Exploring Correlations in the Changes of Stock Sectors Listed on the NYSE and NASDAQ Using Cosine Similarity

Deliar Mohammadi 30072994 (CPSC 572)
Soroosh Esmaeilian 30153011 (CPSC 672)

Fundamentals of Social Network Analysis and Data Mining

Fall 2022

Department of Computer Science

University of Calgary

Abstract

The aim of this paper is to explore the influence of economic sectors on one another and identify the most influential stocks within the market. For our purposes, influence is defined as the ability to move and affect other stock or sector prices. Data was gathered from the NYSE and NASDAQ beginning in January 2010 and ending in December of 2020. This data was converted into monthly percentage changes, and then links for our network were calculated between any two stocks by taking the cosine distance between their percentage change vectors. Armed with this data, we were now able to plot individual sectors as our nodes, and link these nodes together via the calculated cosine distances. The resulting graphs can be seen in Figure 4.1 & Figure 4.3.

By looking at nodes with the highest degree centrality and eigenvector centrality, we were able to determine the most influential stocks in the network, which were BMO and JPM. Additionally, we found that almost all of the influential stocks in our network were in the banking or ETF sector, whereas the least influential stocks came from the biotechnology sector. Besides finding influential stocks, we were also able to leverage average linkage clustering to determine inter and intra-sector correlation. We found the strongest intra-correlated sector to be Mortgage Trusts, and the lowest ranking to be Biotechnology. When it came to inter-sector correlation, we found that Electric utilities paired with Multi-utilities was at the top of our list, other close pairings include airlines & hotels, and machinery & packaging. By identifying the most influential stocks in the network, along with the inter-sector correlations, we have made good progress in uncovering the behaviors of stocks in response to changes in other stock prices. Going forward, we can develop improved machine learning algorithms that incorporate stock relationships as input for stock price forecasting, resulting in more accurate predictions, and thus more profitable outcomes.

Introduction

Stock markets play a large role in the development and growth of a nation's economy. They allow companies to divide themselves into shares and sell these shares to investors, in order to raise capital for operations. Investors buy shares of a company in hopes that the share value will rise and net them a profit[1]. Over time, stock markets have proven to be very lucrative for investors who could identify and act upon economic trends. This lucrative nature of stock markets draws in many researchers from different backgrounds including physics, mathematics, economics, statistics, and computer science[2], all of which hope to discover new trends and patterns among the many stock indicators.

Research in the field of stock correlations and price prediction is quite extensive, especially with the growth of machine learning algorithms in recent history. Typical research processes use some form of neural networks to "quantify similarities in time series data"[2]. These include PCA-LSTM models used by Wen, Lin, and Nie[3] which predict the dollar value of a stock on any day given its previous performance. The problem with these approaches is that the stock market is extremely chaotic. There is a multitude of factors that can affect the underlying price of stock including news, natural disasters, pandemics, wars, and many more. Attempting to predict stock prices merely on historic information can yield irregular or inconsistent results.

In this paper we attempt to do away with the chaos and volatility of individual stock price changes, and instead, explore how entire sectors relate to each other. Historic stock data from 2010 to 2020 was scraped using python scripts and the YFinance API[5]. This data then went under further processing to be converted into monthly percentage change. After the data was processed, we then go on to calculate the similarity of any two stocks by finding the cosine distance of their respective vectors. These vectors will be either 12 dimensional, to represent a fiscal year. Or they will be 143 dimensions to represent the entirety of the data set from 2010-2020. In addition to obtaining these percentage changes, the individual stocks were also placed into their respective sector. These sectors range from banking to technology, and even pharmaceuticals. Given all the stocks that belong to a particular sector, we can calculate inter-sector similarity using average linkage clustering as defined in Barabasi's Network Science[4], which is the *"average of x_{ij}, over all of the node pairs i and j that belong to different communities"*. In addition to calculating the inter-sector similarities, we will also calculate intra-sector similarities to determine how stocks within a single sector relate to each other.

Besides sector similarities, we will look into which stocks are the most and least influential in our network. These will be determined in two ways, the first being the degree-centrality of nodes. Nodes with higher degree centrality will be considered more influential whereas nodes with lower degree centrality are less influential. In addition to degree-centrality, we will also look at the eigenvector centrality of nodes in our network. Eigenvector centrality may reveal some stocks that have high transitive influence, which were not highlighted by degree centrality. Moreover, having two methods in which we score nodes gives us more confidence in our results.

In the next section we will delve deeper into our research questions, followed by the methodology that we employed, and a thorough description of our data set. Beyond that, we will present our findings, which will include the fundamental statistics of our network, visualizations of our network, and key results that were derived from the data. The paper will then end with a final discussion section that summarizes the entirety of our semester's work.

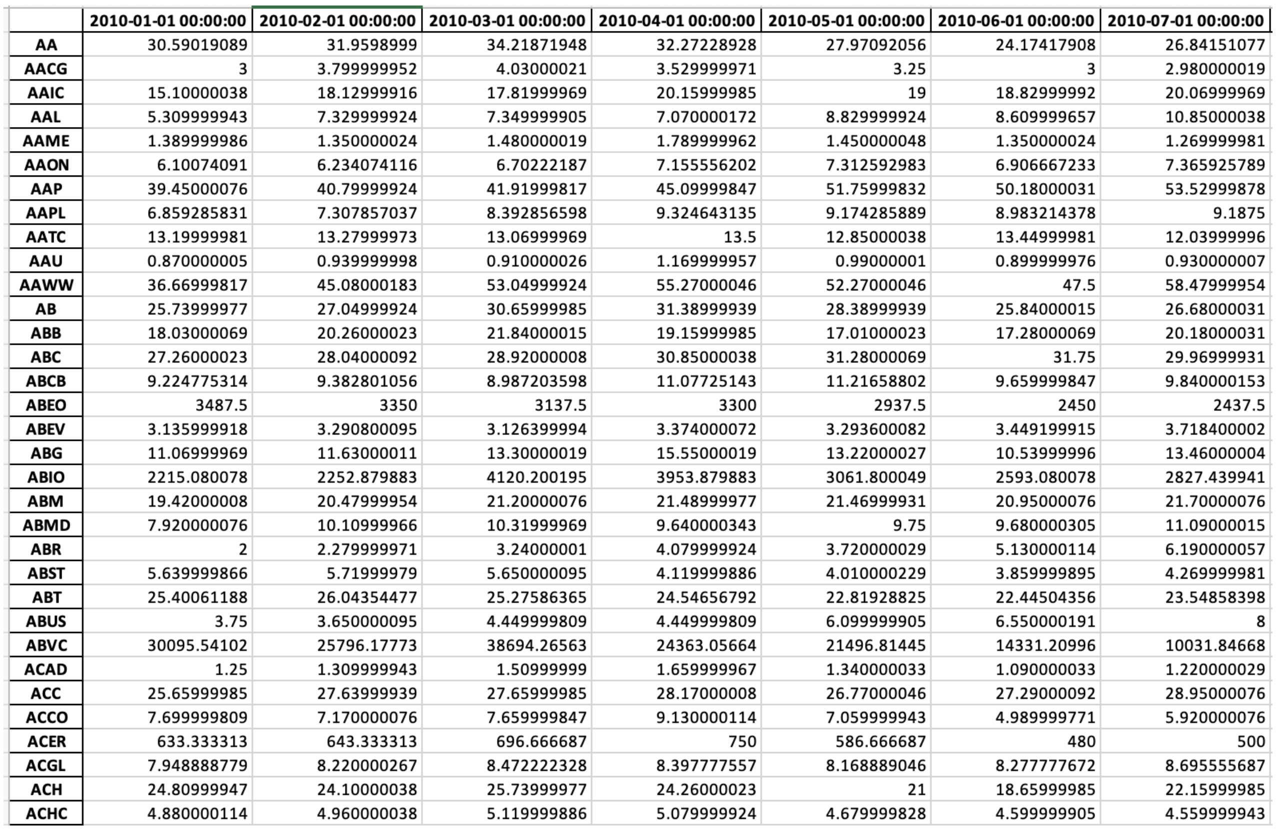
Research Questions

There are four main research questions we intend to tackle for this paper, they are the following:

1. What are the most influential stocks in our network?
   - This is determined by looking at the weighted degree of each node. Nodes with larger degrees have more "linkage" to the rest of the network.
   - Another method to find these nodes is to look at the eigenvector centrality of nodes.
   - Additionally, we would like to look at what sectors these stocks belong to.
2. What are the least influential stocks in our network?
   - This is determined by looking at the stocks with the lowest weighted degree. Having a low weighted degree corresponds to not having a significant impact on the rest of the network
   - Similar to our first research question, we can look at the eigenvector centrality of nodes to determine which are at the bottom.
   - These nodes can also be seen as isolated from the rest of the market. Meaning they don't rely on, or change in response to movement from other stocks.
3. How are stocks within the same sector correlated to each other? Does this differ between sectors?
   - This question will be uncovered by calculating the average linkage clustering between the stocks within a sector
   - Answering this question can give us valuable insight as to which sectors are more volatile, and which are more stable.
4. How are stocks in different sectors related to each other? What are some of the strongest sector-to-sector correlations?
   - This question will be answered using similar methods to question 3. However this time we will be calculating average linkage clustering between two different sectors rather than a single sector.
   - Understanding the results of this question can yield promising predictions of a sector's performance, given the performance of sectors that correlate to it.

