# SOCIAL ANALYTICS

## IDS 564 – ADVANCED LAB 2

Advanced Lab 2: Link Prediction: Inferring the Nasdaq 100 network of correlated social media chatter

Submission By-

Shivam Duseja

UIN: 678895805

## Introduction

In the era of social networking, daily chatters on social media platforms can prove useful in predicting and foreseeing the performance of various firms. By utilizing different features within a network, we can draw meaningful inferences that can further be used for link prediction i.e. to identify pair of nodes that will either form a link or not in the future. As a part of this advanced lab, we will be working on undirected networks that link NASDAQ 100 firms based on their partial correlations as per their Twitter activity.

As described in the paper by Tafti, Zotti and Jank (2015) – *'through monitoring of chatter on Twitter about firms listed on the Nasdaq 100, observing spikes of chatter affords a reliable and non-trivial amount of foresight into oncoming surges in trading volume.'* This can also be seen by an interesting example – wherein a tweet surfaced claiming that President Obama was injured by an explosion in the white house - caused a temporary drop of around 150 points in the Dow Jones industrial average.

As a part of this NASDAQ network, if there is a statistically significant correlation in the number of daily twitter messages between two firms – we can consider the two firms linked. This data shows the number of twitter messages, collected each day during financial trading hours – that mention the NASDAQ firms present in our dataset. In our work, we will be considering 92 firms listed in the NASDAQ 100, for 198 days of trading as described in the paper by - Tafti, Zotti and Jank (2015).

To start our analysis, we will first look into the correlation between different firms. Fig. 1 below displays the correlation between different NASDAQ 100 firms wherein the edge width displays

