us that there is a large amount of stocks that do not have a correlation above our threshold of 0.7 over the course of our 5 year time period. Looking at Figure 5 and Figure 6 we can see that the 2019-2023 Network fits a power law degree distribution quite nicely.

For the 2019-2023 Network, we can see that it displays significantly higher clustering coefficients and average shortest path in comparison to our null models, seen in Figure 7 and Figure 8. For the constructed network, we see an average clustering coefficient of 0.358, a much higher number than found in the Degree Preserving model (0.109 ± 0.00552) and the Erdos-Renyi model (0.0191 ± 0.00196). We also see that the average shortest path is much higher in our constructed network (3.282) when compared to the Erdos-Renyi (2.488 ± 0.0155) and Degree Preserving (2.547 ± 0.0125) models.

The comparison to the Erdos-Renyi model allows us to rule out our network being a random network. On the other hand, the comparison to the Degree Preserving model allows us to confirm that the layout of links in our network does in fact carry significance in and of itself, and changing the structure of the network would destroy information and eliminate its inherent nature. We can infer from this that our network does indeed have a community structure.

Since the average shortest path of our constructed network is higher than both null models, we can confirm that our network is displaying behavior that is not random. In addition to this, the fact that our network displays higher than random average shortest path and higher than random average clustering, lends itself to the idea that the nodes in our network are arranging themselves into communities with hubs being the source of interconnection between these communities.

Table 2: March 2019 Network Statistics

Statistic	**Value**
Nodes	489
Links	7272
Avg. Degree	29.74
Connected Components	41
Modularity	0.339
Avg. Clustering Coefficient	0.457
Avg. Shortest Path	2.890

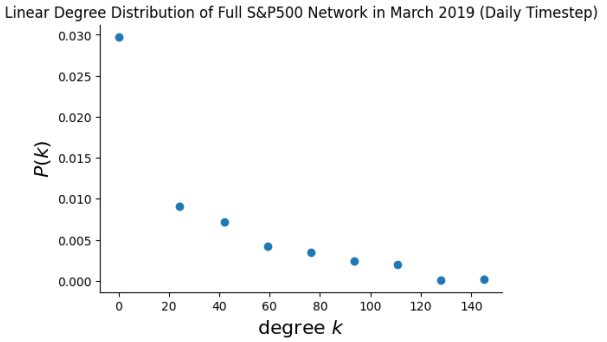

Figure 9: Linear plot of 2019 March Network Degree Distribution

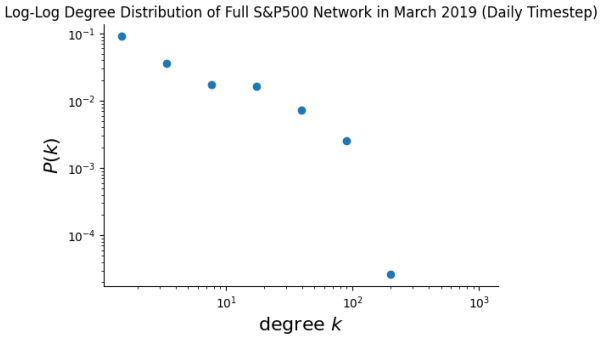

Figure 10: Log-Log plot of 2019 March Network Degree Distribution

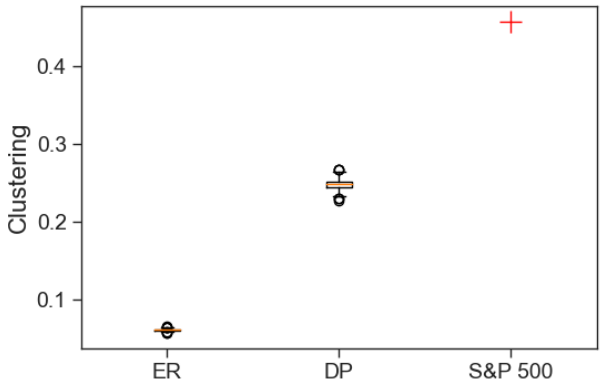

Figure 11: 2019 March Network Null Models Clustering Comparison

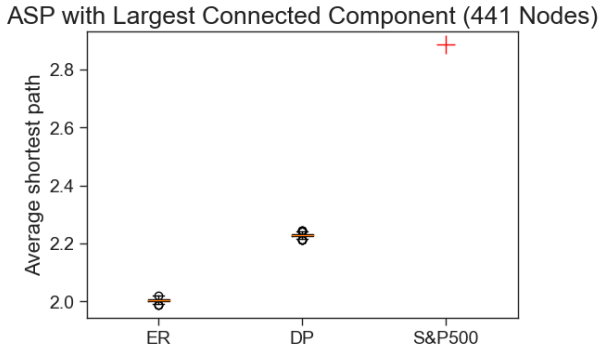

Figure 12: 2019 March Network Null Models Average Shortest Path Comparison

Looking at the 2019 March network shown in Figure 2 with statistics in Table 2, we can see that there is a large increase in links at 7272 in comparison to the 2019-2023 Network at 2290. This is due to the time step in the 2019 March network being daily; we are seeing a higher correlation as stocks move with each other daily. Our modularity also drops to 0.339, indicating that the community structure is less tightly knit than in the 2019-2023 Network. The amount of connected components also drops to 41 which shows us that as a whole our network is better connected, and combining this with the lower modularity we can infer that more stocks are correlated with each other, but are not necessarily forming tight communities. As seen in Figure 9 and Figure 10, we still observe a tight fit to a power law degree distribution.

