Link Prediction: Inferring the Nasdaq 100 network of correlated social media chatter

Vigneshwaran Giri Velumani

This project focusses on building, describing and visualizing undirected networks of Nasdaq 100 firms' correlations in Twitter activity. Two firms are linked if there is a statistically significant correlation in the daily #tweet mentions.

To begin with initial analysis, I have shown only the nodes that are connected to at least one another node in Fig.1. The nodes(firms) are represented in dark green. The width of the edge denotes the strength of the correlation and the gradient color from firebrick to dodger blue denotes the range from negative to positive correlation respectively between the firms. The reference code[1] used for plotting the Fig.1 has been provided in the Reference section.

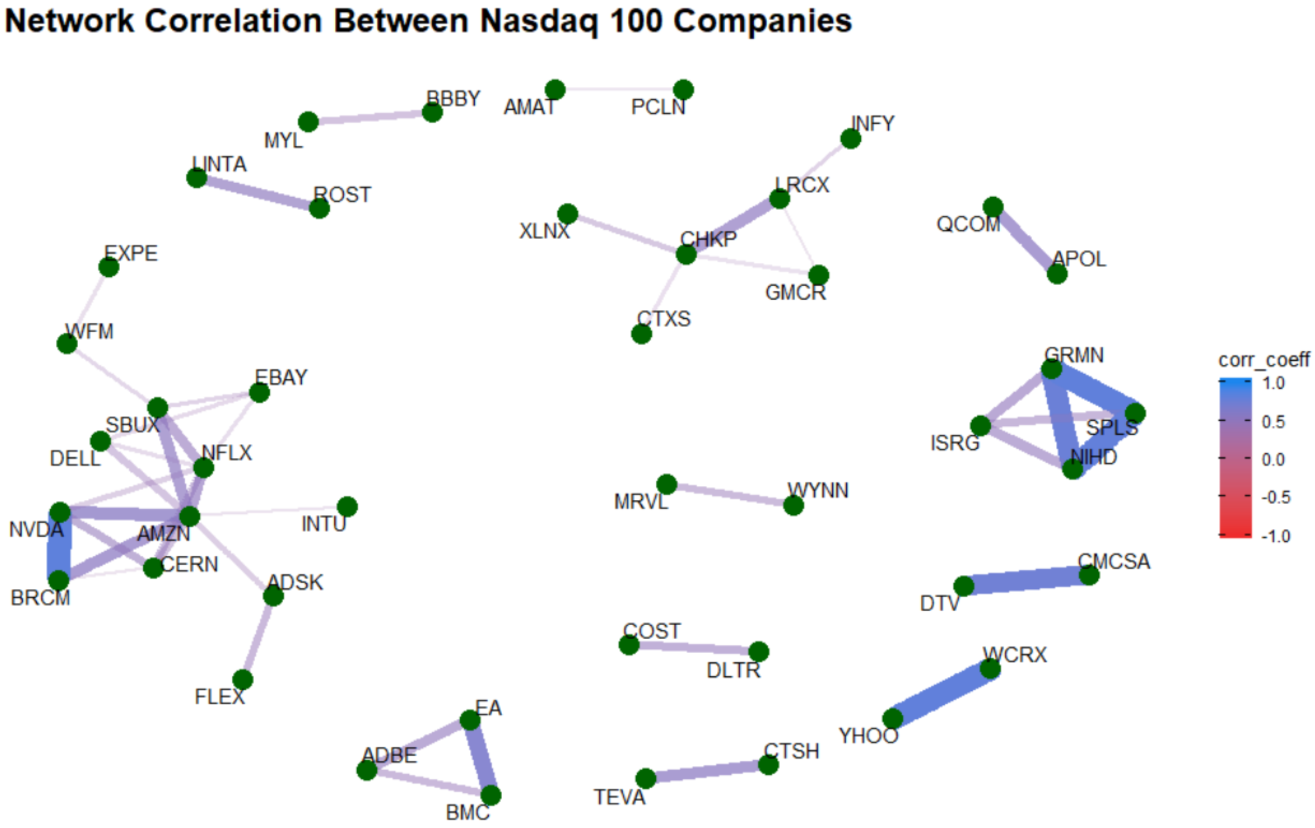

Fig.1. Network Correlation between Nasdaq 100 Companies

The above network has been visualized after filtering the nodes having the absolute correlation coefficient greater than a factor of 0.35 and has 44 Vertices and 90 Edges.

To determine the statistically significant links between the companies given in the dataset, initially I calculated the partial correlation coefficient between each node. Then I computed the Fisher's transformation to approximate the bivariate distributions and estimated the confidence intervals to obtain p-values.

