# Social Media and Network Analysis Pawanjeet Kaur

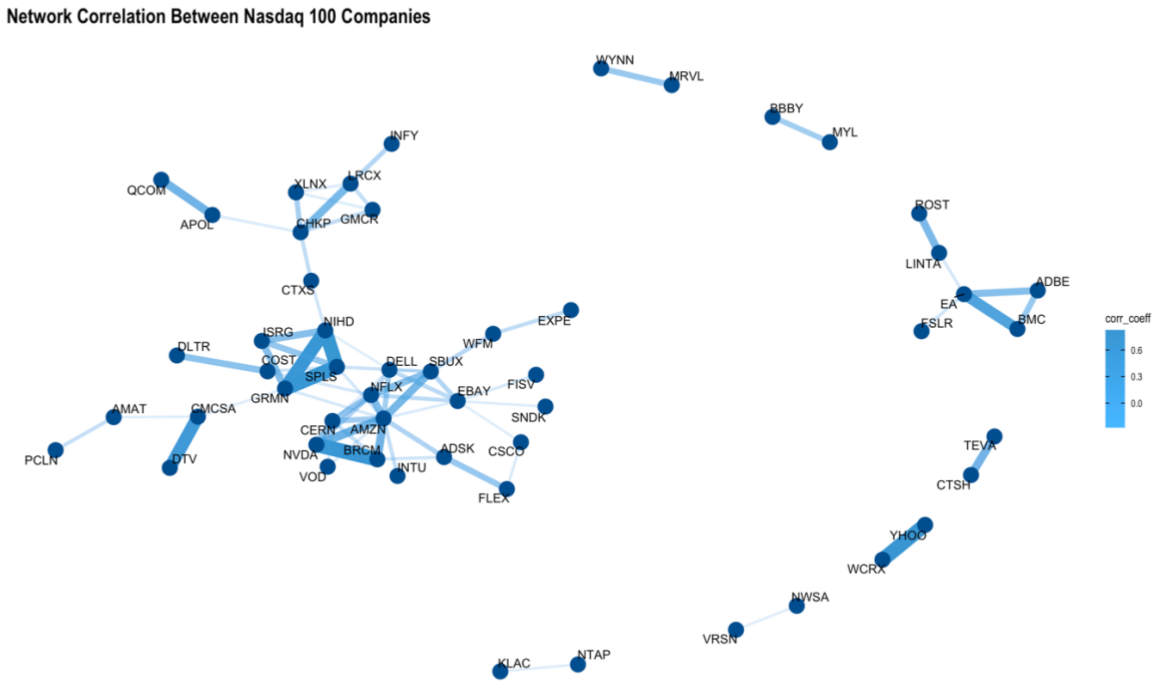
*Advanced Lab 5: Inferring the NASDAQ 100 Network of correlated Social Media Chatter*

This report contains results of analyzing NASDAQ 100 firms' twitter data. It is based on analyzing undirected graph, and statistical significance between two firms which will tell about the likelihood of any connection between the firms. Any two firms are considered to be linked if two are correlated in statistically significant way.

The dataset contains number of twitter messages that mention each of Nasdaq firm. As described in the paper "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" by Tafti, Zotti and Jank (2015) the data is from real-time relationship between chatter on Twitter, and the trading volume of Nasdaq 100 firms during 193 days of trading in the period from May 21, 2012 to September 18, 2013 except weekends and holidays.

As a part of the analysis, initially, network plot with at-least one edge is generated, which tells about the companies which are related to each other by at-least one degree.

In the below figure, every node is representing one firm, and the edge between two nodes define the correlation between the two firms. With the change in magnitude of co-relation the width of the edge varies i.e. width of edge represents the strength of correlation.

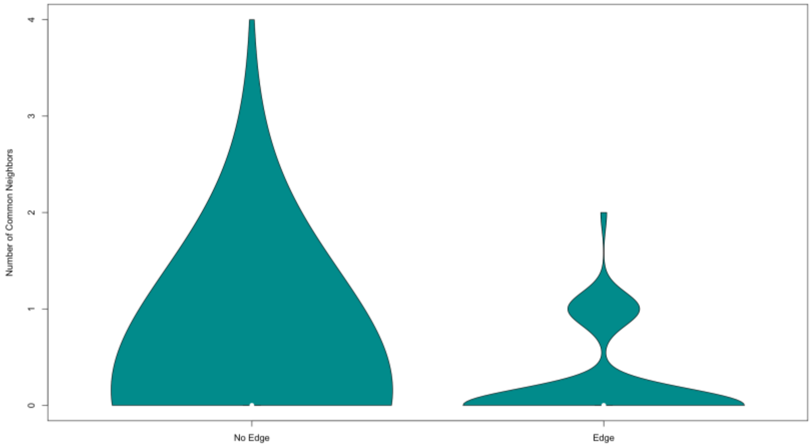


Above network is a filtered network plot with correlation coefficient above 0.30 which has 53 firms as nodes and 142 edges representing the relations.