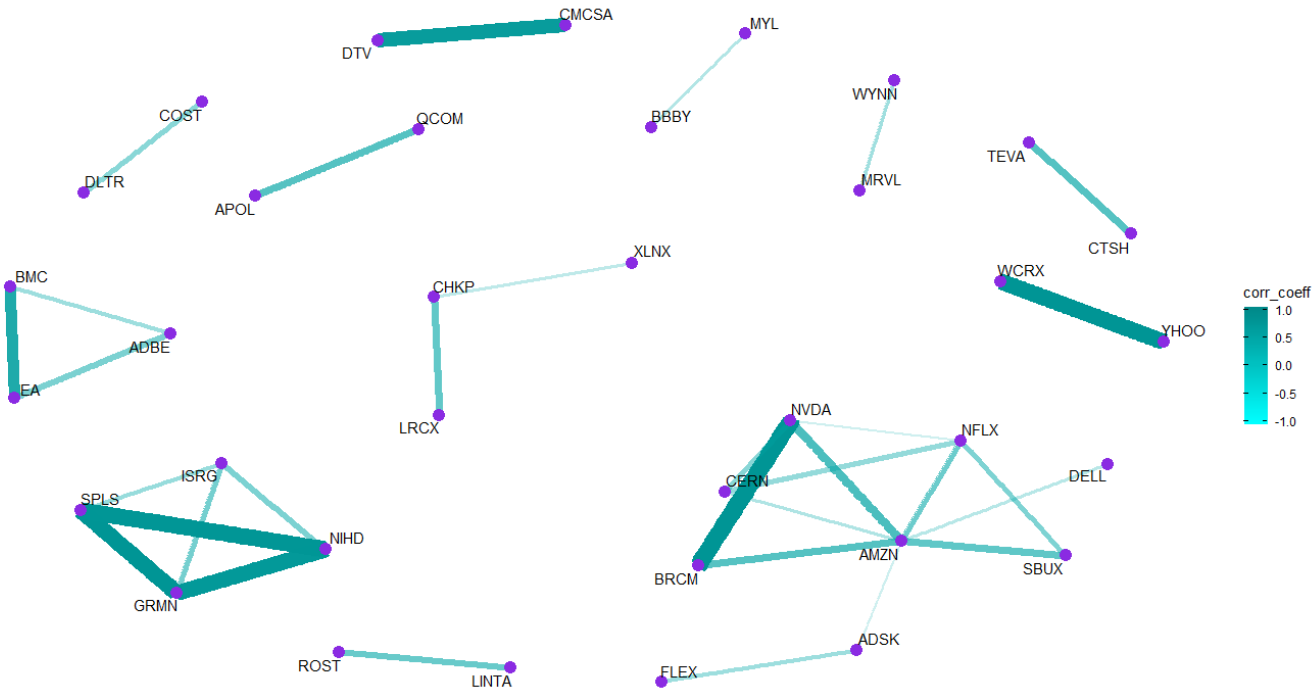


**Fig. 1: NASDAQ 100 firms: Correlation**

the strength of correlation and the edge color from light – dark green displays the range of correlation i.e. from negative – positive. In order to have a better look at the correlation between firms, the nodes have been filtered with an absolute correlation coef. value > 0.4.

**Network Analysis**

Figure 2 on the right, displays a heatmap of NASDAQ 100 twitter chatter data. These firms have been arranged according to their hierarchical clusters of the corresponding vectors of twitter chatter levels.

From the heatmap, we can observe that some of the firms display similar chatter activity across some duration of the trading days. However, in order to further make statistical conclusion from this data, we will use popular measures of association - correlation and partial correlations.

In order to come up with statistically significant results, we will first compute the empirical correlation using cor function. Post this we will compute the Fisher's transformation to approximate bivariate distributions along with the estimation of confidence intervals to obtain p-values.