When looking at our null models in comparison to our 2019 March constructed network, we again see a significant deviation from randomness in our constructed model. Our constructed network again showed higher clustering (0.457) than the Erdos-Renyi (0.061 ± 0.00112) and the Degree Preserving (0.248 ± 0.00608) models. In terms of average shortest path, our model yielded (2.887) whereas the Erdos-Renyi was (2.004 ± 0.00542) and Degree Preservation was (2.227 ± 0.00523).

Similar to the comparison to null models for our 2019-2023 Network, we again see a higher than random clustering and average shortest path. This signifies that our 2019 March network is not random, and that the network is arranging itself into communities with hubs serving as the bridge between communities.

Table 3: March 2020 Network Statistics

Statistic	**Value**
Nodes	489
Links	69506
Avg. Degree	284.28
Connected Components	3
Modularity	0.119
Avg. Clustering Coefficient	0.803
Avg. Shortest Path	1.440

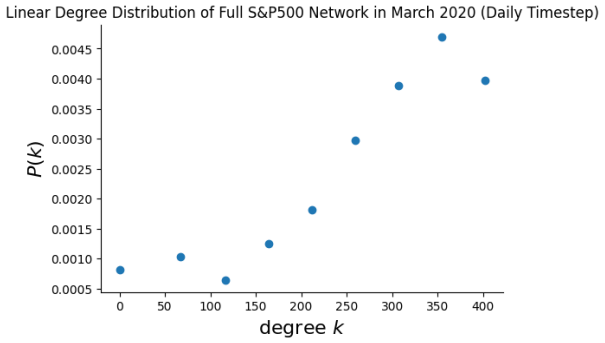

Figure 13: Linear plot of 2020 March Network Degree Distribution

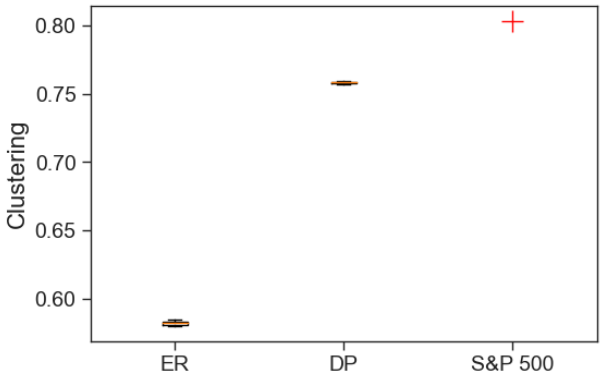

Figure 15: 2020 March Network Null Models Clustering Comparison

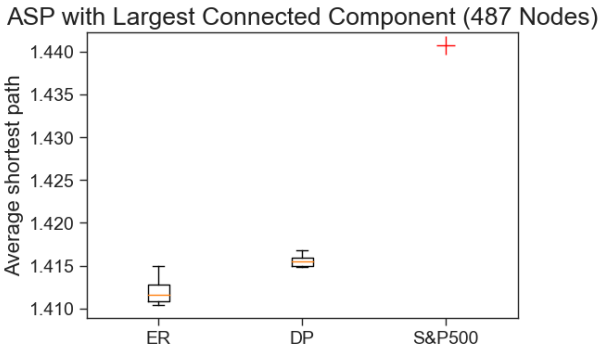

Figure 16: 2020 March Network Null Models Average Shortest Path Comparison

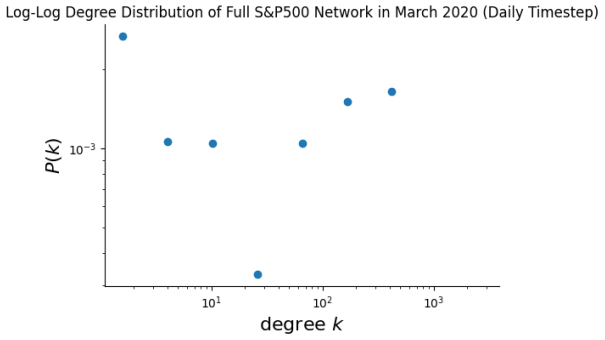

Figure 14: Log-Log plot of 2020 March Network Degree Distribution

The 2020 March Network as seen in Figure 3 and Table 3 visually displays a lack of clear community structure and a much more dense network as a whole. When comparing to the 2019 March Network statistics present in Table 2, we can see that our network has a nearly 10x increase in the number of links. We can also see that the modularity drops very low, and the connected components drops to 3. The degree distribution as seen in Figure 13 and 14 no longer displays any semblance to a power law, rather we see a favoring of very high degree nodes. All of these mentioned factors in combination with the increase in average clustering coefficient and decrease in average shortest path indicates that the

2020 March Network no longer has a meaningful community structure.

When looking at our null models displayed in Figure 15 and Figure 16 in comparison to our March 2020 network, we see a different trend as compared to the previous two networks discussed. Our March 2020 Network average clustering coefficient was (0.803), the Erdos-Renyi model was (0.582 ± 0.00145) and the Degree Preserving model was (0.758 ± 0.000681). For the average shortest path our network produced a value of (1.441) whereas the Erdos-Renyi model produced (1.412 ± 0.00141) and the Degree Preserving model produced (1.416 ± 0.000610).