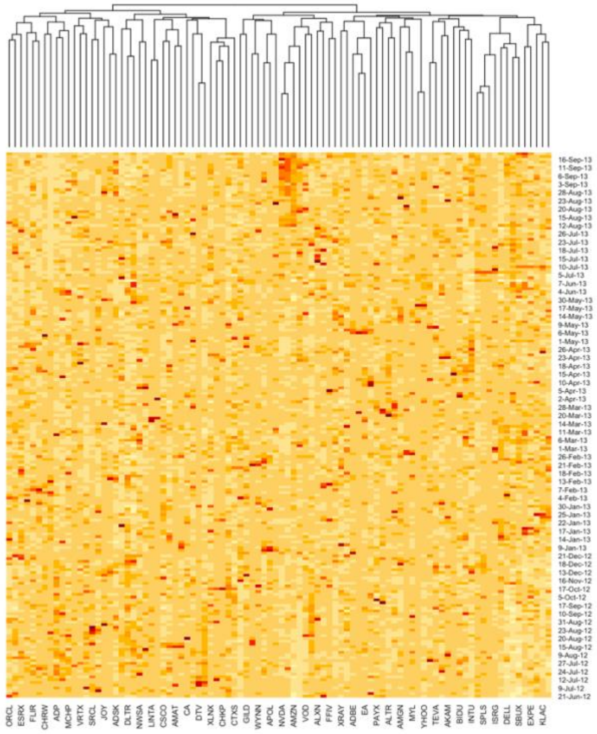
The Degree Distribution Table and Viola plot with 0.05 threshold for overall correlation plot is as below:

Overall Correlation: p<0.05

| **Degree** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 11 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|
| **# of Firms** | 12 | 15 | 13 | 17 | 11 | 5 | 12 | 2 | 1 | 1 |



The next figure represents the heatmap visualization of Nasdaq 100 twitter chatter data. The firms are shown as columns and every row represents new day. The columnar data is clustered hierarchically which represents the activity level of the nodes (i.e. firms in this case).



The graph shows, the firms grouped together have activity level variation. Also, there is a relationship between the firm, when two firms are mentioned in the same tweet, those are grouped together and are considered linked in some way.

Since it is tough to read through the heatmap plot, further analysis is performed based on its statistical significance using correlation and partial correlation.

At first, the partial correlation coefficient between each node is calculated, and Fisher’s approximation is performed. It is done in the code from line 54 from chunk 9.

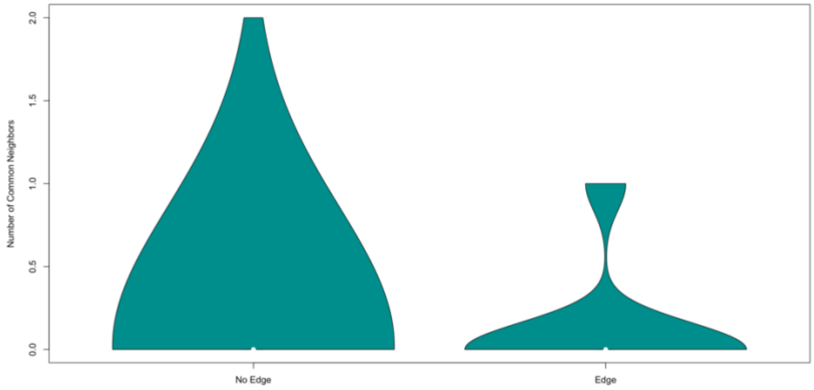
This transformation approximates bivariate distribution and calculates p-value by estimating confidence interval. Over that - to control the