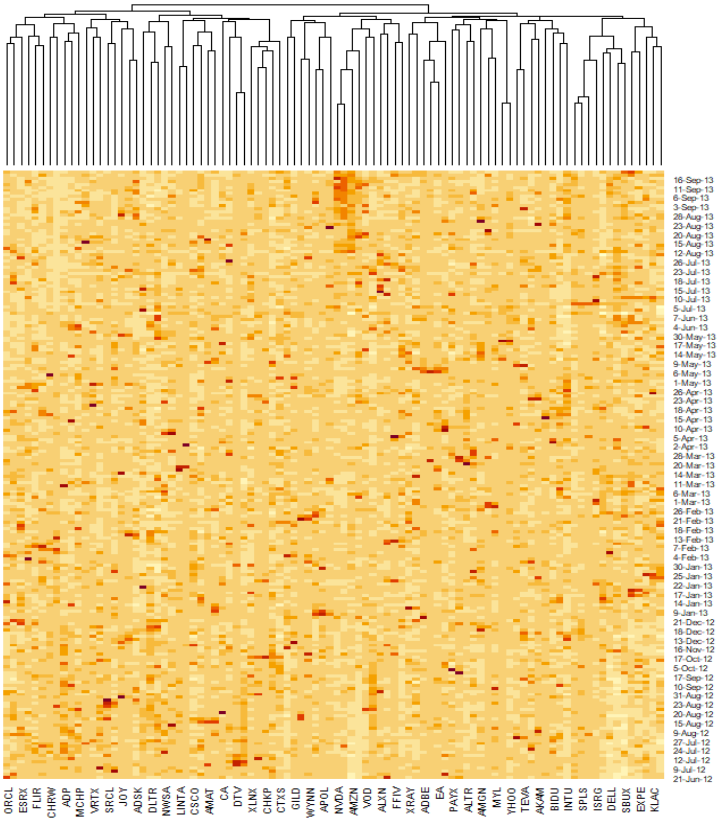


**Fig.2: Heatmap of NASDAQ 100 Firms**

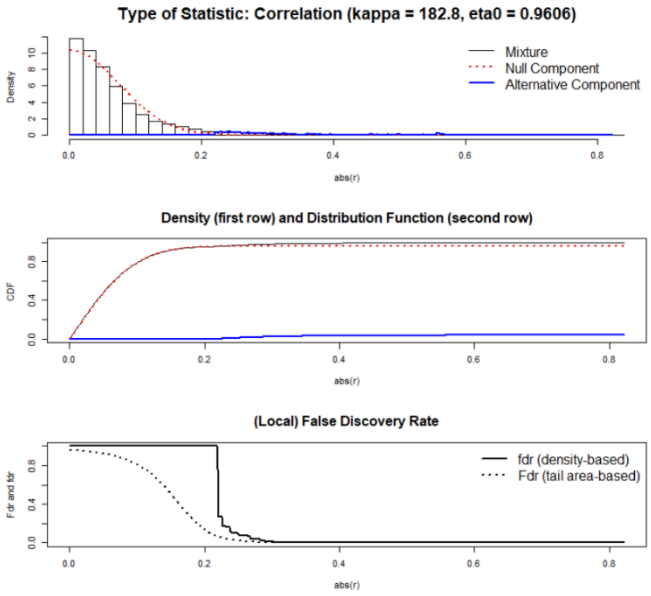


**Fig.3: Empirical Correlation Coefficients - Analysis**

Now, in order to adjust the False Discovery rate, Benjamini-Hochberg adjustment will be applied and a threshold of 0.05 is used to identify statistically significant partial correlations. Figure 3 on the left, shows the density, CDF and FDR for this twitter chatter dataset.

Post applying the above calculations, Fig 4 below, shows the network of firms post removing the nodes with degree less than 1. This network contains 48 edges and 56 vertices.