When comparing 2020 March's average clustering coefficient we see that it is much higher than what is produced by the Erdos-Renyi model, leading us to the conclusion that our network is not displaying random characteristics. However, when we compare to the Degree Preserving model, we see that the average clustering coefficients are much closer than any other network has been to its respective null models. With a difference of only 0.045 between the average clustering coefficients, this tells us that the structure of our network has far less significance than in other time periods and that the way in which nodes are connected to each other does not uncover any underlying information.

We also see a much tighter spread of the values when observing average shortest path length. There is an overall decrease in the average shortest path length when compared to the values found in 2019 March; the difference between 2020 March and the null models is much smaller than it was the previous year. The difference in average shortest path can be attributed to a large increase in the density in the graph, but the closeness of values between the network and the null models lends itself to the conclusion that this network does not have hubs bridging communities, and rather arranges itself in a more random fashion. Combining the conclusions drawn from the average clustering coefficient and the average shortest path, it seems as though our March 2020 network does not display much of a meaningful community structure, and rather has high density with the arrangement of nodes and links being closer to random than in March 2019 or any other year.

Table 4: 2022 Network Statistics

Statistic	**Value**
Nodes	489
Links	4621
Avg. Degree	18.9
Connected Components	84
Modularity	0.537
Avg. Clustering Coefficient	0.539
Avg. Shortest Path	3.493

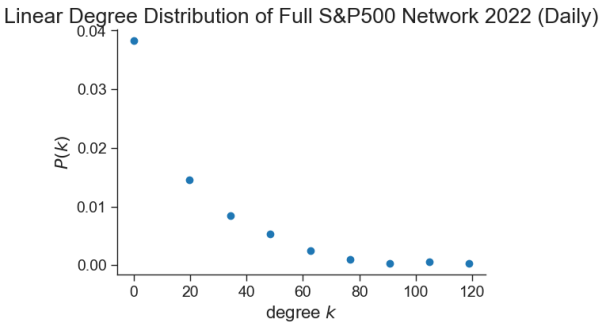

Figure 17: Linear plot of 2022 Network Degree Distribution

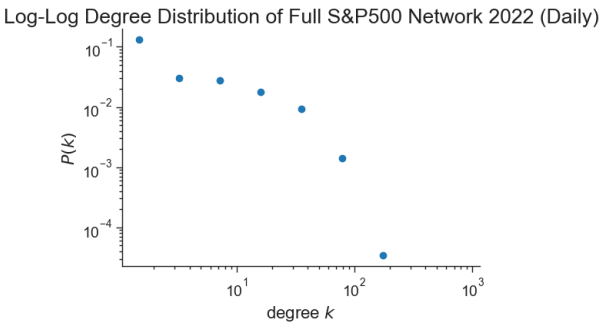

Figure 18: Log-Log plot of 2022 Network Degree Distribution

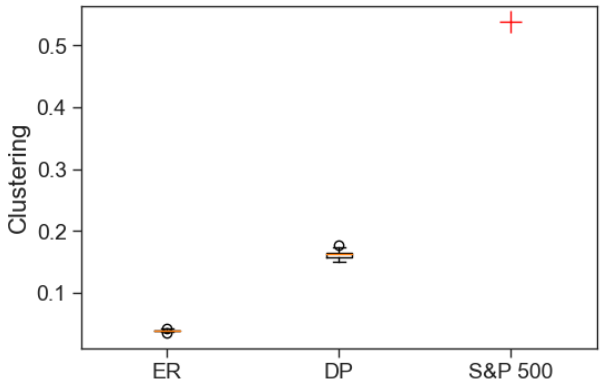

Figure 19: 2022 Network Null Models Clustering Comparison

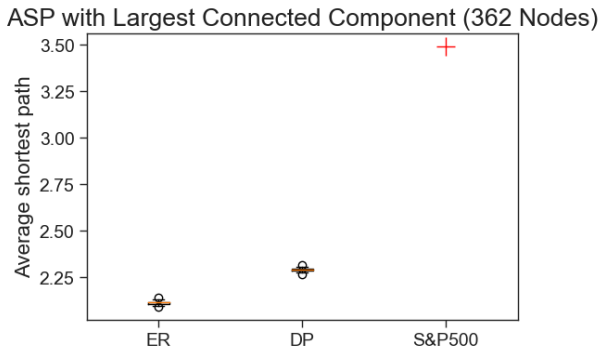

Figure 20: 2022 Network Null Models Average Shortest Path Comparison

Looking at the 2022 Network visualized in Figure 4 and with statistics seen in Table 4, we can see that the network appears to be forming meaningful communities once again. With a moderate amount of links at 4621 and connected components at 84, the network much more closely resembles the 2019-2023 Network. We can also see the return of a power law in the degree distribution, as seen in Figure 17 and Figure 18, although it is not as close of a fit as it was in the 2019-2023 Network's (Figure 5 and Figure 6). The number of connected components is 84, showing that in 2022 there is much more variation in the behavior of stock prices. The modularity is 0.537 which closely resembles the modularity of 0.500 seen in the 2019-2023 Network. We do see a higher average clustering coefficient and a lower average shortest path in the 2022 Network than in the 2019-2023 Network, indicating that in 2022 the network had more nodes bridging communities, and a lower prevalence to high degree hubs.