Data Set Description & Methods

The data for this project was gathered using a python script along with the YFinance API[5]. This allowed us to obtain the price of all stocks listed on the New York Stock Exchange NYSE or NASDAQ. This data ranged from January 2010 to December 2020. The data points we obtained were from the first day of each month, this was to reduce the data set size and make later computations more feasible. Figure 1.0 is a snippet from the data set.



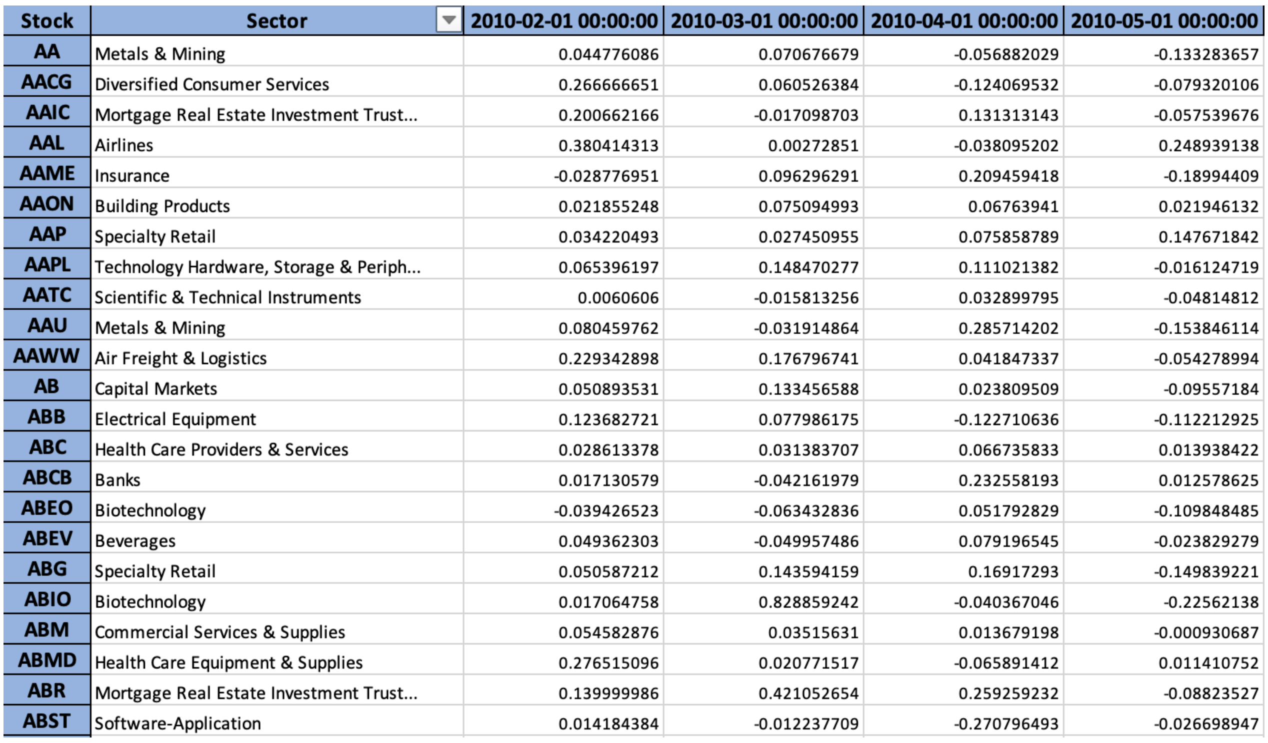
	2010-01-01 00:00:00	2010-02-01 00:00:00	2010-03-01 00:00:00	2010-04-01 00:00:00	2010-05-01 00:00:00	2010-06-01 00:00:00	2010-07-01 00:00:00
AA	30.59019089	31.9598999	34.21871948	32.27228928	27.97092056	24.17417908	26.84151077
AACG	3	3.799999952	4.03000021	3.529999971	3.25	3	2.980000019
AAIC	15.10000038	18.12999916	17.81999969	20.15999985	19	18.82999992	20.06999969
AAL	5.309999943	7.329999924	7.349999905	7.070000172	8.829999924	8.609999657	10.85000038
AAME	1.389999986	1.350000024	1.480000019	1.789999962	1.450000048	1.350000024	1.269999981
AAON	6.10074091	6.234074116	6.70222187	7.155556202	7.312592983	6.906667233	7.365925789
AAP	39.45000076	40.79999924	41.91999817	45.09999847	51.75999832	50.18000031	53.52999878
AAPL	6.859285831	7.307857037	8.392856598	9.324643135	9.174285889	8.983214378	9.1875
AATC	13.19999981	13.27999973	13.06999969	13.5	12.85000038	13.44999981	12.03999996
AAU	0.870000005	0.939999998	0.910000026	1.169999957	0.99000001	0.899999976	0.930000007
AAWW	36.66999817	45.08000183	53.04999924	55.27000046	52.27000046	47.5	58.47999954
AB	25.73999977	27.04999924	30.65999985	31.38999939	28.38999939	25.84000015	26.68000031
ABB	18.03000069	20.26000023	21.84000015	19.15999985	17.01000023	17.28000069	20.18000031
ABC	27.26000023	28.04000092	28.92000008	30.85000038	31.28000069	31.75	29.96999931
ABCB	9.224775314	9.382801056	8.987203598	11.07725143	11.21658802	9.659999847	9.840000153
ABEO	3487.5	3350	3137.5	3300	2937.5	2450	2437.5
ABEV	3.135999918	3.290800095	3.126399994	3.374000072	3.293600082	3.449199915	3.718400002
ABG	11.06999969	11.63000011	13.30000019	15.55000019	13.22000027	10.53999996	13.46000004
ABIO	2215.080078	2252.879883	4120.200195	3953.879883	3061.800049	2593.080078	2827.439941
ABM	19.42000008	20.47999954	21.20000076	21.48999977	21.46999931	20.95000076	21.70000076
ABMD	7.920000076	10.10999966	10.31999969	9.640000343	9.75	9.680000305	11.09000015
ABR	2	2.279999971	3.24000001	4.079999924	3.720000029	5.130000114	6.190000057
ABST	5.639999866	5.71999979	5.650000095	4.119999886	4.010000229	3.859999895	4.269999981
ABT	25.40061188	26.04354477	25.27586365	24.54656792	22.81928825	22.44504356	23.54858398
ABUS	3.75	3.650000095	4.449999809	4.449999809	6.099999905	6.550000191	8
ABVC	30095.54102	25796.17773	38694.26563	24363.05664	21496.81445	14331.20996	10031.84668
ACAD	1.25	1.309999943	1.50999999	1.659999967	1.340000033	1.090000033	1.220000029
ACC	25.65999985	27.63999939	27.65999985	28.17000008	26.77000046	27.29000092	28.95000076
ACCO	7.699999809	7.170000076	7.659999847	9.130000114	7.059999943	4.989999771	5.920000076
ACER	633.333313	643.333313	696.666687	750	586.666687	480	500
ACGL	7.948888779	8.220000267	8.472222328	8.397777557	8.168889046	8.277777672	8.695555687
ACH	24.80999947	24.10000038	25.73999977	24.26000023	21	18.65999985	22.15999985
ACHC	4.880000114	4.960000038	5.119999886	5.079999924	4.679999828	4.599999905	4.559999943