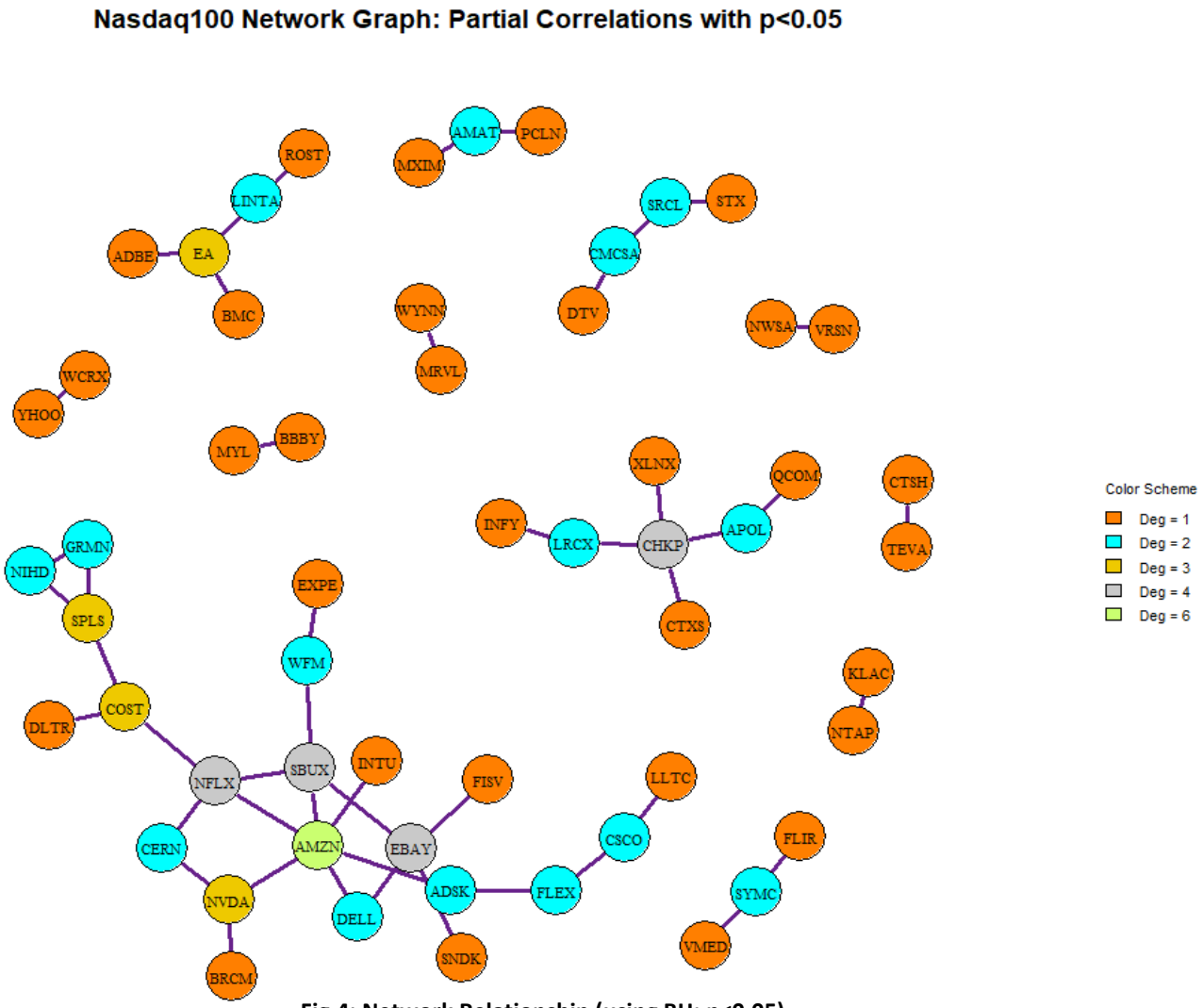


**Fig.4: Network Relationship (using BH: p<0.05)**

From the above graph, we can see that Amazon (AMZN) has the highest degree of 6. It is connected to NFLX, SBUX, INTU, ADSK, DELL, NVDA. These results show that a spike in twitter chatter about AMAZON may signal a surge in trading activity of the firm's stocks. A similar impact might be visible on the trading volume of the firm's linked to Amazon as well. The table below shows the Degree distribution of the firms (48 edges and 56 vertices).

| **Degree** | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| **# of Firms** | 32 | 15 | 4 | 17 | 1 |

Some of the key inferences that we can make from **Fig.4** above –

- Amazon's Prime video was launched in competition with Netflix, hence, a link between Amazon and Netflix makes sense in terms of competitive relationship

- Similar to the above case, relationship between Amazon and SBUX (Starbucks) looks to be competitive in a sense that Amazon is trying to replace the conventional coffee experience by selling products through amazon pantry.
- Amazon-Dell looks more like a cooperative relationship – as Amazon sells Dell Laptop's and other electronics products.
- Similar to the above point, relationship between Amazon and NVDA (Nvidia) looks to be cooperative in a sense that AWS and Nvidia are working together build powerful GPU-accelerated cloud solutions.
- Nvidia is also linked to Broadcom which looks more like a competitive relationship within the semiconductor industry.
- eBay is linked to Scandisk and Dell – which looks like a cooperative relationship as eBay sells Dell and Scandisk products.
- Costco seems to be connected to Staples and Dollar Tree which looks like a competitive relationship (with both of them) in a sense that all of them are brick and mortar stores and there is an overlap in the type of products being sold by Costco and both Staples and Dollar Tree.
- The viola plot for this network is shown on the right – which provides neighborhood information i.e. the pair of nodes with fewer neighbors and no edges at all.

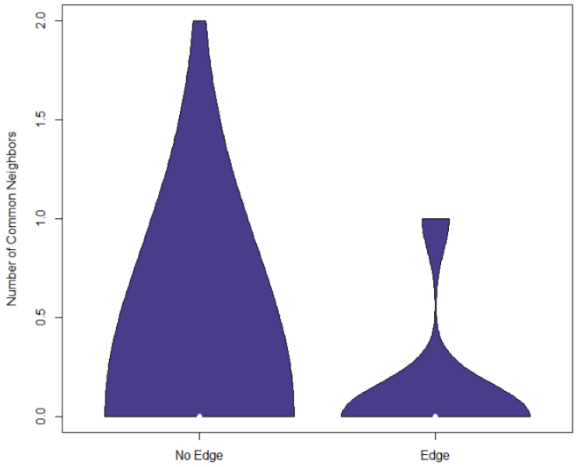


**Fig.5: Common Neighbors comparison (using BH: p<0.05)**

In order to better understand the network, with a higher statistical significance, I updated the threshold to p<0.01 and Fig6. below shows the network graph of the same. From the graph – we can see that Amazon now has a degree of 4 and its link with Nvidia and Intuit are not significant. The table below highlights the degree distribution of this network –

| **Degree** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **# of Firms** | 28 | 10 | 4 | 1 |

Some of the key inferences from Fig6. below are –

- eBay is still connected to Dell as it is a big seller of Dell products, however, eBay is not linked to scandisk at higher level of statistical significance.
- Connection between Staples – Garmin – NIHD is still intact, giving a sense of triadic closure.