When comparing our null models to our 2022 Network, we see our network with an average clustering coefficient of (0.539) whereas the Erdos-Renyi produced (0.0338 ± 0.00122) and Degree Preserving produced (0.162 ± 0.00541). For the average shortest path we have our model at (3.493), Erdos-Renyi (2.1112 ± 0.00898) and Degree Preservation (2.289 ± 0.00729).

Both the comparison of average shortest path length and average clustering coefficient between both null models and the 2022 Network show us that our network is not displaying traits characteristic of random networks.

5 Results

In this section, we present the findings of our study, specifically by providing analysis for each of the previously outlined research questions.

5.1 Evolution of Stock Price Correlations and Network Structure

How do stock price correlations (between and within sectors) evolve over the pre, during, and post-COVID-19 pandemic periods, and what are the underlying structural changes within the S&P500 network?

In order to answer this question, we analyzed the evolution of stock price correlations within the S&P500 over distinct temporal stages - pre-, during-, and post-COVID-19 pandemic, which revealed subtle dynamic shifts in market behavior and in turn, structural changes.
Comparisons of metrics such as centrality measures, average degree, degree distributions, clustering coefficients, as well as additional methods like

Katz centrality provided valuable insights into the changing stock correlation landscape, which we cover more extensively in the statistics section.

The structural changes visible in the network visualizations underscored the impact of the pandemic on market dynamics at the time, with major shifts in community structure and clustering patterns, which only served to highlight the inter-connectedness of stocks and the vulnerability of the entire market to external shocks, even for those stocks and sectors that have a general tendency towards stability.

To give a more specific example of the structural changes observed, as outlined in the null models section, the network in the time period of March 2020 lost most meaningful structure, becoming statistically closer to the random null models. Even traditionally stable and isolated sectors such as Energy, for instance, were heavily impacted by the onset of the pandemic, leading to a dense and deeply interconnected graph.

However, this observation is diluted by the fact that we see a near full recovery within the next two years, which leads us to believe that the market as a whole has a tendency to return to its pre-COVID-19 state, characterized by a degree distribution that follows the power-law and a tendency to form mostly isolated communities. While the network does revert back to having a distinct community structure, there are differences in what communities are forming, such as the IT sector forming a more prominent community, which is unlike what we observed in March 2019. Cross-checking with market trends, we know that the IT sector experienced a 'boom' period post-pandemic, which accounts for this change[11]. Another change we see is that the Consumer Cyclical sector, which is a subset of Consumer Discretionary and encompasses such things as hotel companies and cruise lines, also formed its own small, but distinct community in the years following the onset of the pandemic.

The dynamic nature of this network makes it difficult to assert any behavior with certainty, but the observed quality of the network returning to a 'favoured' structure is a good indicator of future outlooks for general network dynamics, making it easier for an investor to make assumptions regarding "safe" sectors and stocks to invest in. However, during times of major disruption it is more difficult to make any predictions regarding the network's structure as we have seen that the structure tends towards randomness in these cases.

5.2 Portfolio Diversification Strategies

How can an investor create a diversified, relatively risk-free portfolio as best as possible by looking at the correlation trends between sectors?

In order to find how the average investor can minimize risk within a portfolio, we investigated the ways in which an investor can 'hedge' their bets; essentially, insure themselves against losing all of their money during periods when the market is full of uncertainty and price fluctuations. One way in which we did this is by looking at communities within the S&P500 and trying to spread investments across communities that are not highly correlated with each other. Since we have already found that within the network, stocks do tend to form communities, frequently defined by what GICS sector a stock is in, we will investigate which sectors are least correlated. By forming sub-networks of each sector, we found the number of links within each individual sector network and then created a network of two sectors to find how correlated they are. For example, suppose we have sector A and sector B; we can find the number of links between the two sectors with the following calculation: $\Delta = |AB_{\text{links}}| - (|A_{\text{links}}| + |B_{\text{links}}|)$ Since links are defined by the value of the Spearman correlation coefficient, the higher the Δ value, the higher the correlation is between these two sectors. Using a monthly time scale in order to capture larger trends in the market, we find that in general, the Industrials and Financials sectors are very highly correlated with other sectors. We also see that Industrials and Financials have the highest correlation with each

other, whereas the lowest correlation is between the sectors Utilities and Energy.