Figure 1.0

In total there are 2935 stocks and 144 data points per stock, 1 data point for each month. In order to process the data, we take the difference of the next month and the current month, then divide that amount by the previous month.

$$\%Change_i = \frac{month_i - month_{i-1}}{month_{i-1}}$$

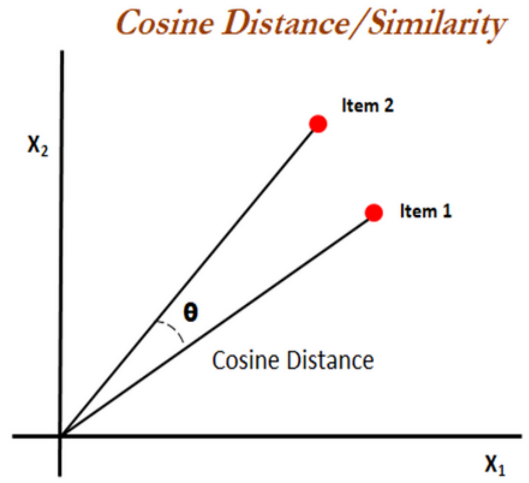
After computing this month-to-month percentage change, we have 143 columns, since the first month has no previous month to be compared to. Figure 1.1 shows this processed data set. In addition to transforming the raw data, we added a *Sector* column that classifies each stock into a sector. These sector-to-stock mappings were obtained from BarCharts [7], there are a total of 104 unique sectors, the full list can be found in the *sectorComp.ipynb* file located in our repository. The largest among these economic sectors are ETFs which account for 10.3% of all the stocks in our data set. Followed by banks at 8.5% , biotechnology at 4.31% , and Oil&Gas at 4.25%



Stock	Sector	2010-02-01 00:00:00	2010-03-01 00:00:00	2010-04-01 00:00:00	2010-05-01 00:00:00
AA	Metals & Mining	0.044776086	0.070676679	-0.056882029	-0.133283657
AACG	Diversified Consumer Services	0.266666651	0.060526384	-0.124069532	-0.079320106
AAIC	Mortgage Real Estate Investment Trust...	0.200662166	-0.017098703	0.131313143	-0.057539676
AAL	Airlines	0.380414313	0.00272851	-0.038095202	0.248939138
AAME	Insurance	-0.028776951	0.096296291	0.209459418	-0.18994409
AAON	Building Products	0.021855248	0.075094993	0.06763941	0.021946132
AAP	Specialty Retail	0.034220493	0.027450955	0.075858789	0.147671842
AAPL	Technology Hardware, Storage & Periph...	0.065396197	0.148470277	0.111021382	-0.016124719
AATC	Scientific & Technical Instruments	0.0060606	-0.015813256	0.032899795	-0.04814812
AAU	Metals & Mining	0.080459762	-0.031914864	0.285714202	-0.153846114
AAWW	Air Freight & Logistics	0.229342898	0.176796741	0.041847337	-0.054278994
AB	Capital Markets	0.050893531	0.133456588	0.023809509	-0.09557184
ABB	Electrical Equipment	0.123682721	0.077986175	-0.122710636	-0.112212925
ABC	Health Care Providers & Services	0.028613378	0.031383707	0.066735833	0.013938422
ABCB	Banks	0.017130579	-0.042161979	0.232558193	0.012578625
ABEO	Biotechnology	-0.039426523	-0.063432836	0.051792829	-0.109848485
ABEV	Beverages	0.049362303	-0.049957486	0.079196545	-0.023829279
ABG	Specialty Retail	0.050587212	0.143594159	0.16917293	-0.149839221
ABIO	Biotechnology	0.017064758	0.828859242	-0.040367046	-0.22562138
ABM	Commercial Services & Supplies	0.054582876	0.03515631	0.013679198	-0.000930687
ABMD	Health Care Equipment & Supplies	0.276515096	0.020771517	-0.065891412	0.011410752
ABR	Mortgage Real Estate Investment Trust...	0.139999986	0.421052654	0.259259232	-0.08823527
ABST	Software-Application	0.014184384	-0.012237709	-0.270796493	-0.026698947

Figure 1.1

After the data set in Figure 1.1 was obtained, we moved on to calculating the links of our network. To calculate the links, we used cosine similarity, which takes the dot product of two given vectors and divides it by the product of their euclidean distances. We drew inspiration from Coronnello's sector similarity paper[6], in which they used Pearson correlation. One thing to note is that these values all fall in the range $[-1, 1]$. A value of -1 denotes two stocks that are negatively correlated, meaning that one goes up as the other goes down. A value of 1 denotes two stocks that are strongly correlated, meaning their price changes are nearly identical. Values at or close to 0 represent links with very weak or no association.



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Figure 2

This computation was carried out for all pairs of stocks in our data set, using entire rows as input vectors. These links were then stored in */data/Raw_links.xlsx*. Since we are technically dealing with a complete network that has 2935 nodes, and we know the graph is undirected, the number of links can be calculated with:

$$L = \frac{n*(n-1)}{2} = \frac{2935*2934}{2} = 4305645$$

In total there are 4,305,645 links, this is an extremely large number and placing this many links in a network far exceeds our computational limits. In order to extract the meaningful links and reduce the computational strain of plotting over 4 million edges, we sorted the raw links by weight and took only the top 100,000. These filtered links can be found in the *filtered_links/* folder.

We are prepared to build our network now that all the links have been calculated and we have our list of stocks with their respective sectors. In Gephi, we import two separate csv files, the first being *stocks.csv*. This contains all the individual stocks along with their unique ID & sector and will serve as the nodes of our network. After that, we import *data/Filtered_links/links_total.csv* as our edge list. This yields our resulting network which can be seen in Figure 4, under the Network Visualization section.

In order to answer our first and second research questions, we took the filtered data described above and imported it into a Networkx graph. Once the graph was created, we used the built-in degree() method to obtain the degree of all nodes in the Graph, which is simply the sum of all the edge weights which go through a node. The exact formula implemented by the degree() method is:

$$k_i = \sum_j a_{ij}$$

Where k_i represents the degree of node i, and A is the adjacency matrix of our graph G. One thing to note is that since our network is undirected, there is no in-degree or out-degree, there is only a single-degree value. Additionally, the edge weights in our network take on a range of values from $[-1, 1]$, so degrees can take on non-integer values. Our weights will be converted to absolute weights since negative values mean that a node negatively impacts another node. In this way, both negative and positive impacts and the extent of the impact will be considered. Hence the degrees will take on a range of values from $[0, 1]$. Once the degree of each node is computed, we sort the nodes by degree value and obtain the top and bottom nodes for further inspection.

Similar to degree centrality, we will also repeat the above process, but with Eigenvector centrality. Eigenvector centrality is similar to degree centrality, with the addition that connections to high-scoring nodes contribute more to the score than equal connections to low-scoring nodes[4]. Rather than making our own algorithm for this, we used networkx's built-in eigenvector_centrality() method which implements the following formula:

$$Ax = \lambda x$$

Where A is the adjacency matrix filled with absolute weights of our graph G, with eigenvalue λ [8]. In a similar fashion to degree centrality, once our calculations of eigenvector centrality are done, we take the top and bottom scoring nodes to inspect further.

Stocks which are more connected to other market movers are more powerful compared to stocks that are connected to less influential nodes. That's why we expect to see better results with eigenvector centrality compared to degree centrality.

In order to answer our third and fourth research questions, we need a way to calculate similarity within, and between communities. Thankfully we are armed with average linkage clustering, which is a formula to determine *The similarity between two communities*[4]. Average linkage clustering is quite simple, we take the average of all the links between two communities. Figure 3 illustrates this formula. Likewise, we will calculate intra-sector similarity by applying average linkage clustering to a single community, in which we take the average of all the links within that community alone.