Then later I applied the Benjamini-Hochberg adjustment to control for the false discovery rate; and used a threshold of $p < 0.05$ to identify statistically significant partial correlations. After performing these calculations, I determined the edges and the nodes. I used the code given in the textbook[2] chapter 7 for performing all the above steps. Later I constructed and visualized the network of firms as shown in Fig.2 after removing the nodes whose degrees are less than 1. The network contains 56 Vertices and 48 Edges.

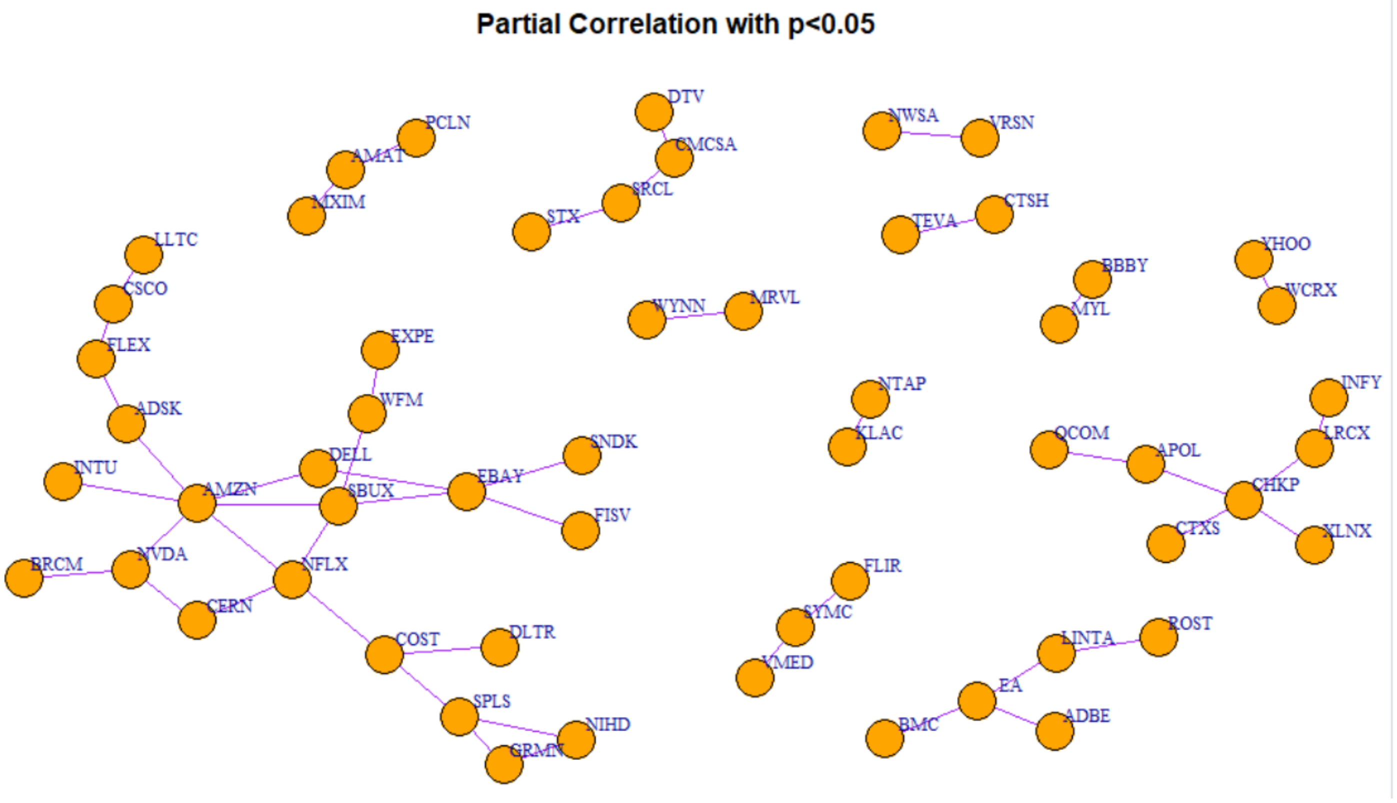

Fig.2 Partial Correlation with p<0.05

Since this network contains considerably a large network of firms with a balanced edge network, I have dived deep into finding the relationship between the firms in this partial correlation network.

If we start from the left to right, the first biggest network that we can see is

Amazon's network with 6 other firms.

Netflix's, Ebay, Starbucks & CHKP each linked with 4 other firms.

NVIDIA, EA Sports, Costco each linked to 3 other firms. Rest all are connected with one or 2 firms.