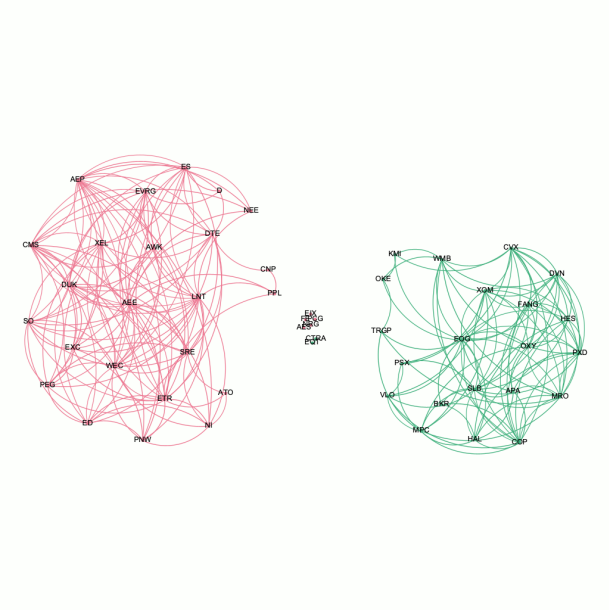

Figure 21: Energy(Green) & Utilities(Pink) Sectors (Subset of 2019-2023 Network)

In Figure 21, we can see that there are no links between the different colors because the two networks are perfectly isolated from each other. Here, Utilities are pink and Energy is green.

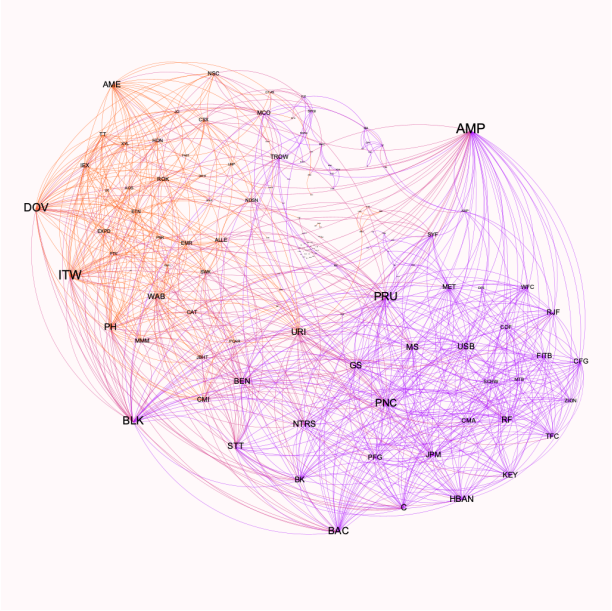

Figure 22: Financial(Purple) & Industrial(Orange) Sectors (Subset of 2019-2023 Network)

In Figure 22, we can clearly see high interconnectedness between the stocks. Here, Financials is in pink and Industrials are in orange. We can see that the two sectors are very correlated with each other. Certain nodes, especially in the Financials sector, do not link at all with the Industrials side; this does not necessarily mean they will not be impacted by activity from other stocks and sectors, albeit indirectly.

A less naive approach for diversifying a portfolio is to assume that sectors do tend to influence each other, and thus diversifying by sector may not always be the best strategy when taking into consideration market dynamics. With this assumption in mind, we will use the Louvain method[2] to create partitions in order to determine whether or not the sectors are an optimal way of categorizing stocks in the context of correlation.

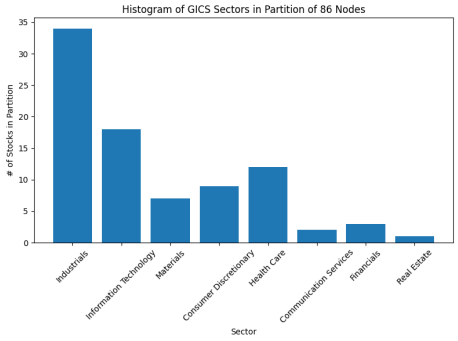

Figure 23: Histogram showing frequency of individual stocks belonging to each GICS Sector present in largest partition

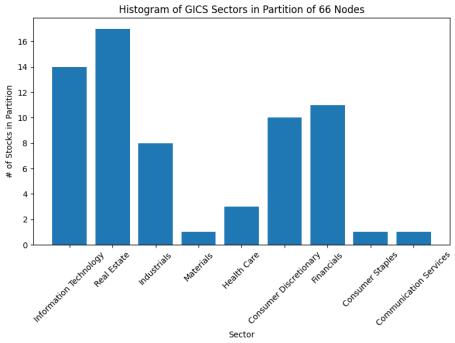

Figure 25: Histogram showing frequency of individual stocks belonging to each GICS Sector present in third largest partition

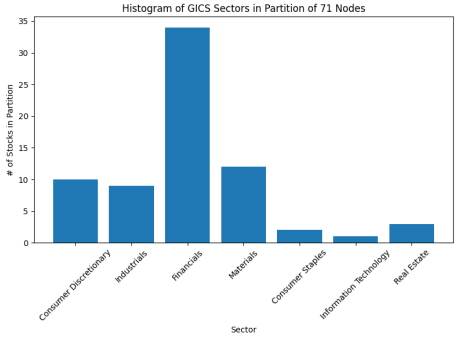

Figure 24: Histogram showing frequency of individual stocks belonging to each GICS Sector present in second largest partition