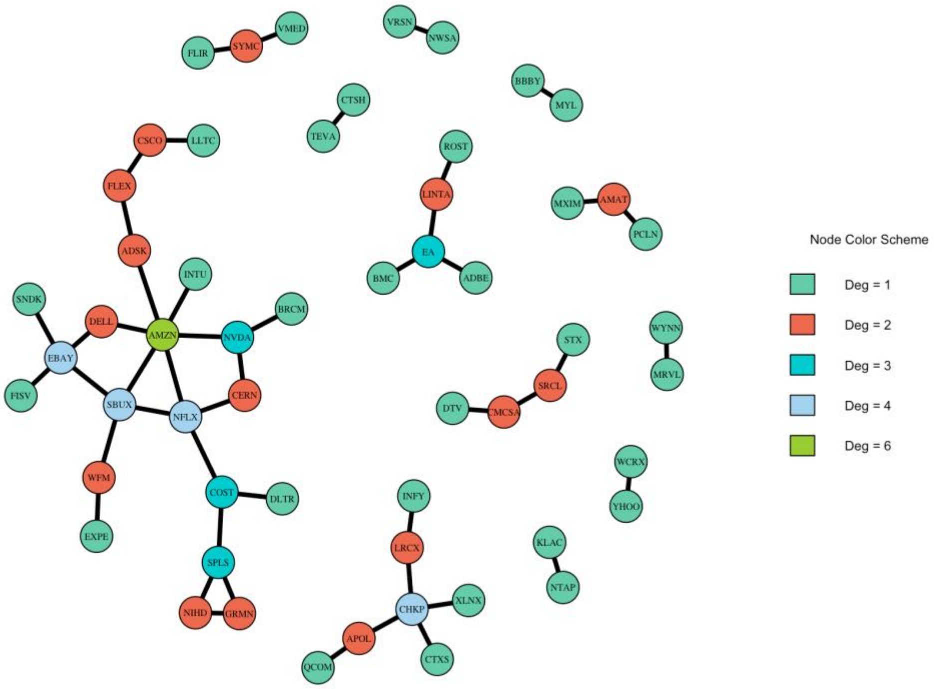
False discovery rate, “Benjamini -Hochberg” adjustment with threshold of 0.05 is applied, identifying significant partial correlations. It gave the number of edges and nodes. Then, using code from Regular Lab 5 network graph of 48 edges and 56 vertices is plotted. The degree distribution for same is shown below.

Using BH - Partial Correlation: p<0.05

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

The next figure shows the Viola plot of partial correlation with p<0.05, and degree(node) >= 1. The plot shows the pairs of nodes having fewer neighbors and no edges at all.





From the degree distribution table above, it is noticed that among these AMZN is having the highest degree, so in the next step, some of the neighbors of AMZN are looked up i.e. to which other 6 firms it is associated with. It is found that it connects to ADSK, DELL, INTU, NFLX, NVDA and SBUX.

AMZN relations could be justified and categorized as below:

ADSK(Autodesk)/AMZN(Amazon) and INTU(Intuit)/AMZN(Amazon) can be categorized as Cooperative Relationship. The relationship of Amazon and Dell could be due to electronics products since Amazon serves as platform for online sale for Dell products.

Some of the relations of AMZN can be described as industry level. Amazon is related to Netflix probably because both fall under entertainment sector.

NVDA (Nvidia) and AMZN(Amazon) probably also could be cooperative because both provide cloud services.

AMZN(Amazon) and Starbucks (SBUX) could be related because of both having physical stores; Amazon Go stores are gaining huge popularity these days just like Starbucks.

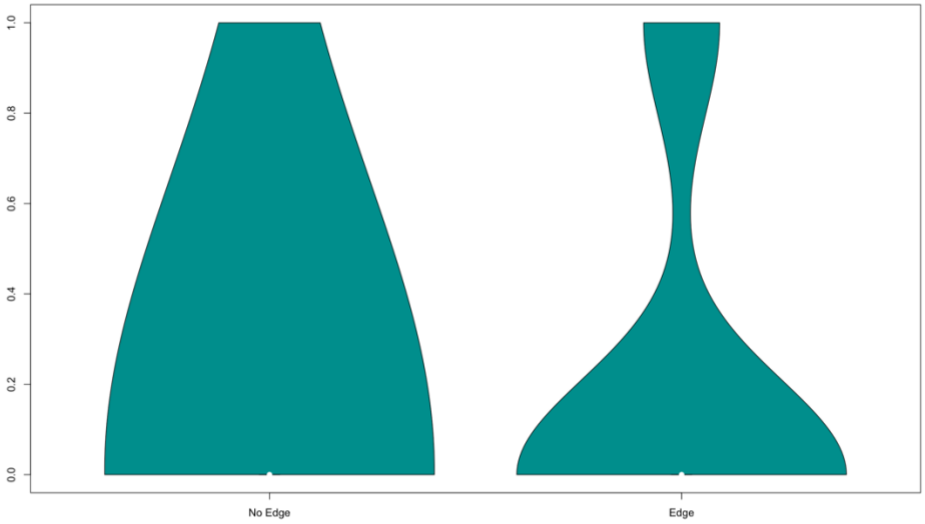
Further, it can be seen that this network still has huge number of firms. To understand the relationship better, we can play around with the threshold, and analyze the impact on nodes and edges. In next step, the threshold is set to 0.01 and the below network graph along with viola plots is obtained.