Let’s look at one by one:

- The relationship between Autodesk(ADSK) and Amazon(AMZN) looks like a cooperative relationship because the standalone software and services provided by Autodesk once has moved to the Cloud provided by Amazons AWS. To streamline development and time to market, Autodesk has been steadily expanding its use of AWS and decreasing its data-center footprint.[3]
- The relationship between Intuit(INTU) and Amazon(AMZN) looks like a cooperative relationship because this financial software company that develops and sells financial, accounting, and tax preparation software and related services uses Amazon SageMaker to Manage Machine Learning at Scale.[4]
- The relationship between Nvidia(NVDA) and Amazon(AMZN) looks like a cooperative relationship since this partnership is advancing the hybrid cloud to power modern enterprise workloads. AWS and NVIDIA have partnered to deliver the most powerful and advanced GPU-accelerated cloud to help clients build a more intelligent future.[5]
- Amazon are Netflix are in a competitive relationship since amazon’s prime video was launched to take on the streaming giant.
- Amazon – Starbucks looks like a competitive relationship since Amazon tries to take on the traditional brick & mortar coffee giant using its Amazon prime, amazon go products. First, Amazon destroyed retail --now, it's coming for Starbucks.[6]
- Amazon-Dell is a cooperative relationship since they sell their Laptop and other electronics accessories both in Amazon and eBay.
- SanDisk and eBay – cooperative relationship since they sell their products on eBay.
- Costco and Dollar Tree are in a competitive relationship since both of them run a traditional brick & mortal stores with overlaps in the products which they are selling.
- Costco and Staples are competitive since staples specializes in office supplies and Costco also has a market share in it.
- Garmin, NII Holdings and staples form a triadic closure. Since Garmin and NII holdings both are in the Telecommunications industry, they share a collaborative relationship. Since Staples sells both of their products, they all share a collaborative link.
- CHKP looks like its sharing cooperative relationship with Citrix Systems since its sharing its Citrix compatible products in Citrix’s Marketplace.
- CHKP may be sharing a competitive relationship with Xilinx and Lam Research since all these companies are in the security and network and semiconductor industry.
- AMAT and MXIM share a competitive relationship since both are in the integrated circuit and semiconductor manufacturing industry
- Starbucks partnered with Wholefoods by selling their beverage products thereby sharing a cooperative relationship. Starbucks makes a move into Whole Foods.[7]
- Comcast and DirecTV can be considered as rivals since they are competitive in the cable industry.
- NVIDIA and Broadcom can be considered as rivals since both are competitive in the Semiconductor manufacturing industry.

To gain further insight, I have tried to compare my results with an alternative threshold of statistical significance i.e. p<0.01 and built and visualized the network as shown in Fig.3. The network contains 43 Vertices and 32 Edges.