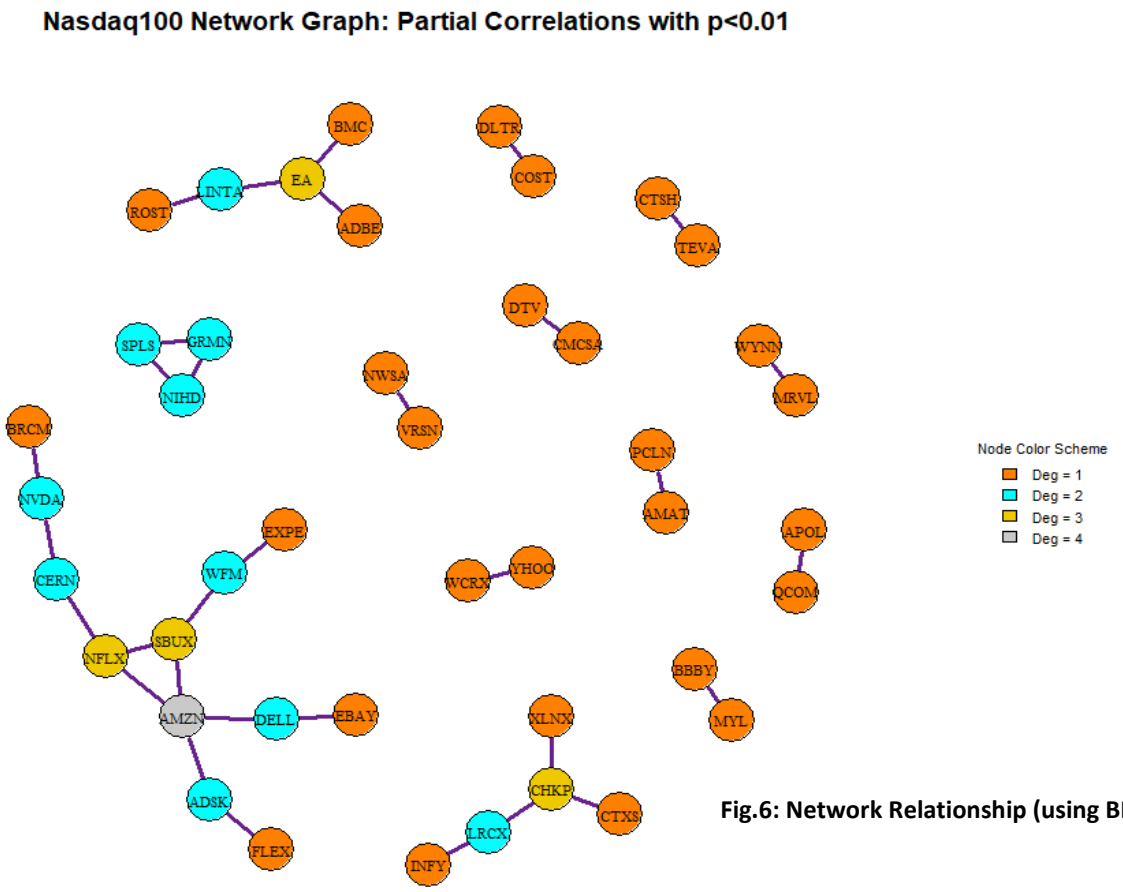


**Fig.6: Network Relationship (using BH: p<0.01)**

- As Starbucks (SBUX) partnered with Whole Foods (WFM) in order to sell beverage products – their link is intact even at higher level of statistical significance, showing deep cooperative relationship.
- The competitive link between Nvidia and Broadcom is intact – displaying a significant competitive relationship in the semiconductor manufacturing industry.
- Figure7. on the right shows the viola plot providing neighborhood information about the network.