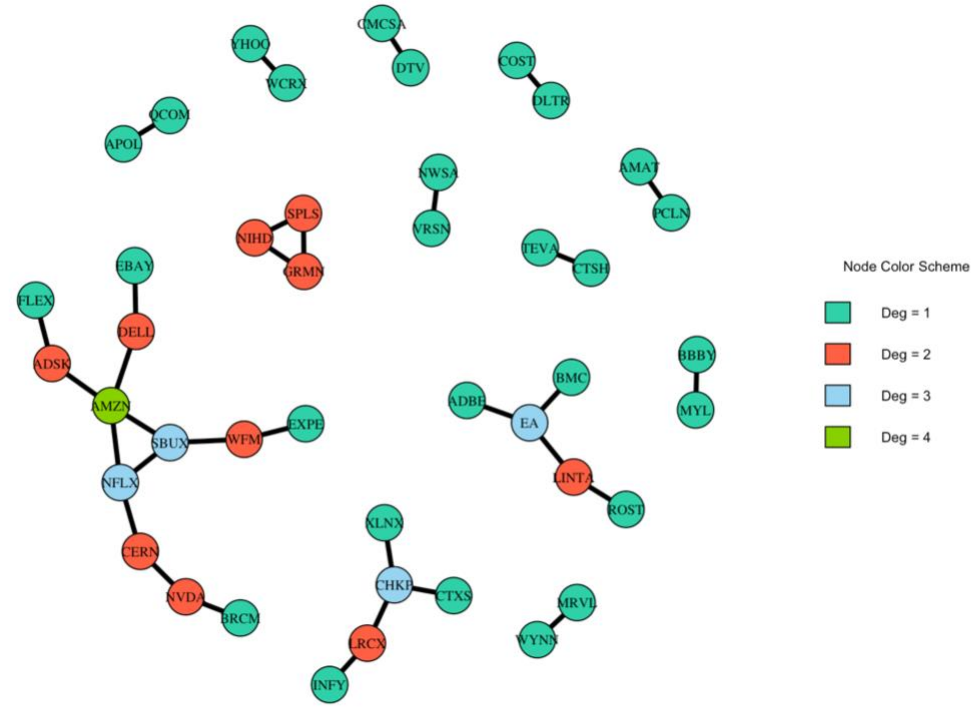
These networks under 0.01 threshold have 32 Edges and 43 Vertices which are statistically significant. Degree distribution table for same is as below.

Using BH - Partial Correlation: $p<0.01$

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

Below image represents the network plot using similar layout settings with 0.01 threshold and it can be seen that AMZN has now links to 4 other firms. This gives the subset of edges that were created with regular correlation coefficient.





Again, under new condition, AMZN has highest degree of 4 and neighboring firms to AMZN are NFLX, SBUX, DELL and ADSK. The links to INTU (Intuit) and NVDA(Nvidia) no longer significant. But there are NIHD, SPLS and GRMN in the plot shows triadic closure between them. In earlier analysis, this triadic closure was linked with COST, and COST was linked with DLTR (i.e. Costco and Dollar Tree) but that link is no longer significant under new threshold. COST and DLTR fall under similar category of market (retail) and hence justifiable link.