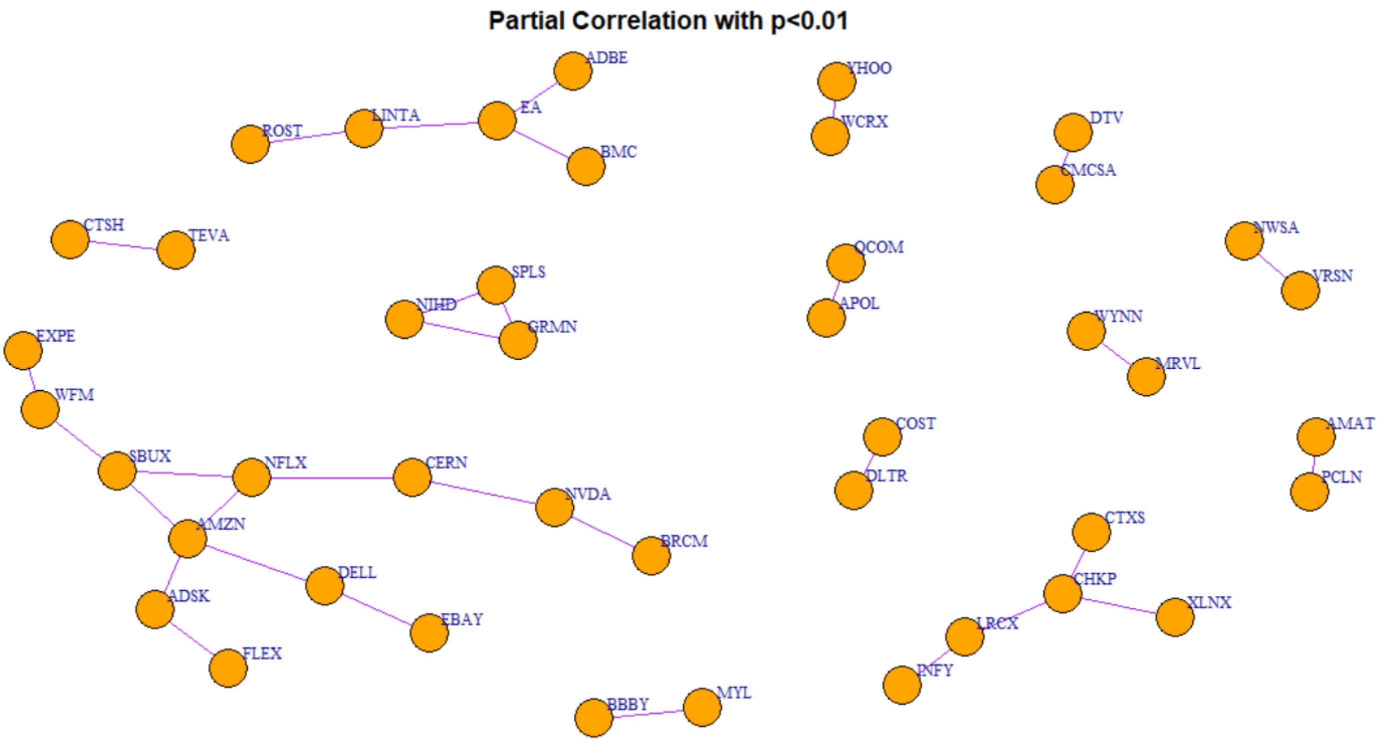

Fig.3 Partial Correlation with p<0.01

We can see the 2 links of Amazon with Nvidia and Intuit is broken. But the triadic link between Garmin, Staples and NIHD still looks intact signifying a strong cooperative relationship. Costco has broken from Netflix and Staples but shares a strong connection with Dollar Tree since they both are competitive in the retail market. The CHKP's link with APOL is broken but rest other three links are still intact. Also, the EA's chain with its neighbors also looks intact.

Later, I have also tried an alternative method for constructing the edges – the overall correlation method and used a threshold of p < 0.05 to identify statistically significant correlations. The graph of this network has been given below in Fig.4. This network has 89 Vertices and 180 Edges. From Fig.4 we can infer that using overall correlation we can see the links of all the companies in the network which are correlated to be more than the ones using partial correlation. The network also looks dense, with more companies in the network compared to the previous networks, hence showing that it's a strongly correlated network.