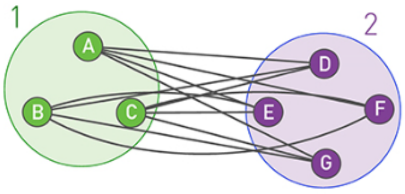


Figure 3

Network Visualization

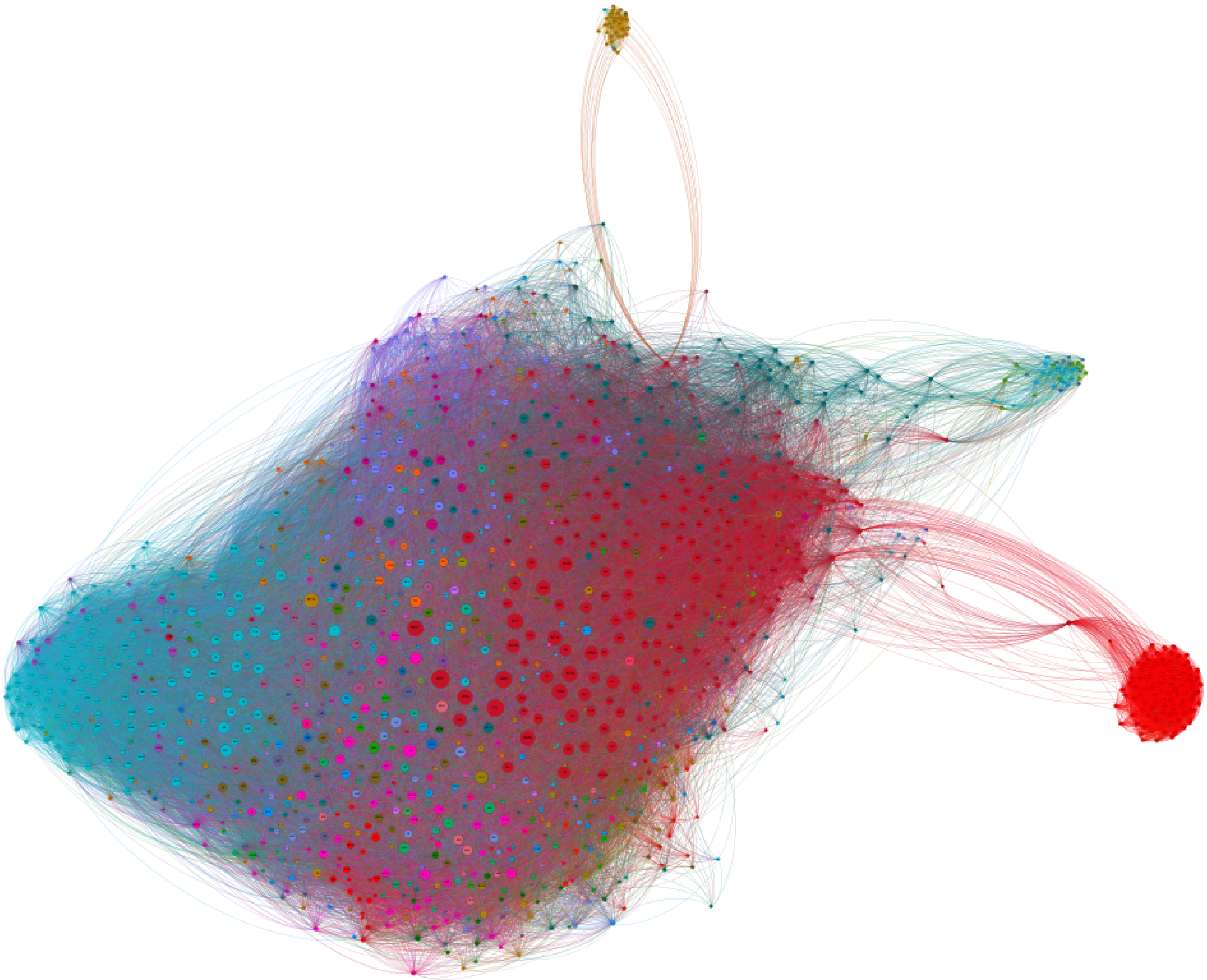
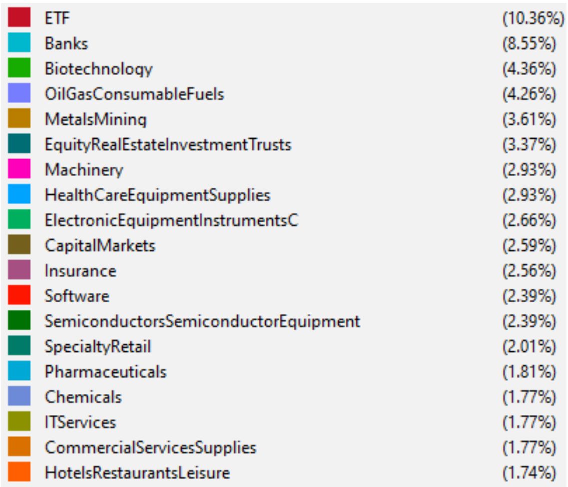


Figure 4.1 . Network for data spanning the entire data set (2010-2020)



🟥	ETF	(10.36%)
🟦	Banks	(8.55%)
🟩	Biotechnology	(4.36%)
🟪	OilGasConsumableFuels	(4.26%)
🟫	MetalsMining	(3.61%)
🟦	EquityRealEstateInvestmentTrusts	(3.37%)
🟪	Machinery	(2.93%)
🟦	HealthCareEquipmentSupplies	(2.93%)
🟩	ElectronicEquipmentInstrumentsC	(2.66%)
🟫	CapitalMarkets	(2.59%)
🟪	Insurance	(2.56%)
🟥	Software	(2.39%)
🟩	SemiconductorsSemiconductorEquipment	(2.39%)
🟩	SpecialtyRetail	(2.01%)
🟦	Pharmaceuticals	(1.81%)
🟦	Chemicals	(1.77%)
🟨	ITServices	(1.77%)
🟧	CommercialServicesSupplies	(1.77%)
🟧	HotelsRestaurantsLeisure	(1.74%)