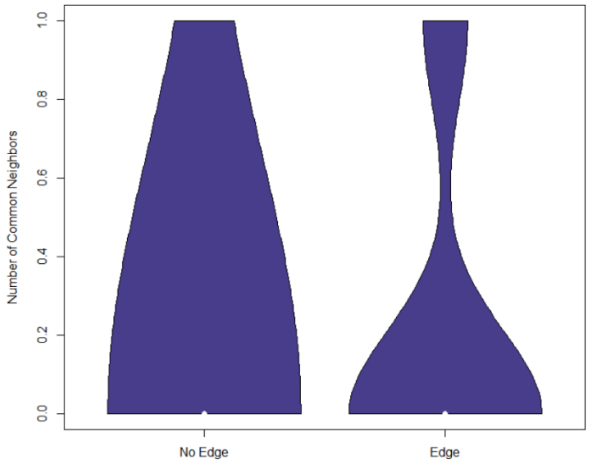


**Fig.7: Common Neighbors comparison (using BH: p<0.01)**

Now, in order to adjust the False Discovery Rate, I have used the FDR tool to generate links based on the statistical significance value of p<0.05. This analysis (47 edges & 56 firms) yields a result that is pretty similar to our initial analysis. The table below highlights the degree distribution of this network:

| **Degree** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **# of Firms** | 32 | 16 | 3 | 4 | 1 |

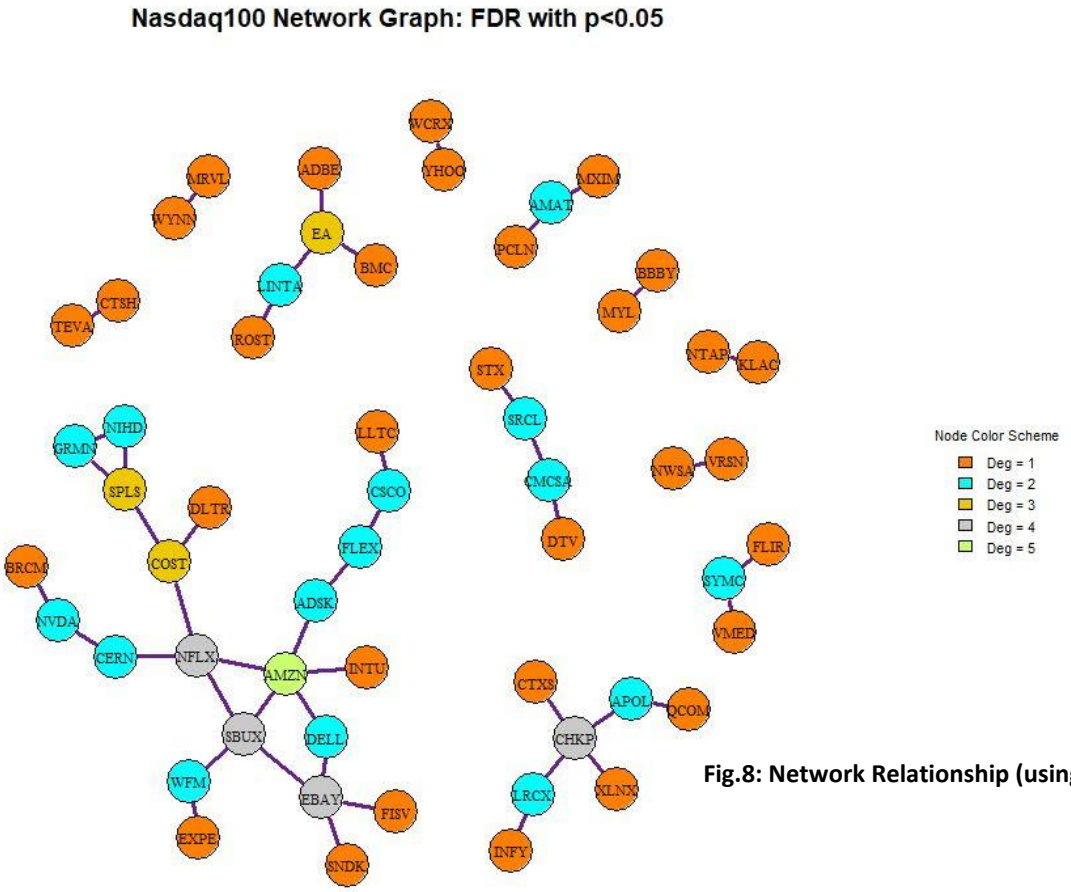


**Fig.8: Network Relationship (using FDR: p<0.05)**

From Fig8. above, we can see that again, Amazon has the highest degree of 5 with its neighbors as – DELL, INTU, SBUX, NFLX, ADSK, however, its link with NVDA is not statistically significant.