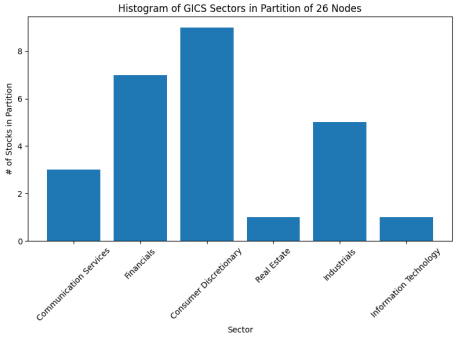

Figure 26: Histogram showing frequency of individual stocks belonging to each GICS Sector present in fourth largest partition

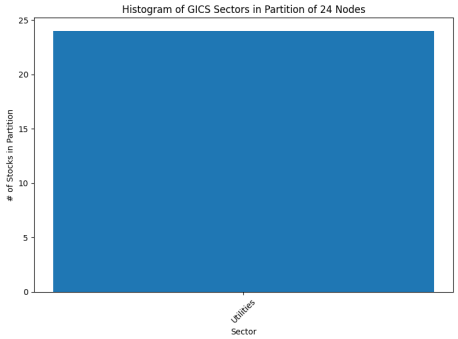

Figure 27: Histogram showing frequency of individual stocks belonging to each GICS Sector present in fifth largest partition

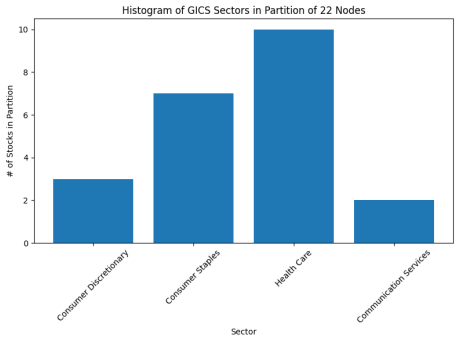

Figure 28: Histogram showing frequency of individual stocks belonging to each GICS Sector present in sixth largest partition

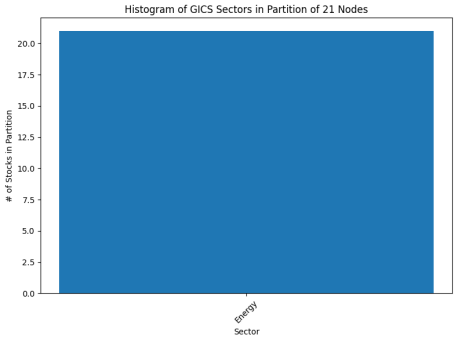

Figure 29: Histogram showing frequency of individual stocks belonging to each GICS Sector present in seventh largest partition

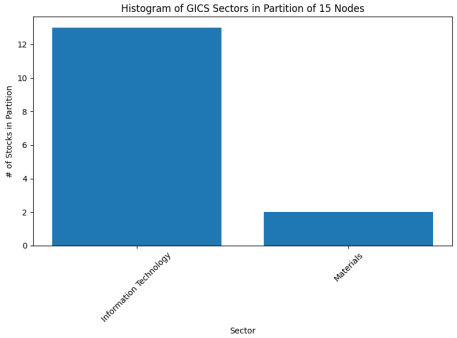

Figure 30: Histogram showing frequency of individual stocks belonging to each GICS Sector present in eigth largest partition

When using the Louvain method for our network of five years with monthly timesteps, we have around 8 significant partitions, with the majority of the partitions having less than five nodes. Within the partitions that consist of less than five, but more than one node, the vast majority of them include only stocks of the same sector. Looking at our 8 largest partitions we see the following: Our largest partition (Figure 23) is dominated by the Industrials sector, with a significant amount of IT stocks as well. The second largest partition (Figure 24) is dominated by the Financials sector, with smaller portions consisting of Consumer Discretionary, Industrials, and Materials. Our third largest (Figure 25) is

dominated by IT and Real Estate, with Industrials, Consumer Discretionary, and Financials also making up large portions. We once again see that Utilities and Energy are two very significant partitions, and they each comprise 100 percent of the stocks in their respective partitions (Figure 27 & Figure 29).

While these findings disprove the theory that the sectors do not impact each other on large scales, it further proves that the diversification of a portfolio can be achieved by investing in Utilities and Energy. The overall stability of these sectors due to their inherent nature, combined with the fact that our sector analysis shows that they are extremely minimally impacted by the movements of other sectors makes them a good choice for lower-risk investing. It also provides a new strategy for diversification of a portfolio by providing us with groups of stocks (partitions) that minimize the correlation between each group, allowing for an investor to diversify their portfolio by spreading out funds between partitions.

5.3 Influential Stocks and Predictive Power

How can we predict the influence of nodes and how does that help with predicting future performance for investing purposes?

In order to find the most influential nodes in this network, we applied the Katz centrality[12] measure on the network as a whole. The Katz centrality measure was chosen as it takes into account both the direct and indirect influence of a node, as opposed to a pure degree centrality measure which only accounts for direct connections.