Figure 4.2 . Color map for our network diagrams

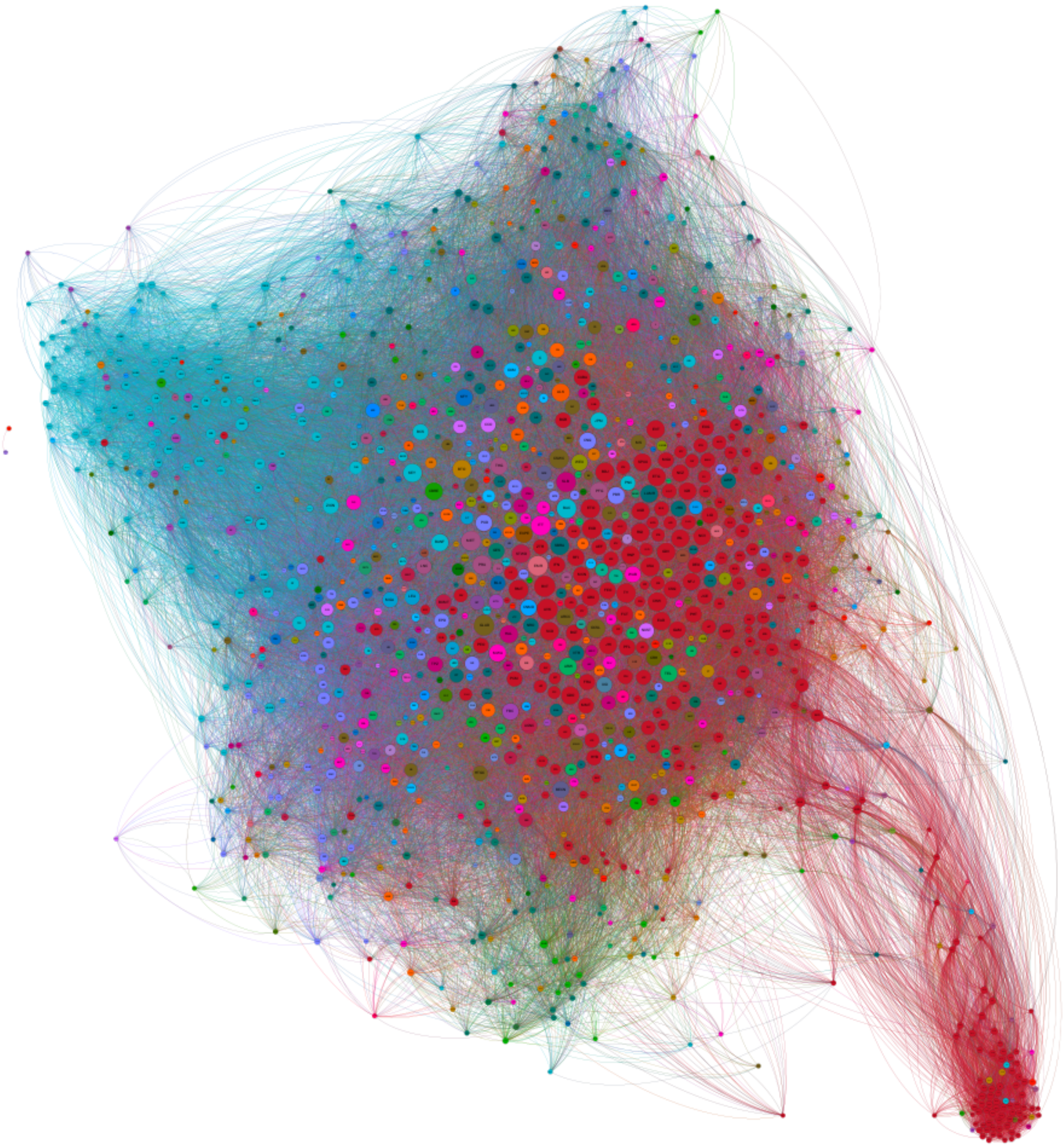


Figure 4.3. Network for 2020 data

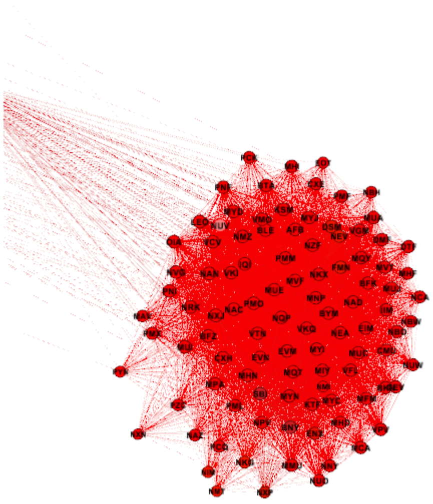


Figure 4.4 . Stand-alone Community from Figure 4.1

Figure 4.1 and 4.3 depict two network diagrams, the first being the links calculated across our entire data set (2010-2020) and the second being the links calculated using only data from 2020. Gephi was the software of choice when it came to constructing these diagrams. There were several useful features that allowed us to extract more meaning from the graphs. The first of many useful features was the ability to color by sector. Figure 4.2 shows the color mapping, and this allows us to recognize key sectors. The dark reds are ETFs and other investment funds, this sector made up a large portion of the graph in both our diagrams. The large turquoise sector represents Banks, which again was a big segment of our graph. Another tool that allowed us to better inspect the graph was node resizing. Within Gephi we had the ability to resize nodes based on their degree, this made stocks with higher degrees much more pronounced in our networks.

The reasoning behind plotting two graphs was to ensure that our results were not random. There are some striking similarities between the two generated graphs, one of which is the general structure. Both graphs have a large central component, with 1 or smaller communities on the outskirts. One notable community that stands out in both graphs is the red chunk near the right-hand side, see Figure 4.4. Upon inspection, these nodes were all found to be ETFs. Yet they don't have strong bonds to the rest of the ETF sector. This is suggesting that there is a smaller sector within the ETF sector that all of these stocks are a part of. In addition to the structural similarities between Figure 4.1 and 4.4, we also see that many of the large nodes in one graph are also quite large in the other. This tells us that there is some level of consistency between the 10-year and 1-year networks.

Basic Statistics

The below table shows the basic statistics for the graph constructed from 2010-2020. For this statistic, we assumed each link has weight=1. The average clustering coefficient and average path length are calculated for the largest connected component. The Average degree, links, and nodes are calculated for the entire graph.

Table 1

Nodes	Links	Connected components	Average Degree	Average clustering coefficient	Average Path length
1898	100000	42	105.37	0.667	2.728

In Figure 5, the degree distribution plot is shown on a logarithmic scale, both on the x-axis and y-axis. This degree distribution is for the data from 2010-2020 and the weights are between $[0, 1]$ (Absolute weights). There are a total number of 20 bins in the plot which means there are 20 logarithmically spaced bins between the smallest and biggest degree. The green diagonal line is the best fit for power low distribution with gamma 1.28.

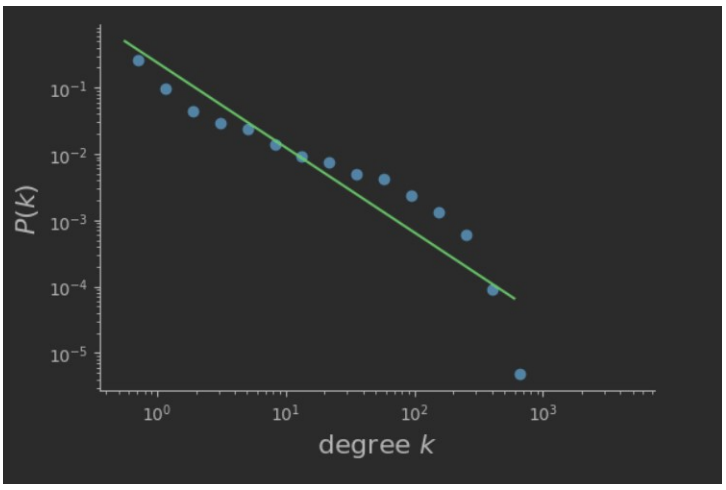


Figure 5

Figure 6 shows the degree distribution for each year from 2010 to 2020. Each year's degree distribution is differentiated by its own color shown in the plot. Like the previous Figure, The graph weights are absolute and there are 20 bins in total.

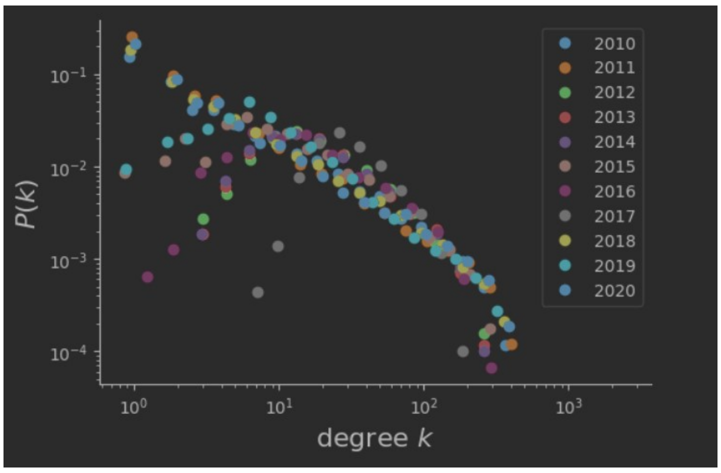


Figure 6

The green line in Figure 5 shows that our graph has a closer structure to scale-free networks rather than random networks.

Results

Q1 & Q2) What are the most and the least influential stocks in our network?

To answer these questions, it's reasonable to think that stocks with a higher degree centrality have more influence based on the things we said in the methods section. Therefore, degree centrality seems a good method to determine degree importance.