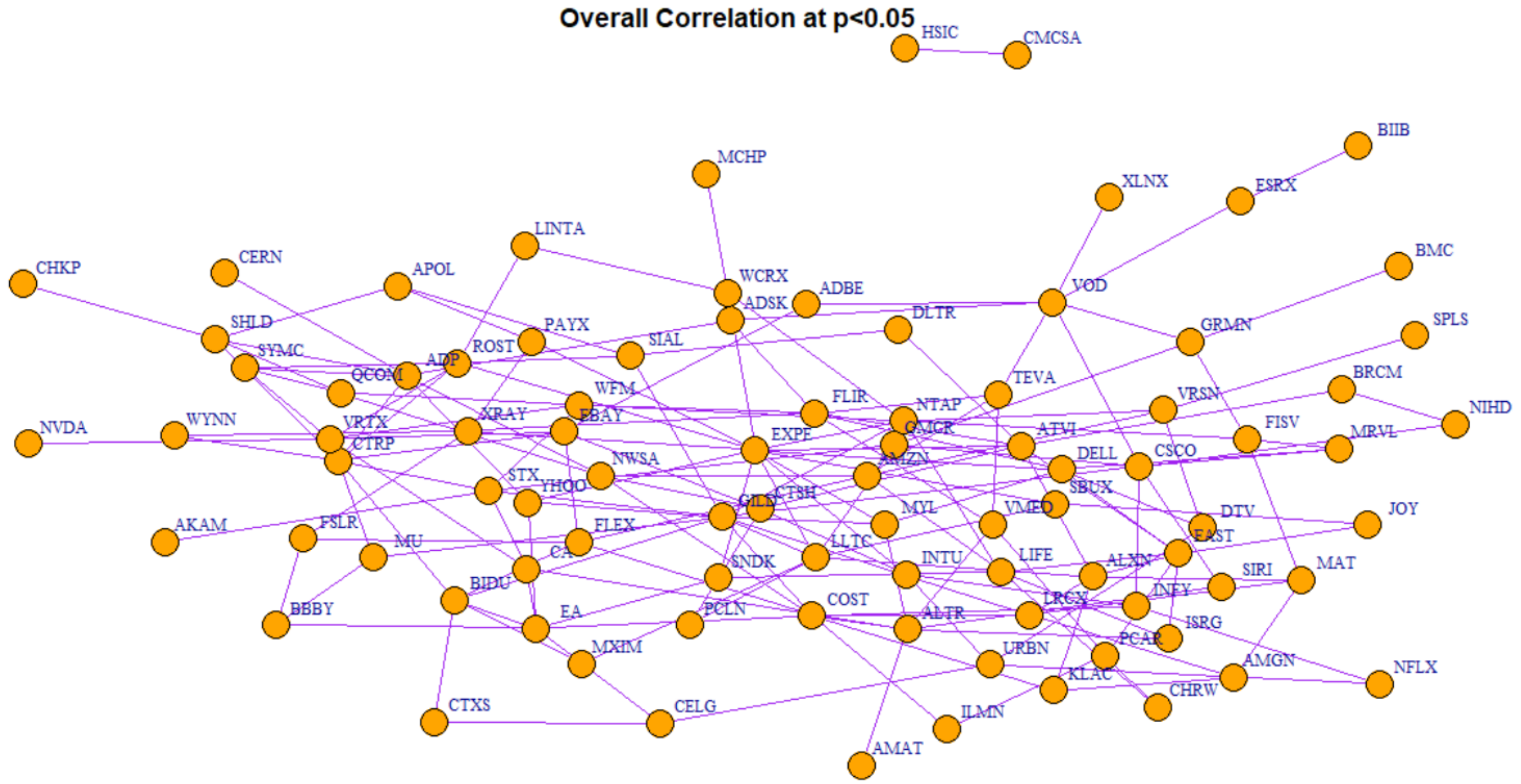

Fig.4 Overall Correlation with p<0.05

Finally, I have tried to compare my results with an alternative threshold of statistical significance i.e. p<0.01 and built and visualized the network as shown in Fig.5. The network contains 85 Vertices and 114 Edges.

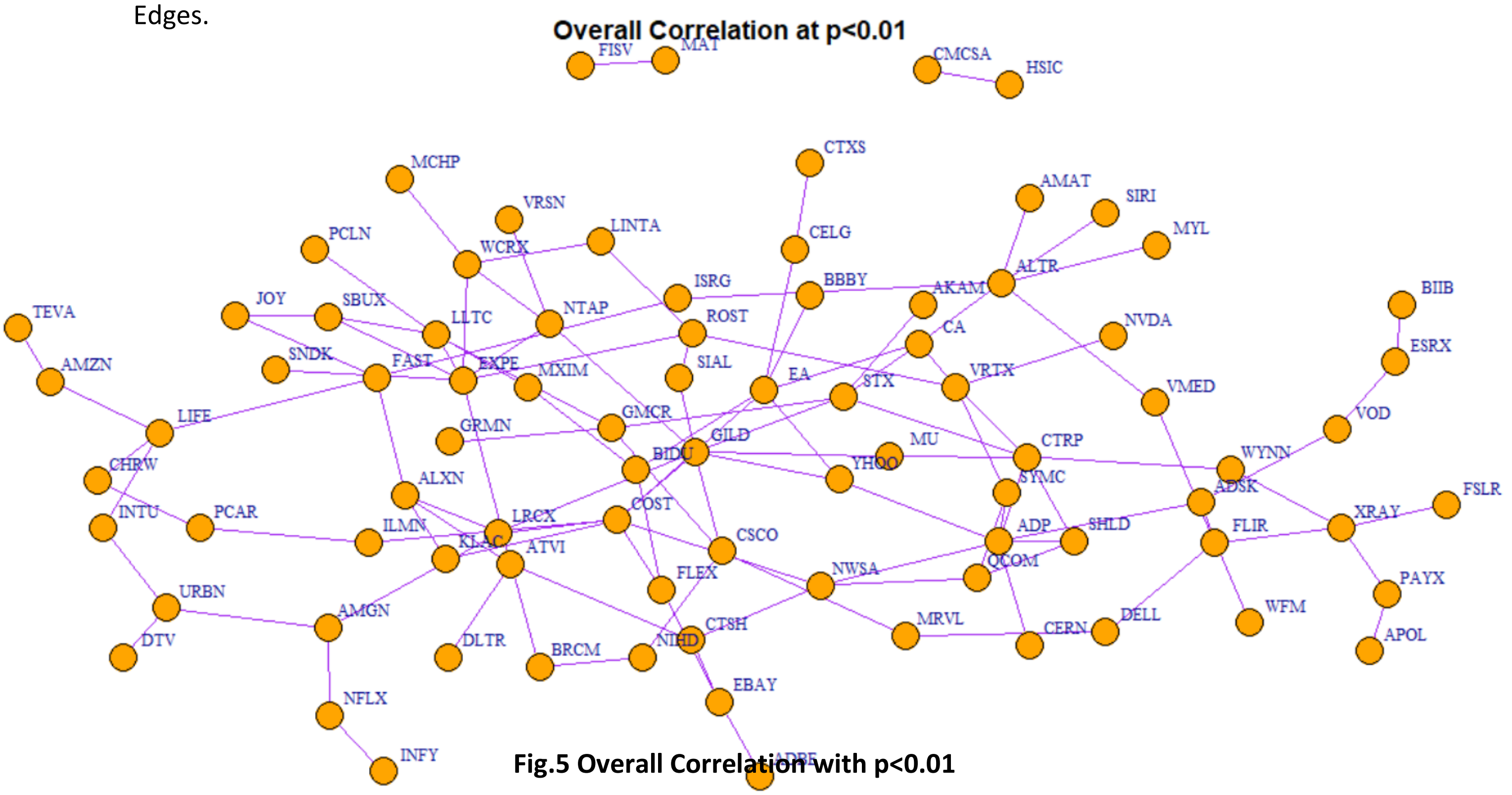

Fig.5 Overall Correlation with p<0.01

We can infer from fig.5 that with the threshold value of <0.01, the weaker connections have been removed since Edge count dropped from 180 to 114 but Vertices remaining almost the same and most of the companies are connected together as one big cluster of interconnected network.