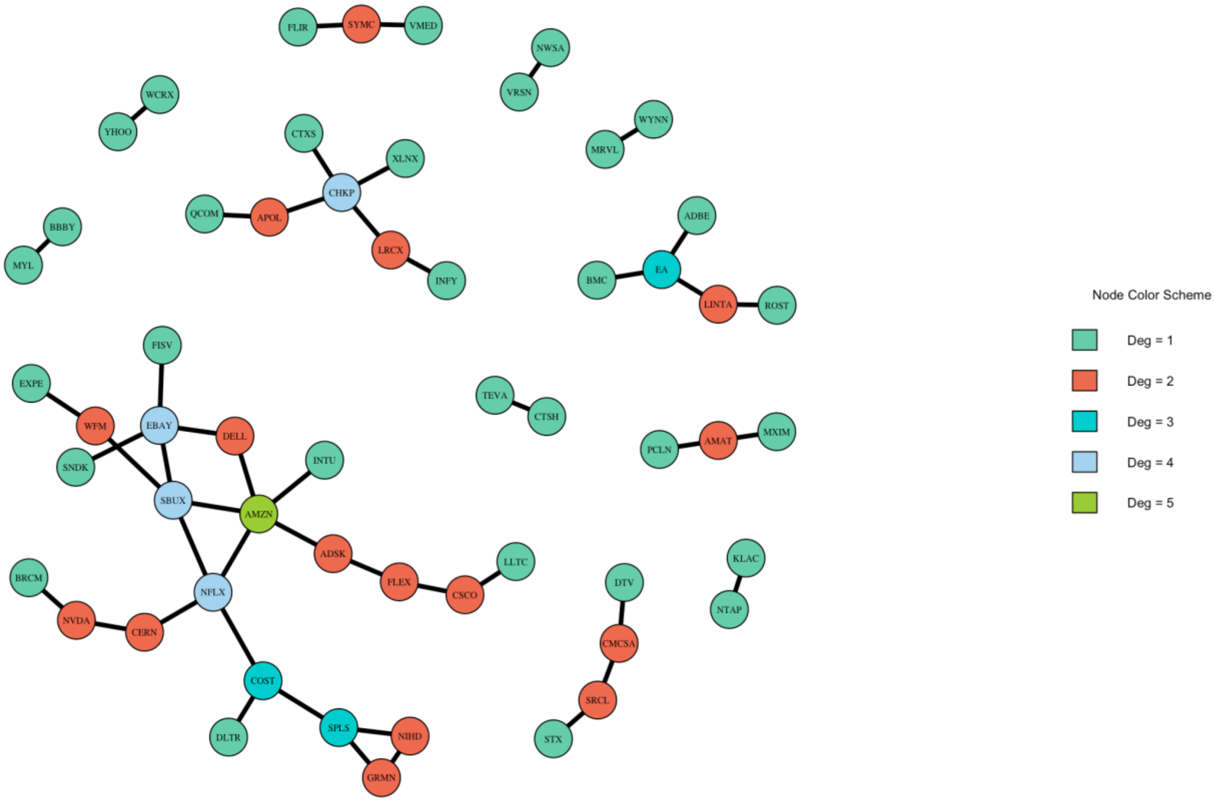
Links of EA, LINTA, ADBE, BMC and ROST are still significant under 0.01 and 0.05 threshold. None of the link is broken in these firms.

After this, a different approach is used to generate the network. In this, FDR are used tools to generate links, and statistically significant links are found under the threshold of 0.05. Below is the degree distribution table, network and viola plot for the same.

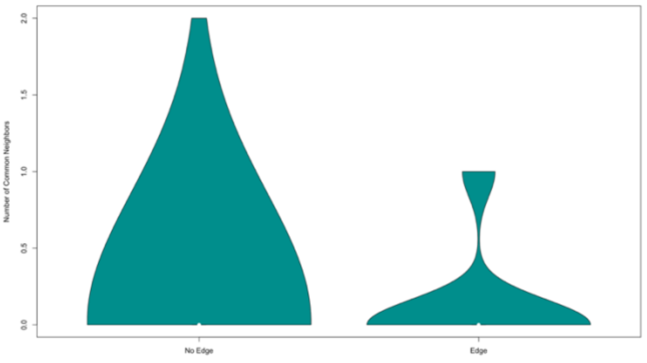
Using FDR Tool: p<0.05

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

This network has 56 vertices and 47 edges which are significant. This is approximately similar to the one generated in the initial analysis.



AMZN has highest degree of 5 and its neighbors are ADSK, DELL, INTU, NFLX, SBUX. In this graph link between AMZN and NVDA is broken.