By calculating the degree centrality of each node, sorting them, and choosing the top and bottom 100 stocks we came to these results:

Table 2

	Famous Stocks	Most frequent Sectors	Most important Sectors
Top Stocks (2010-2020)	BMO (Bank of Montreal)	ETF(46) , Banks (14)	ETF(15.13%), Electrical Equipment(10.71%)
Bottom Stocks(2010-2020)	AZN(AstraZeneca), NVDA(Nvidia), Ebay(eBay)	Banks (9), Biotechnology (8)	Internet & Direct Marketing Retail(20%), Wireless Telecommunication Services(14.28%)
Top Stocks (2020)	HP(Hewlett-Packard), JPM(JPMorgan Chase)	ETF(49) , Banks (7)	Mortgage Real Estate Investment Trust(18.75%), ETF(16.11%)
Bottom Stocks(2020)	NVDA(Nvidia), AMZN(Amazon)	Biotechnology (12), Banks (6)	Gas Utilities (25%), Life Sciences Tools & services (15.78%)

We compared the Stock graph constructed from the whole period (2010-2020) with the Stock graph only from the year 2020 to see their differences. In the above table, the Famous stocks column are some of the famous stocks that almost everybody is familiar with. Companies like Nvidia and eBay are examples. In the table, Information is based on the top or bottom 100 stocks that are sorted by the degree centrality of nodes. The most frequent sectors are the sectors that are repeated the most in these 100 data. The most important Sectors are the Sector with the highest fraction in these 100 data. Sector fraction is a more accurate criterion than Sector frequency to analyze the data. For example, we have a total of 251 banks in whole stocks. In row 2 of the above table, although we have the highly frequent sector, Banks, with 9 occurrences, The fraction will be low (9/251 = 3.5%) which means only 3.5% of Banks are represented in these 100 bottom stocks. On the other hand, Internet & Direct Marketing Retail is less frequent but 20% of them are represented in these 100 bottom data which makes this sector more important than banks.

Note that we only considered sectors that had at least a size of 10 in the whole data.

The key results of the above table are:

1. In both Top Stocks 2020 and Top Stocks 2010-2020, ETFs are playing an important role.
2. Biotechnology Sector is highly frequent in both Bottom Stocks 2020 and Bottom Stocks 2010-2020.
3. AstraZeneca company is one of the least influential companies from 2010-2020 which makes sense because pharmaceutical companies are not usually following the trend of other stocks.
4. HP company is among the top influential stocks in 2020 which is a little strange since software companies are not known as market movers.

According to the things we discussed in the methods section of this article, eigenvector centrality seems a better approach for answering research questions 1 and 2. We will calculate the eigenvector centrality and then compare it to the degree centrality.

By calculating the eigenvector centrality of each node, sorting them, and choosing the top and bottom 100 stocks we came to these results:

Table 3

	Famous Stocks	Most frequent Sectors	Most important Sectors
Top Stocks (2010-2020)	BMO (Bank of Montreal)	ETF(47) , Banks (13)	ETF(15.46%), Insurance(9.3%)
Bottom Stocks(2010-2020)	EA (Electronic Arts)	Biotechnology (16), Health Care Providers & Services(11)	Health Care Providers & Services(23.4%), Food Products(21.73%)
Top Stocks (2020)	JPM(JPMorgan Chase)	ETF(54) , Banks (5)	Mortgage Real Estate Investment Trust(18.75%), ETF(17.76%)
Bottom Stocks(2020)	NVDA(Nvidia)	Biotechnology (17), Health Care Equipment & Supplies(6)	Water Utilities(15.38%), Biotechnology(13.28%)

With degree centrality results in mind, The key results of the above table are:

1. Again, in both Top Stocks 2020 and Top Stocks 2010-2020, ETFs are playing an important role.
2. Again, the Biotechnology sector is highly frequent in both Bottom Stocks 2020 and Bottom Stocks 2010-2020.
3. Health Care Providers & Services have a significant impact on bottom stocks 2010-2020.
4. HP is removed from the Top Stocks 2020 which is more desirable.

Although we expected to see oil and gas as the most important sector, It was the third important sector in Top Stocks(2020) and for Top Stocks (2010 to 2020) there were no oil companies among the top 100 stocks. Our theories as to why this is will be mentioned in the discussion section.

Q3) How are stocks within the same sector correlated to each other? Does this differ between sectors?

To obtain the results for this question we applied the previously mentioned average linkage clustering formula within each sector. The tables below are the top 10 and bottom 10 results for the entirety of the data set (2010 to 2020).

Table 4

Sector	Average Correlation
Mortgage Real Estate Investment Trust...(16)	0.583776452
Multi-Utilities(18)	0.540964664
Containers & Packaging(12)	0.486124604
Industrial Conglomerates(6)	0.480197185
Road & Rail(22)	0.478792203
Airlines(14)	0.450195247
Construction Materials(8)	0.448189502
Equity Real Estate Investment Trusts ...(98)	0.435802967
Electric Utilities(28)	0.429634785
Energy Equipment & Services(38)	0.422664845

Table 5

Sector	Average Correlation
Biotechnology(128)	0.115406
Pharmaceuticals(53)	0.123715
Personal Products(19)	0.141648
Entertainment(22)	0.16131
Food & Staples Retailing(16)	0.163015
Food Products(46)	0.165864
Real Estate Management & Development(23)	0.167805
Health Care Equipment & Supplies(86)	0.170972
Diversified Consumer Services(17)	0.188487
Software(70)	0.195832

These results that we obtained are both interesting, and explainable. Within the top 10 categories, we have sectors that are competitive, in the sense that they sell similar products within similar markets. Take, for instance, Mortgage Investment Trusts, which are placed first among all the sectors. The vast majority of the companies within that sector are buying and selling similar mortgage bonds and other forms of securities. So it is expected that the stocks within that sector perform similarly, which is why this sector is so highly ranked. Another great example in the top 10 would be airlines. Although airline companies try to stand out by offering incentives such as rewards programs and other in-flight luxuries, the products they sell are nearly identical to all other airline companies. Since these products are nearly identical, and these airline companies are targeting the same markets, we expect the stocks within this sector to once again be highly correlated. In addition to the previously mentioned points, the stocks within these top 10 sectors generally react to news and events in a similar manner. A prime example of this would be the early stages of Covid-19 and its effects on airline companies. When lockdowns first took place, airline companies all took massive hits to their stock prices. During this time there was not a single airline company that went against this downward trend.

On the other end of the spectrum we have the bottom 10 stocks in intra-sector correlation. The lowest of which is Biotechnology, followed closely by Pharmaceuticals. A pattern that stood out among most of these bottom 10 stocks was the fact that they have very stringent laws on intellectual property and patent protection. To give an example, let's look at Pharmaceuticals. When a pharmaceutical company makes a breakthrough drug that cures some disease or illness, that drug is heavily patented and any attempts to violate these patents are met with severe legal and financial consequences for the violating party. With this in mind, the company that originally came up with this miraculous new drug reaps all the profits from that particular drug, whereas all the other competitors are left stranded. This results in a great disparity among the stocks in the Pharmaceutical sector and results in a low correlation value. Another sector on the bottom list that follows this trend is entertainment. Since there are very strict copyright laws on things like tv shows and movies, not all of the stocks within the sector can flourish to the same degree. If a large entertainment company such as Marvel or Disney comes out with a blockbuster movie, they are going to be the sole recipient of all the profits that the movie generates. Another entertainment company can't just come along and copy this movie in the hopes of turning a profit, so the remaining stocks are once again left in the dust, causing a large disparity within the stocks of the sector.

In addition to calculating the intra-sector correlation values for the entirety of the data set, we also did it for each year from 2010 to 2020. The results for each year can be found in the */data/-sector_comp* folder. We parsed through this data and found that mortgage trusts, multi-utility, airlines, packaging, oil & Gas, and the road & rail sector were almost always present. Similarly, the bottom 20 ranking stocks in terms of intra-sector correlation almost always included biotechnology, pharmaceuticals, entertainment, and food products for each fiscal year. This indicates that the data we are obtaining is consistent year over year and reaffirms our theories as to why these results are true. Lastly, to better grasp the concept of intra-sector correlation, we made a plot that includes all the major sectors within our data set, along with their respective intra-correlation values. See Figure 7 for details.