Visualizing Network Correlation Between Companies by their Industries:

I wanted to understand and visualize how the industries were themselves correlated and so I identified each company's specific industry and reduced them to 13 broader industry categories like **Entertainment, Hardware, Hardware & Software, Healthcare, Internet, IT Consulting, Others, Retail, Semiconductors, Software, Technology, Telecom, Travel** and then created a list containing their company indexes to use for my visualization using qgraph package. To do this I referred to this qgraph example website[8].

Initially I visualized the industries Fig.6 without creating industry segmentation to see the positive and negative correlations indicated by green and red. From this visualization I was only able to understand that there are some strong positive correlations are present in the network and negative correlations don't seem to be very strong and overall the positive correlations are predominant in the network.

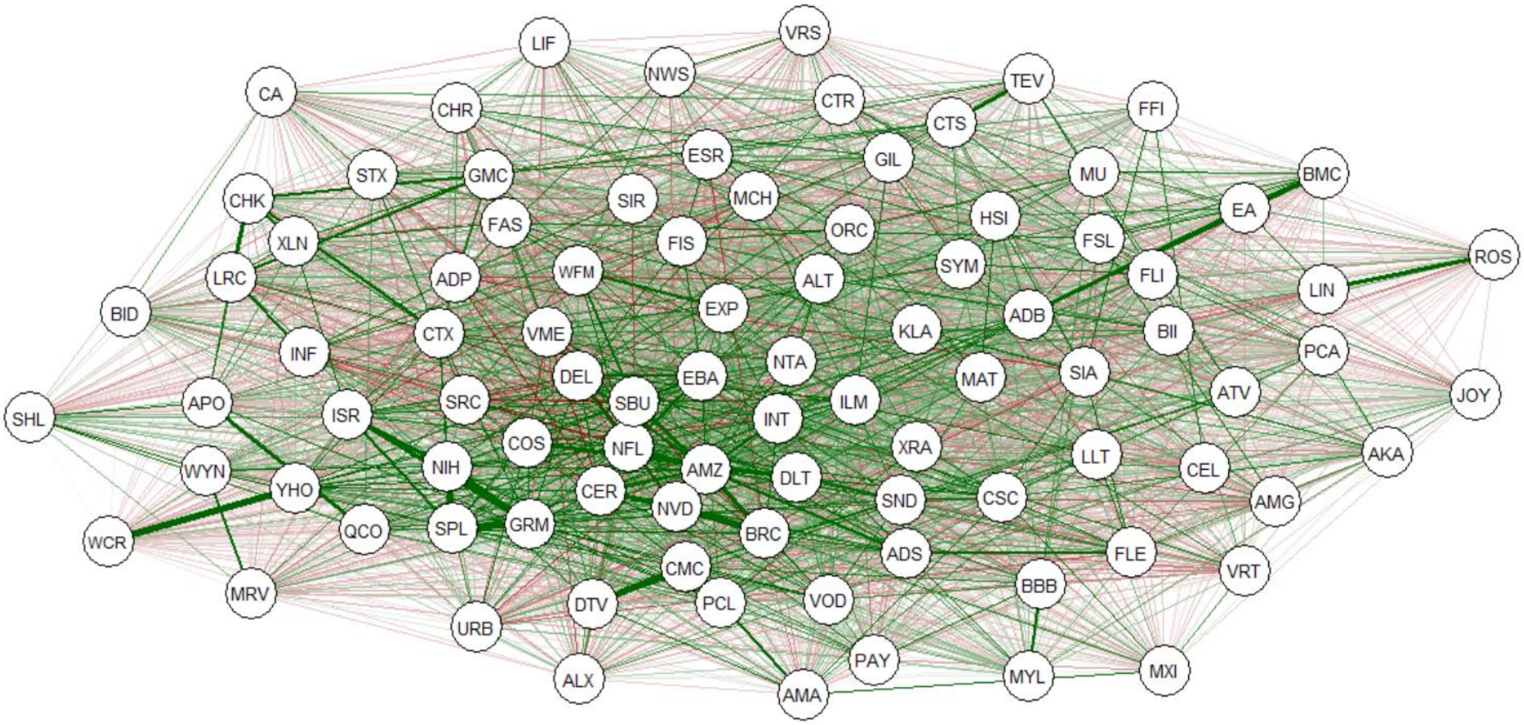

Fig.6 Correlation between companies (Without Industry Segmentation)

Later, I tried a different layout for the visualization Fig.8 using the spring layout where the companies were distributed with the color coding indicating their industries. But I was not able to infer much about the correlation between industries.