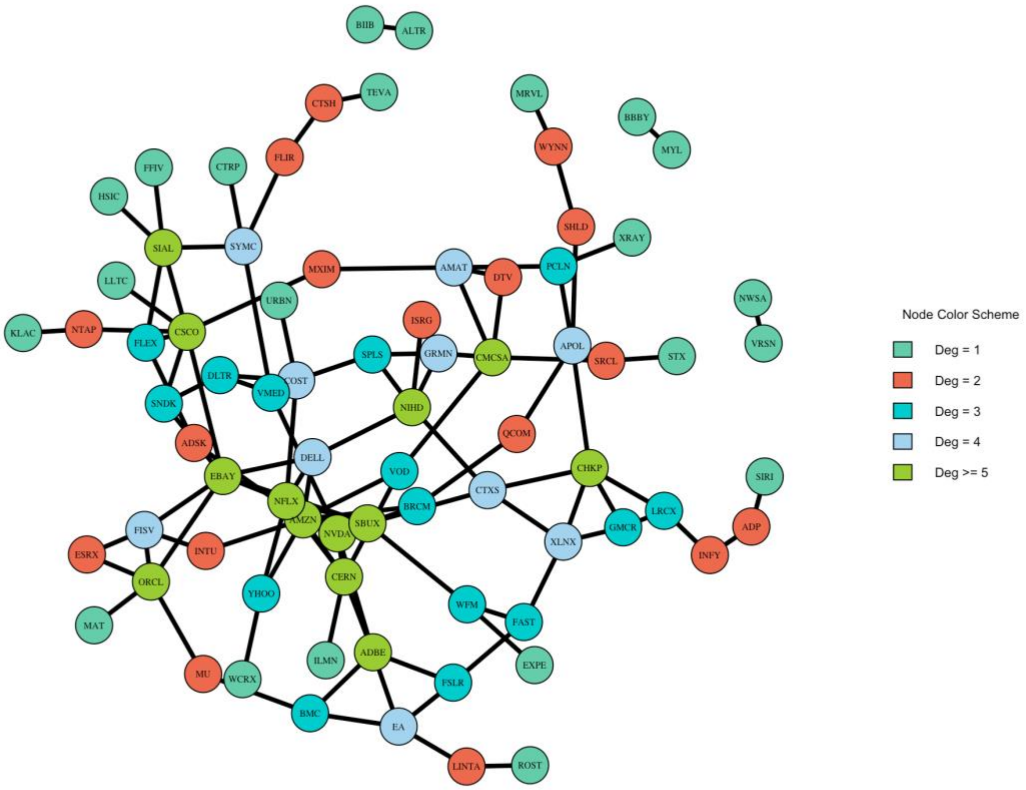
The results using FDR-tools library are similar to the ones generated using partial correlation with threshold of 0.05.

Then as the last analysis, Huge library is being used, the network is generated using this library with the threshold of 0.05. It produced network of Edges 109 and Vertices 77 which are statistically significant. The degree distribution, network plot and viola plot are as below.

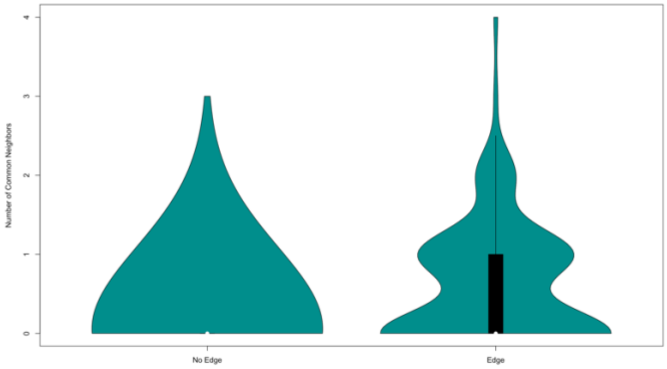
Using HUGE Tool: $p<0.05$

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

This graph gives the highest number of significant nodes and links between them. Again, AMZN has the highest degree and it is connected to 10 other firms. As Amazon is expanded in different markets of retail, ecommerce, digital, physical stores, cloud, it gives us the reason to see the highest degree in all the analysis.



AMZN is linked to ADSK, BRCM, CERN, DELL, INTU, NFLX, NVDA, SBUX, VOD, YHOO. This is highly connected graph among all the graph produced in this report.



Hence, it can be seen that the links between the firms are highly likely due to their industry. If two firms have similar industry or, are a competitor to each other - they are more likely to have an edge between them more significantly. e.g. if a new firm is introduced in retail business, it is more likely to see its links with Costco, Amazon with more significance.

To further increase the scope of this analysis, industry level analysis can be run, and firms can be mapped to corresponding industries and different significant plots can be generated which are beyond the scope of this advanced lab.

To conclude our analysis, we can say that link prediction can be useful in the financial markets in a sense that a spike in the twitter chatter can have an impact on the trading volumes of that firm along with the firms it has statistically significant connections to.

**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

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

Statistical Analysis of Network Data with R.pdf

https://aws.amazon.com/solutions/case-studies/