In our financial network, we frequently observed that the nodes which have a high Katz centrality, such as the stock DOV, also have a higher degree centrality. We make the assumption that in this stock network, these high centrality nodes represent market leaders; the companies that commonly have many connections across many industries, and which may have the power to impact trends in the market. Market capitalization is a good indicator for the size of the company, stability, and market share, which is why we originally assumed that those stocks with higher market capitalization would be a one-to-one match with the most influential nodes. However, this proved untrue. Stocks like AAPL (Apple), MSFT (Microsoft), and AMZN (Amazon), are the heaviest 'weighted' stocks in the S&P500 by market capitalization[10], but they did not even manage to crack the top ten of most influential nodes by the Katz centrality measure. This discordance accentuates the constraints of relying solely on financial measures like market capitalization as a predictor of influence.

Looking at sector-specific dynamics and the inherent nature of the companies and sectors themselves draws more nuanced insights in determining why a node is influential, which helps us draw conclusions with stronger certainty as to the causal effect between stocks and why it occurs that way.

Notably, we found that the companies within the Industrials sectors consistently had the highest Katz centrality values. This suggests that companies in this sector are the most influential in the network as a whole, but as outlined in the sector analysis section, it does not mean that every other sector and stock is equally impacted by the stocks in this sector. Looking at the market as a whole, we can see why; the companies in the industrial sector are major players in manufacturing, infrastructure, and other essential goods and services. These also tend to have extensive supply chain connections, such as with materials suppliers and customers both. These connections are one reason why shifts in this sector could potentially have a ripple effect through the rest of its network, attesting to its influence in the network.

Understanding the influence that this sector has consistently had on the market, prompts us to suggest that a risk-averse investor diversify away from stocks and sectors that are heavily influenced by the industrial sector, in order to mitigate risk in the chance of a market shift in the future. Since

these Katz centrality values were found over the full time scale from 2019-2023, these stocks have consistently remained the most influential over the whole time period, and therefore it is a reasonable assumption that they will stay at a similar level in the future, barring another event of total economic collapse.

While our findings provide valuable insights, it is important to take into consideration other economic factors as well, as the stock market does not operate in a vacuum. Factors like market sentiment, interest rates and other economic indicators, regulations, and overall performance have an impact. However, the results provided here are a complementary indicator on top of the factors listed above, consistently demonstrating sector influence over time, excluding major disruptors.

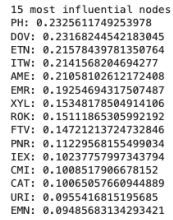

Figure 31: Top 15 Most Influential S&P500 Stocks

Figure 21 shows the top 15 most influential stocks by Katz centrality, in descending order.

6 Discussion

Overall, our analysis of the S&P500 network provides valuable insights into the changing landscape of stock price correlations and structural changes over different time periods, including the 'before and after' picture of the COVID-19 pandemic, in order to shed light on the interconnection between stocks and the network's structural response market shocks. We also endeavored to provide portfolio recommendations for an average risk-averse investor, considering factors such as sectors versus partitions and influence of nodes.

While we believe that our analysis was beneficial in uncovering newer understandings of the ever-changing nature of the stock market, there are inherent limitations to our approach. One such limitation, and potentially the most significant, lies in the fact that we are only taking into account correlations between stock prices. This approach limits our ability to capture certain nuances that come from more qualitative data such as market sentiment, political events, regulatory changes, etc. Thus, our findings represent a niche of overall market behavior but will not perfectly capture the complexity of such a financial system as the US stock market.

Despite this limitation, we do think we were able to successfully address our research questions regarding structural changes in the S&P500 and portfolio recommendations for a risk-averse investor. Our findings do align with the existing literature, which also highlights the sensitivity of the stock market to changes that impact the structure of the network.

In regards to future work, we may want to consider implementing a multiplex network in which Granger causality and Spearman correlation are applied as link definitions, as well as other layers that could represent different modes of interaction such as the impact of regulatory changes, social media, or politics. As well, applying machine learning techniques for better predictive strength could also be a subject of focus in future iterations.

7 Methods

Throughout the course of the project, we made use of several Python libraries and pieces of pre-made code.

To obtain our data we used Pandas to scrape a Wikipedia page containing the list of stocks in the S&P500 index along with their corresponding company names and GICS sectors. We also used the open-source yfinance API to obtain all of our time

series data.
Once we had the relevant data, we calculated the percent change between each time step which allowed us to calculate Spearman correlation. Spearman correlation was calculated using the spearmanr function from the SciPy Stats module. The Spearman correlation between two sets of time series data is given as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where n = the number of pairs of data points and d_i = the difference in ranking between two corresponding data points in the two datasets. The final step was to filter out any sectors that were not wanted in the network being constructed. NumPy was also used throughout the process as a means to increase efficiency. Final network visualizations were constructed using Gephi, with intermediate visualizations also created in NetworkX.

Plotting for null models and degree distribution was done using Matplotlib and code created by Dr. Emma Towlson.

In order to create our networks, we used the Python library NetworkX. NetworkX allowed us to take Pandas dataframes and convert them into network format. Once in network format, we were able to use NetworkX alongside NumPy to find our basic statistics on each network.