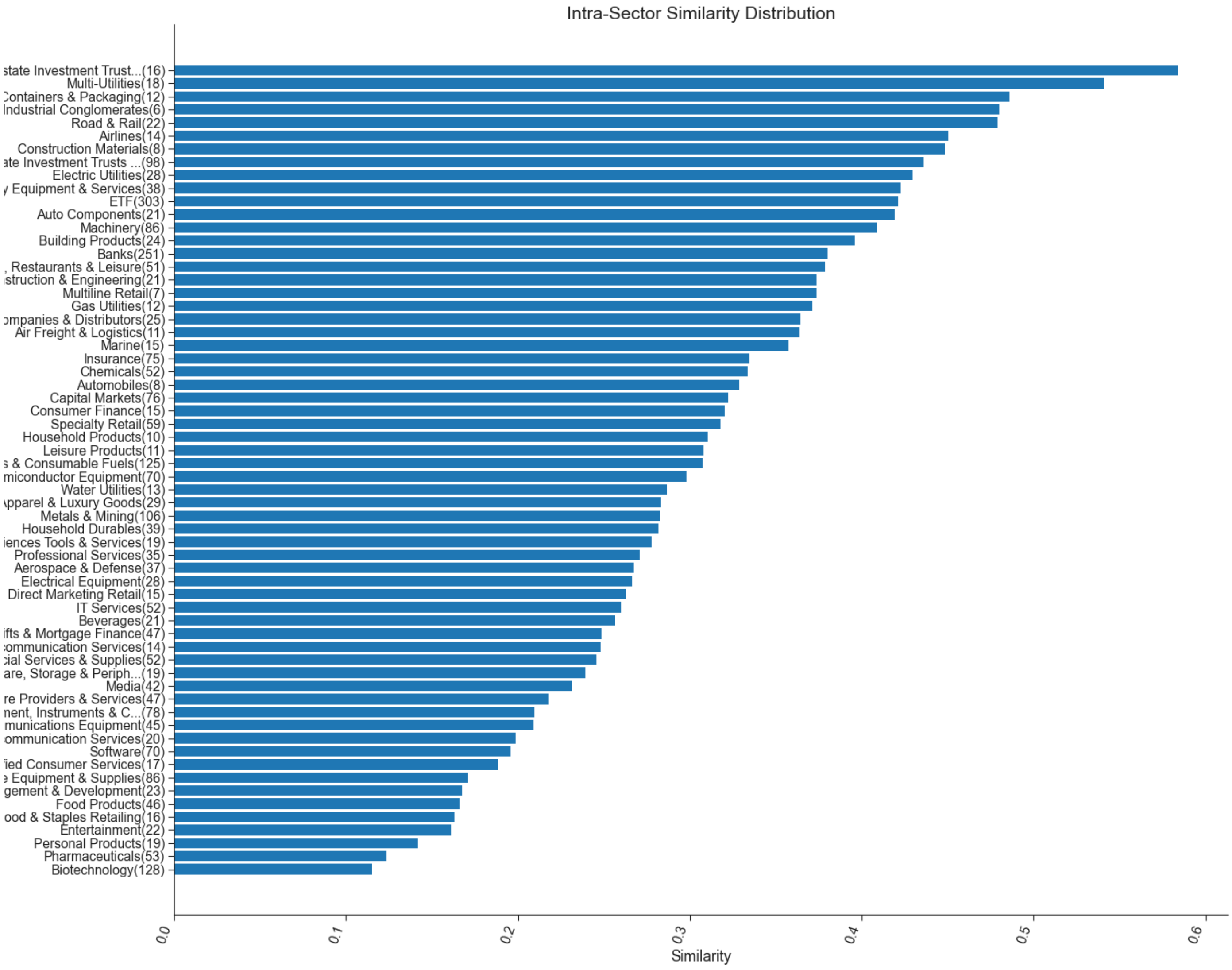


Figure 7

Q4) How are stocks in different sectors related to each other? What are some of the strongest sector-to-sector correlations?

The results for this question were obtained in a similar fashion to Q3, however this time we calculated the average linkage clustering between two sectors rather than within a single sector. This yielded the following results:

Table 6

Sector1	Sector2	Average Correlation
Electric Utilities(28)	Multi-Utilities(18)	0.476321421
Machinery(86)	Industrial Conglomerates(6)	0.430563433
Road & Rail(22)	Industrial Conglomerates(6)	0.424757762
Machinery(86)	Containers & Packaging(12)	0.406432882
Multi-Utilities(18)	Gas Utilities(12)	0.405279226
Building Products(24)	Industrial Conglomerates(6)	0.404283152
Auto Components(21)	Industrial Conglomerates(6)	0.403522184
Trading Companies & Distributors(25)	Industrial Conglomerates(6)	0.402016684
Equity Real Estate Investment Trusts ...(98)	Mortgage Real Estate Investment Trust...(16)	0.398903767
Road & Rail(22)	Air Freight & Logistics(11)	0.396712476

These results are slightly more difficult to interpret compared to the previous question, it requires a solid understanding of economics and a background in some of these industries to understand why these results are generated. With that said we can look at some of the more obvious correlations such as the very top spot, which has electrical utilities paired with multi-utilities at a correlation of 0.476. This can be explained by the simple fact that multi-utility companies offer electrical utilities as one of their services. So as the stocks for electrical utility companies rise, so do the multi-utility companies, as they offer similar services. There are several examples that are not in the top 10 but still rank very high relative to the rest of the pack. Among them is the pair between Hotels and Airlines, which has a correlation of 0.336. This is quite easily explained by the fact that people generally buy products from airline companies and hotels simultaneously. Whether it be a vacation or a business trip, booking an airline ticket and making a hotel reservation are almost synonymous. Now there are even websites that offer travel packages, which include flight and stay.

In addition to looking at particular pairings and trying to come up with coherent theories on their ranking, it's also important to zoom out and look at the big-picture influence of each sector. Figure 7 shows exactly that. Along the x-axis, each of the major sectors is listed, and for each of these sectors, there are a series of dots in the graph that accompany it. Each dot represents another sector that it correlates to, and one thing to note is that the higher the position of the dot, the greater the inter-sector correlation. Near the left-hand side of Figure 7, we see a few highly ranked sectors, including machinery, road & rail, electric utilities, and ETFs. These stocks have higher average inter-sector correlations, whereas the stocks near the right-hand side have much lower sector-to-sector correlations. On the right-hand side, we have pharmaceuticals, biotechnology, leisure products, and personal products. What this tells us is that stocks further to the left are more likely to affect the movement of the entire market. Whereas stocks near the right have significantly less impact in determining the outlook of the rest of the market.