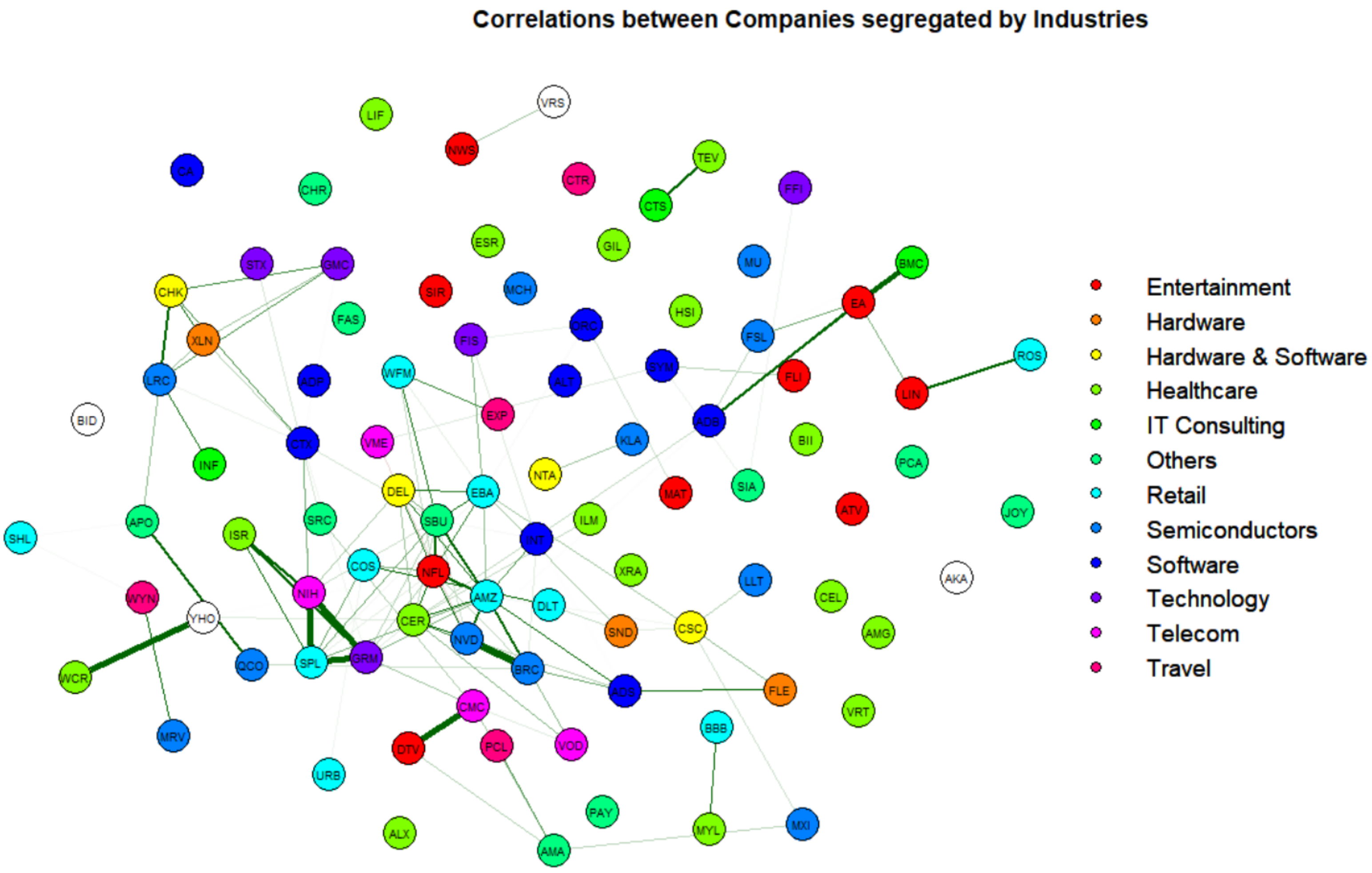

Fig.8 Correlation between companies (Spring layout)

Finally, I segregated the companies by their industries and visualized the correlation between them Fig.7 and found some interesting correlations like:

Comcast and DTV	EA & BMC	NIH & SPL
NIH & Garmin	Garmin & SPL	Amazon with More Others