Given the nature of our constructed networks and the amount of information contained in the degree and structure of the network, the null models we chose to use for comparison were the Erdos-Renyi and Degree Preserving models. All of our networks were disconnected, so in order to measure our average shortest path, we found the largest connected component and then used this to calculate the average shortest path and create our null models.

To create our Erdos-Renyi null models we created ensembles of networks using NetworkX's *erdos_renyi_graph*() function that had the same amount of nodes as our constructed network that we wanted to compare, and the probability for edge creation was given as the constructed network's number of edges divided by the maximum amount of edges possible in that network. To calculate the average clustering coefficient we would find the clustering coefficient for each node using NetworkX's clustering() function and then take the mean using NumPy's mean() function. We would then store this in an array, which we then used NumPy's mean() and std() functions on to find the mean and standard deviation for the ensemble. To find our average shortest path, we needed to find the largest connected community since our networks were disconnected. In order to do this we used NetworkX's *connected_components*() and subgraph() functions to create a subgraph of our largest connected subgraph. Once we had this, we could again create an ensemble of Erdos-Renyi networks in the same way as before, only now we were using our subgraph for the amount of nodes and probability of edge creation. For each network produced we found the average shortest path using NetworkX's *average_shortest_path_length*() function, and stored it in an array, where we calculated the mean and standard deviation the same as before.

To create our Degree Preserving null models we created ensembles of networks using NetworkX's *double_edge_swap*() on our constructed network. We performed one double edge swap per edge in the network. The process of finding average clustering coefficient and average shortest path length were exactly the same as for our Erdos-Renyi models.

Due to computational limits, we were only able to create ensembles of size 1000 for our full network that spanned over the entire time frame of 2019-2023. For the 2019 March and 2020 March networks the null models had to be constructed using ensembles of 10. For the 2022 network the null models were constructed using ensembles of 100.

To find our partitions we used NetworkX's *louvain_partitions*() function which uses the Louvain Community Detection algorithm and produces partitions. To calculate Katz Centrality for

measuring node influence, we used the NetworkX and Numpy function *katz_centrality_numpy*(). Here, we set α, the attenuation factor, to 0.1, in order to reduce the influence of indirect connections on the influence, and focus more on direct connections (shortest paths). We then sorted the results in descending order, focusing on the top fifteen nodes outputed with the highest Katz Centrality values.

8 Codebase

Shared Github repository link:

- https://github.com/haydenac11/572-Project

Citations are included within code files and discussed in the methods section.

Acknowledgements

A special thanks to Dr. Emma Towlson for helpful constructive feedback and getting us on the right track for this project

References

Article Sources

[2] Vincent D. Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 10 (2008), p. 10008. DOI: 10.1088/1742-5468/2008/10/P10008.

[4] Mariana Durcheva and Pavel Tsankov. "Granger causality networks of S&P 500 stocks". In: *AIP Conference Proceedings: Applications of Mathematics in Engineering and Economics* 2333.110014 (2021). DOI: https://pubs.aip.org/aip/acp/article-abstract/2333/1/110014/1027782/Granger-causality-networks-of-S-amp-P-500-stocks.

[5] Minjun Kim and Hiroki Sayama. "Predicting stock market movements using network science: an information theoretic approach". In: *Applied Network Science* 2.35 (2017). DOI: https://link.springer.com/article/10.1007/s41109-017-0055-y#citeas.

[6] Yun-Jung Lee and Gyun Woo. "Analyzing the Dynamics of Stock Networks for Recommending Stock Portfolio". In: *Journal of Information Science & Engineering* 35.2 (2019). DOI: https://openurl.ebsco.com/EPDB%3Agcd%3A1%3A20307361/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A135133483&crl=c.

[8] Janusz Miskiewicz and Dorota Bonarska-Kujawa. "Analyzing the Dynamics of Stock Networks for Recommending Stock Portfolio". In: *Evolving Network Analysis of SP500 Components: COVID-19 Influence of Cross-Correlation Network Structure* 24.1 (2022). DOI: https://www.mdpi.com/1099-4300/24/1/21.

[13] Takeo Yoshikawa Yuta Arai and Hiroshi Iyetomi. "Dynamic Stock Correlation Network". In: *Procedia Computer Science* 60 (2015). DOI: https://www.sciencedirect.com/science/article/pii/S1877050915024205.

Other Sources

[1] Ran Aroussi. *yfinance*. URL: https://pypi.org/project/yfinance/.

[3] Wikipedia Contributors. *List of SP 500 companies*. URL: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

[7] Lund Research Ltd. *Spearman's Rank-Order Correlation.* 2018. URL: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

[9] MSCI. *The Global Industry Classification Standard (GICS)*. URL: https://www.msci.com/our-solutions/indexes/gics.

[10] Nathan Reiff. *The Top 25 Stocks in the S&P500*. 2023. URL: https://www.investopedia.com/ask/answers/08/find-stocks-in-sp500.asp.

[11] UNCTAD. *How COVID-19 triggered the digital and e-commerce turning point*. 2021. URL: https://unctad.org/news/how-covid-19-triggered-digital-and-e-commerce-turning-point.

[12] Unknown. *Katz Centrality*. URL: https://www.sci.unich.it/~francesc/teaching/network/katz.html.