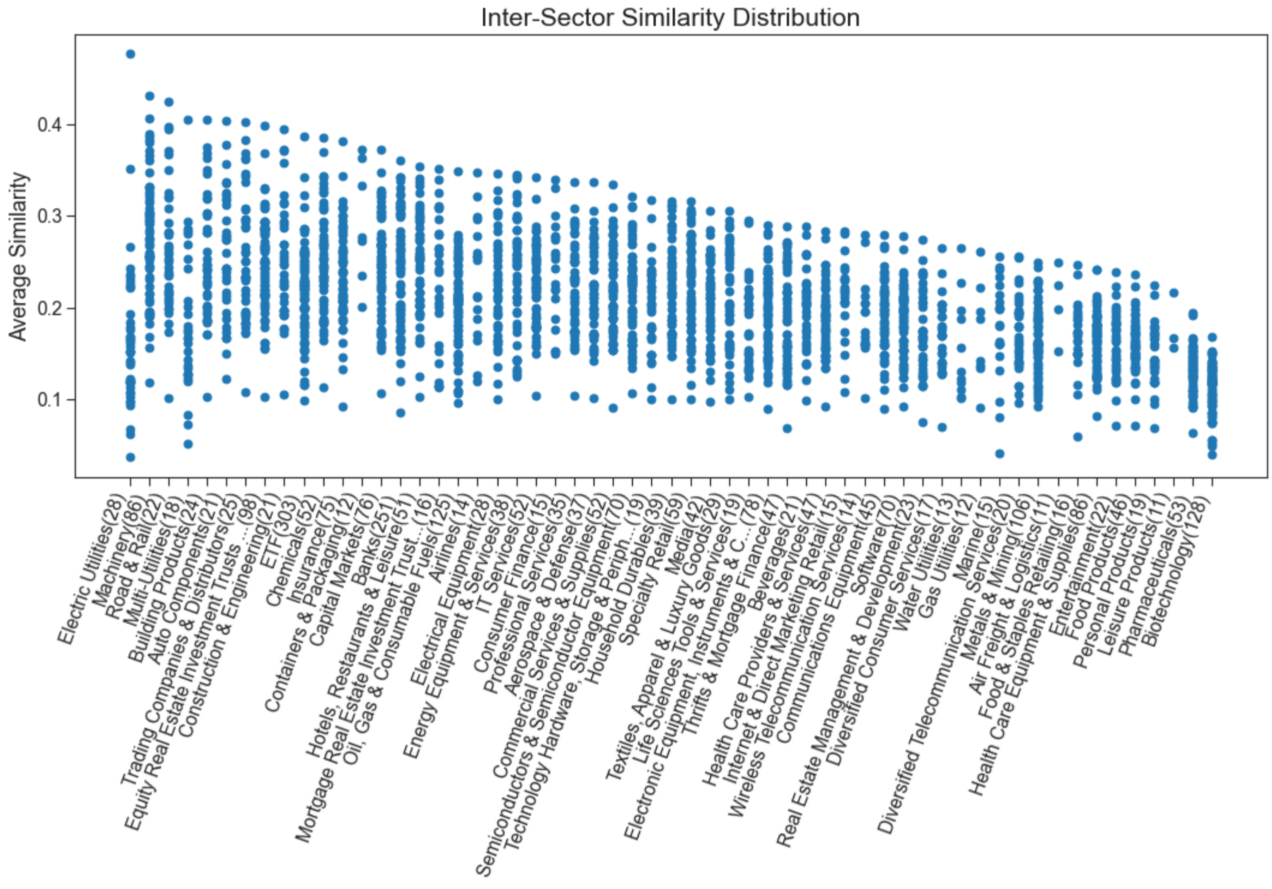


Figure 7

Comparison to Null Models

For comparison to suitable null models, we compared the Stocks graph to both Erdos-Renyi random network and degree preserved random network. We created an ensemble of 1000 networks for each of these random networks.

For creating the Erdon-Renyi network, we used the same number of nodes, and the probability of link addition is calculated by dividing the number of our edges (100000) by the number of edges if our graph was a complete graph.

For creating the Degree preserved network, 100000 edge swapping were done on our graph which is equal to the number of links.

Note that in this comparison, we did not consider the weights of our graph which means each link will have a weight of 1.

Table 7

Attribute	Stocks Graph	ER	DP
Average Clustering coefficient	0.667	0.062±0.0002	0.428±0.003
Average shortest path	2.728	1.938±0.0002	2.189±0.0002

The numbers in the table above show that the Stocks graph has the highest average clustering coefficient and average shortest path among the others. This means that nodes in our graph tend to cluster together more often than the other null models. Also, the nodes in our graph are further away from each other compared to other null models.

Discussion

In this project, we first created a graph in which each node represented a stock from NASDAQ or NYSE stock exchange. Links were defined as cosine similarity between each pair of stocks. After data cleaning and size reduction we created a graph from 2010 to 2020 and also 10 yearly graphs for each year between 2010 to 2020.

We used eigenvector centrality and degree centrality in order to answer question research 1 and 2. For questions 3 and 4, we used the average linkage clustering method to both realize the inter and intra-sector correlations.

There were many outputs that aligned with our expectations. We saw that eigenvector centrality could give better choices of top and bottom networks compared to degree centrality. Also, these network measures gave us many good insights like the most influential sectors are banks and ETFs and the sector with the least influence is Biotechnology. These results were not far from our expectations. However, we anticipated to see Oil gas sectors be the most important ones, but this wasn't the case in our results.

To Analyze the reason behind this, we should first consider several facts. Stocks are hard to model. So many factors can impact the stocks such as global news, famous people, economic features, etc. By considering only price values, we will miss some important features from other sources. Secondly, our model calculates correlation similarities based on cosine similarity. There are plenty of known and unknown mathematical models that can model these similarities in a better way. The final reason is that we only used monthly data which is collected on the first day of each month. For a better result, we can use daily or even hourly data which can show the price fluctuations more accurately.

For our third research question, we found that the sectors with the highest intra-sector correlation were mortgage trusts, multi-utilities, and packaging. The 3 bottom ranking sectors were biotechnology, pharmaceuticals, and personal products. These results were almost directly in line with our predictions. The top sectors are all in competitive markets, and they all offer products that are similar. Whereas the bottom sectors are heavily reliant on patent protection and intellectual property rights. These results make sense since the stocks that rely on patent protection are able to sell unique products and reap all the profits while leaving their competitors behind. On the other hand, stocks in competitive markets don't have too many factors that can set them apart, this leads to stocks within sectors being highly correlated to one another as they share the market.

The fourth research question we set out to answer did not go as well as we hoped. We were expecting oil & gas to be one of, if not the most influential sector in our entire data set, however, this was not the case. We found Utility companies and Industrial conglomerates to be the most commonly found sectors in our top 10 inter-sector correlations. We believe that the reason for this is there are a small number of stocks within these sectors, looking at multi-utilities we have only 18 stocks within that sector. Similarly, there are only 6 stocks in the industrial conglomerate's sector, which is an extremely small number when compared to the number of stocks in the oil&gas sector which is a staggering 128. Even this pales in comparison to the number of stocks in the bank's sector, which is 251. Because some sectors have a small number of stocks within them, there is the possibility that the average linkage to these sectors is higher by random chance. In order to resolve this and get more reliable results, we can try to eliminate sectors from our data set which don't meet a minimum requirement in size. Although this research question was not as clear-cut as the rest, we were still able to dig out some valuable information from Figure 7. Namely that biotechnology and pharmaceuticals are the weakest sectors affecting the rest of the market. And some of the strongest market-moving sectors are electric utilities, machinery, road & rail, and building products. These results make far more sense, as they serve to be the backbone for most of our economy and society.

One modification that we could have made to ensure more reliable results would be to increase the granularity of our data. For this paper, we took one data point per month, per stock. With this relatively small data set, the total time spent on data processing, graph creation, calculating intra-sector and inter-sector correlation, and network measurements exceeded 20 hours. So shifting the data points from monthly data to daily data might multiply the computational complexity by up to 30x which makes it computationally infeasible for us. For future work in this area of research, it would be best to dedicate large computational capabilities in order to perform calculations on much greater scales. Addressing this issue may reveal more reliable and interesting results. In addition to this, it would be nice to consider more complex similarity calculations for computing the links between two nodes. In this paper, we used cosine distance, but perhaps there are more complex, and computationally expensive functions that can better represent the similarity of stock price changes.

Moving forward, we can build more complex machine learning algorithms that incorporate stock relationships as input features to price forecasting models. This will result in more accurate predictions, and thus more profitable outcomes.

Code

The entire code base for this project can be found on GitHub at the following link:

https://github.com/deliarm/672$_{final}$

The raw data and final data can all be found in the *data* folder, along with the python scripts that were used to obtain them. Additionally, the repository holds all the code used to develop our network and gathers statistics that were presented in this paper.

Sources

[1] - Chen, J. (2022, November 4). What is the stock market, what does it do, and how does it work? Investopedia. Retrieved November 16, 2022, from https://www.investopedia.com/terms/s/stockmarket.asp

[2] - Tian, Q., Shang, P. , Feng, G. The similarity analysis of financial stocks based on information clustering. Nonlinear Dyn 85, 2635–2652 (2016). https://doi.org/10.1007/s11071-016-2851-9

[3] - Wen, Yulian, Peiguang Lin, and Xiushan Nie. ”Research of stock price prediction based on PCA-LSTM model.” IOP Conference Series: Materials Science and Engineering. Vol. 790. No. 1. IOP Publishing, 2020.

[4] - Barabasi, A. (2016). Network Science. Albert Barabasi.

[5] - “Yfinance.” PyPI, 16 Nov. 2022, pypi.org/project/yfinance.

[6] - C.Coronnello, et al. (Aug 2005). Sector identification in a set of stock return time series traded at the London Stock Exchange. Retrieved from https://arxiv.org/pdf/cond-mat/0508122v1.pdf

[7] - Barchart.org, B. (2022). Stock market sectors &amp; stock sector finder. Barchart.com. Retrieved December 5, 2022, from https://www.barchart.com/stocks/sectors/rankings

[8] - A Hagberg, D Schult, P Swart, Exploring Network Structure, Dynamics, and Function using NetworkX in Proceedings of the 7th Python in Science conference (SciPy 2008), G Varoquaux, T Vaught, J Millman (Eds.)