# Social Media and Network Analysis Pawanjeet Kaur

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

This advanced lab focuses on analyzing NAQDAQ 100 firms’ twitter data. Analyzing undirected graph and statistical significance between two firms will tell us about the likelihood of any connection between the firms. Any two firms will be linked if two are correlated in statistically significant way.

Our dataset contains number of twitter messages that mention each of Nasdaq firm. As described in paper by Tafti, Zotti and Jank (2015) we have data 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 analysis, initially, we will plot the network with at-least one edge which will tell us about the companies which are related to each other of one degree.

In below figure, every node is representing one Firm and edge between two nodes define the correlation between 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.

A close up of a map

Description automatically generated

Above network is filtered network plot with correlation coefficient above 0.30 which has 53 firms as nodes and 142 edge representing relation.

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 |

A close up of a logo

Description automatically generated

Next figure represents the heatmap visualization of Nasdaq 100 twitter chatter data. The firms are shown as columns and every row shows new Day. The columnar data is clustered hierarchically.

A picture containing food, fence

Description automatically generated

It represents the activity level of the graph.

The firms grouped together has activity level

variation. Means, there is relation between the

firms. When two firms are mentioned in the

same tweet those are grouped together are

linked in some way.

As it is tough to read through heatmap plot

we will perform various analysis further

on its statistical significance using correlation

and partial correlation.

At first, I calculated the partial correlation

coefficient between each node and performed

Fisher’s approximation as done in code from

line 54 from chunk 9.

This transformation approximates bivariate

distribution and calculate p-value by estimating

confidence interval. Over that to control the False discovery rate I applied “Benjamini -Hochberg” adjustment with threshold of 0.05 identifying significant partial correlations. It gave number of edges and nodes and using code from Regular Lab 5 plotted the network graph for the same. Under such conditions, we get network of 48 edges and 56 vertices with degree distribution shown below.

Using BH - Partial Correlation: p<0.05

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

Next figure shows the plot of partial correlation with 0.05 threshold of nodes with degree of 1 and above with viola plot showing the pairs of nodes having fewer neighbors and no edges at all.

A picture containing drawing

Description automatically generatedA picture containing food

Description automatically generated

From the degree distribution table under such constraints we noticed that among these AMZN is having the highest degree, so we tried to look for some of the neighbors of AMZN i.e. to which

other 6 firms it is associated with. It connects to ADSK, DELL, INTU, NFLX, NVDA and SBUX.

We can see Amazon can be related to Netflix as both fall under entertainment sector. Further, Netflix, Ebay is further related to 4 other firms. Some of the relation can be described as industry level like Netflix and Amazon fall under entertainment sector.

ADSK(Autodesk)/AMZN(Amazon) and INTU(Intuit)/AMZN(Amazon) can be cooperative relationship. NVDA (Nvidia and AMZN(Amazon) also cooperative for some cloud services. AMZN(Amazon) and Starbucks (SBUX) can be because of physical stores as Amazon Go stores are gaining huge popularity these days. Amazon and Dell are linked due to electronics link. Dell laptops are available on AMZN platform.

We can see this network has still huge number of firms with balanced edge network. To understand the relationship better, we can play around with the threshold and analyze the impact on nodes and edges. In next step, I have set the threshold to 0.01 and obtained the below network graph along with viola plots.

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 we can see that AMZN has now links to 4 other firms. This gives us the subset of edges we created with regular correlation coefficient.

A close up of a logo

Description automatically generated

A picture containing food

Description automatically generated

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 we have NIHD, SPLS and GRMN in the middle of the plot have triadic closure. In earlier analysis, this triadic closure was linked with COST and COST linked with DLTR

(i.e. Costco and Doller 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, I used different approach to generate the network and used FDR tools to generate links and found statistically significant links under 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 our initial analysis.

A picture containing food

Description automatically generated

A close up of a logo

Description automatically generated

AMZN have 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 we generated using

Partial correlation with threshold of 0.05.

Then as the last analyses, we will run it using Huge library, I generated the network using library huge with 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.

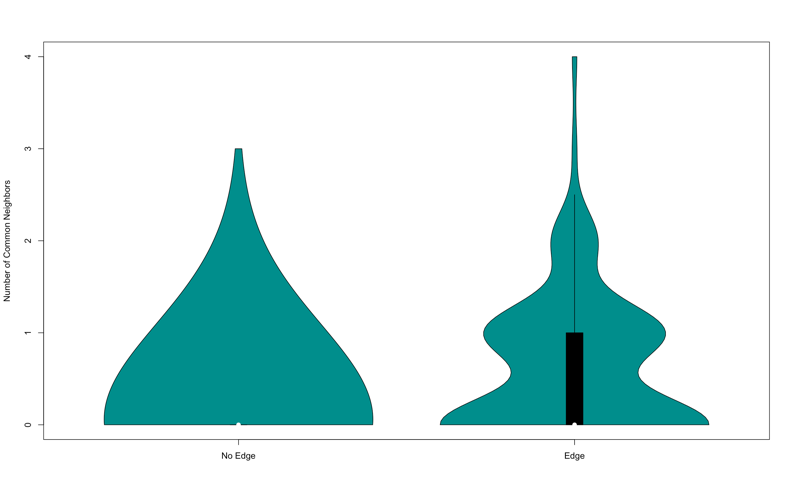
Using HUGE Tool: p<0.05

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 and physical Amazon Go gives us the reason to see the highest degree in all the analysis.A close up of a logo

Description automatically generated

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.



In conclusion, we can see that the links between the firms are highly likely due to its industry impact. If two firms have similar industry or are competitor to each other they are more likely to have edge with them more significantly. Like if we introduce a new firm in retail business, we are more likely to see its links with Costco, Amazon with more significance.

To further increase the scope of this analysis, we can run industry level analysis and map firm to corresponding industries and plot different significant plot which are beyond the scope of this advanced lab.

In general, this lab is useful in analyzing the current twitter trends in their mentions, see the significant link prediction and learn the potential financial impact due to the same.