These results post using the FDR tool are consistent with the results obtained using partial correlations with a threshold of 0.05.

Fig9. on the right shows the viola plot providing neighborhood information about this network.

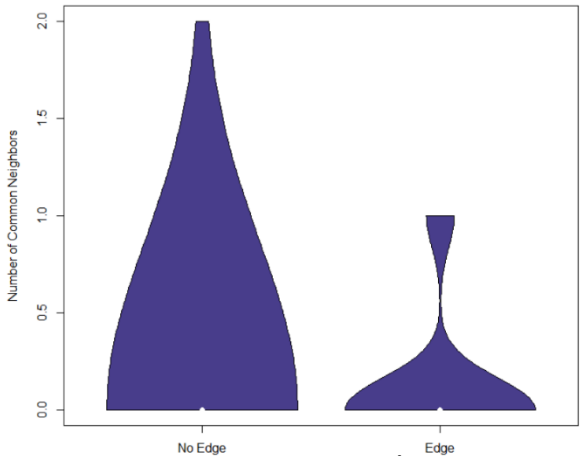


**Fig.9: Common Neighbors comparison (using FDR: p<0.05)**

In order to conclude our analysis, I will be using HUGE (high dimensional undirected graph estimation) to come up with the network graph. This library – transforms the data to have a marginal distribution which is close to normal and then tries to stabilize the overall estimation problem. Using this library – we have 109 edges and 77 vertices – statistically significant with a threshold of p<0.05. The degree distribution table of this network is shown below:

| Degree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 10 |
|---|---|---|---|---|---|---|---|---|
| # of Firms | 22 | 17 | 15 | 10 | 8 | 1 | 3 | 1 |

From this network, we can observe the most number of statistically significant nodes along with their links.