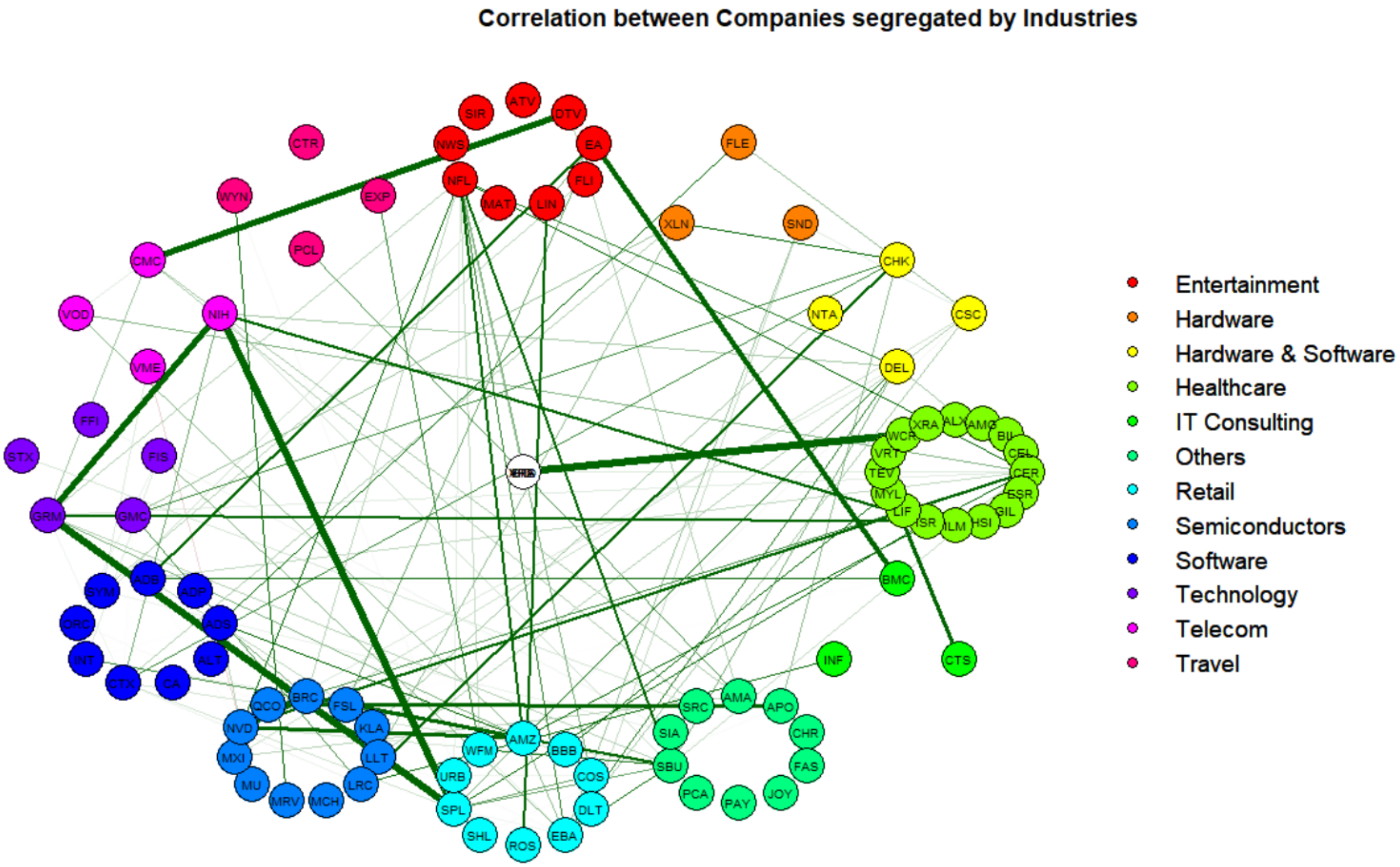

Fig.7 Correlation between companies (After Industry Segmentation)

I was also able to observe that there is a dense network correlation among **Retail, Software, Healthcare, Semiconductor and Entertainment** Industries.

References:

1. https://drsimonj.svbtle.com/how-to-create-correlation-network-plots-with-corrr-and-ggraph
2. **Statistical Analysis of Network Data with R.pdf Chapter – 7**
3. https://aws.amazon.com/solutions/case-studies/innovators/autodesk/
4. https://aws.amazon.com/solutions/case-studies/innovators/intuit/
5. https://www.nvidia.com/en-us/data-center/gpu-cloud-computing/amazon-web-services/
6. https://www.zdnet.com/article/first-amazon-destroyed-retail-now-theyre-coming-for-starbucks/
7. https://www.usatoday.com/story/money/business/2013/08/27/starbucks-whole-foods-evolution-fresh-cold-pressed-juice/2693983/
8. http://sachaepskamp.com/qgraph/examples