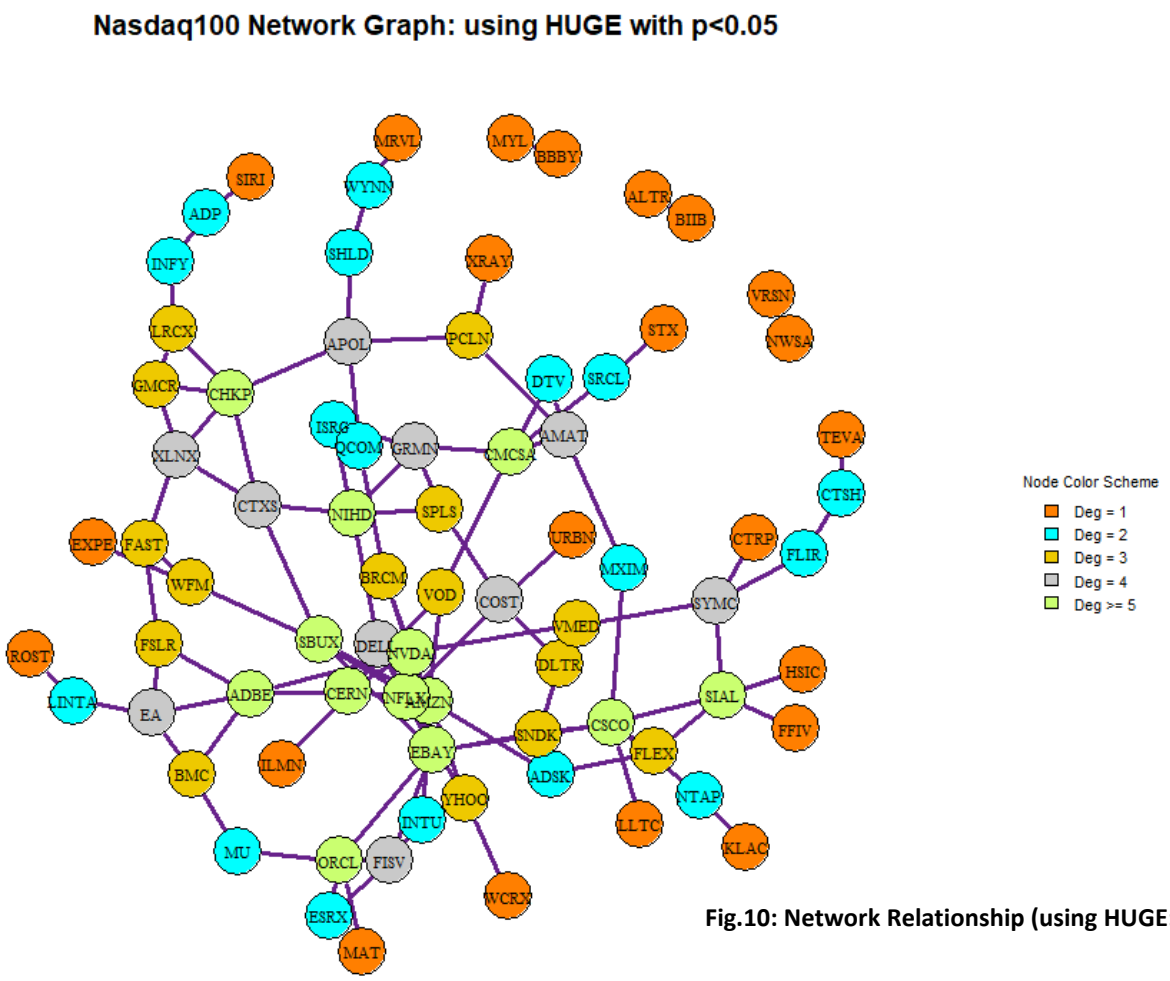


**Fig.10: Network Relationship (using HUGE: p<0.05)**

From the above Fig10. it can be inferred that different firms are linked based on their market/domain of functioning. Amazon being the biggest player in the market across domains like – retail, commerce, cloud, digital products, Amazon go stores - is linked to highest number of firms.

Fig11. on the right shows the viola plot providing neighborhood information about this network

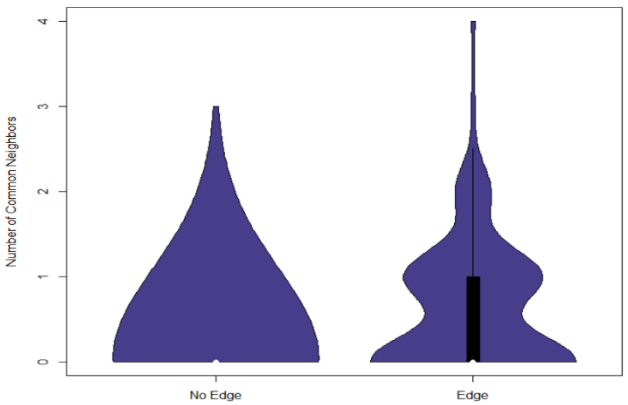


**Fig.11: Common Neighbors comparison (using HUGE: p<0.05)**

From the above analysis, we can conclude that the firms are linked if there is a statistically significant correlation in the daily number of twitter messages in which they are mentioned. This helps us in extracting valuable information for participants in the financial market as well helps us in finding pattern of information diffusion.

These inferences might prove to be useful for financial traders as these chatters in Twitter can be used effectively to predict the stock price/movement. Link predictions also help us in understanding the cooperating/competitive relationship between different firms giving us a sense that if there is chatter about a firm on twitter it might impact the trading volumes of other firms linked to it.

**References**

Tafti, Ali, Ryan Zotti, and Wolfgang Jank, "Real-Time Diffusion of Information on Twitter and the Financial Markets," PLoS ONE 11(8) 2016: e0159226.

https://drsimonj.svbtle.com/how-to-create-correlation-network-plots-with-corrr-and-ggraph

Kolaczyk E.D., Csárdi G. (2014). Statistical Analysis of Network Data with R. vol 65. Springer, New York, NY

Qian Li, Bing Zhou and Qingzhong Liu, "Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion," *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, 2016, pp. 359-364.