Devesh Sharma (UIN: 664049903)

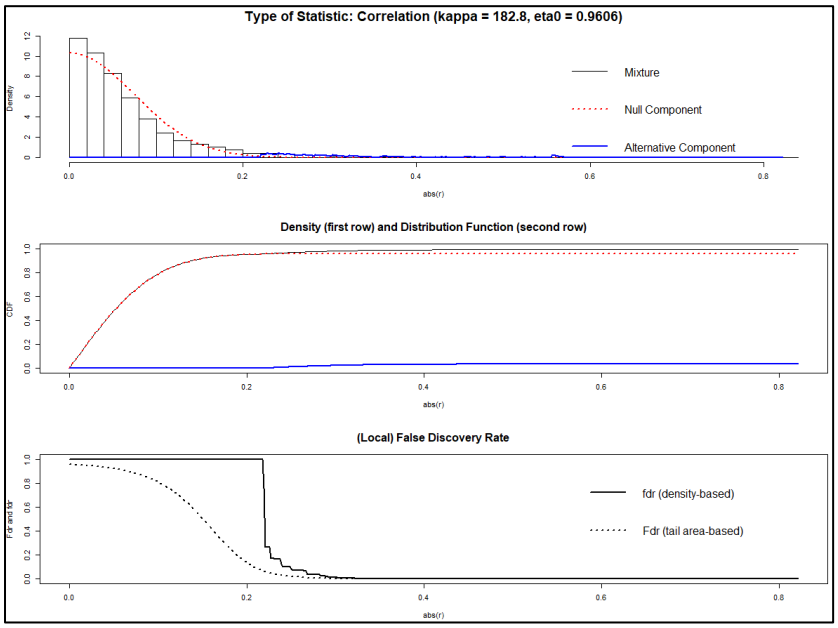
Advanced Lab 5

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

Data Introduction:

The dataset given to us shows the number of Twitter messages, collected each day starting from 21st June 2012 till 18th September 2013 during financial trading hours. This data mentions 92 firms out of Nasdaq 100 and is not a comprehensive count of mentions on Twitter because few specific keywords were used as filter.

Plot after calculating Pearson Coefficient:



Methodology:

I will be identifying statistically significant links between firms and would be using partial correlation coefficient methodology for this. Once, I specify the statistically significant level which is required, then I would be plotting only those links which satisfy the selected significance criteria ($p<0.05$ or $p<0.01$). Then, I will be performing 2 kinds of comparisons, one between the original and either of the statistically significant graph and the another one between 2 levels of significance. The procedure to be followed to determine these statistically significant links is as below:

1. Calculate the partial correlation coefficients between each node; this is the correlation between any pair of two nodes that remains after adjusting for their common correlations with every other node in the graph
2. Compute the Fisher's transformation to approximate the bivariate distributions and to determine the confidence intervals that are used to obtain p-values
3. Apply the Benjamini-Hochberg adjustment to control for the false discovery rate; and use a threshold of $p < 0.05$ to identify statistically significant partial correlations
4. Use the calculations in the above steps to determine the edges among the nodes based, and finally, to construct the network of firms

Questionnaire:

What does the resulting network reveal about the relationship between the firms as they are described in Twitter?

The left graph showcases the original network without any edges removed whereas the right graph showcases the network with statistically significant edges only (with $p<0.05$).

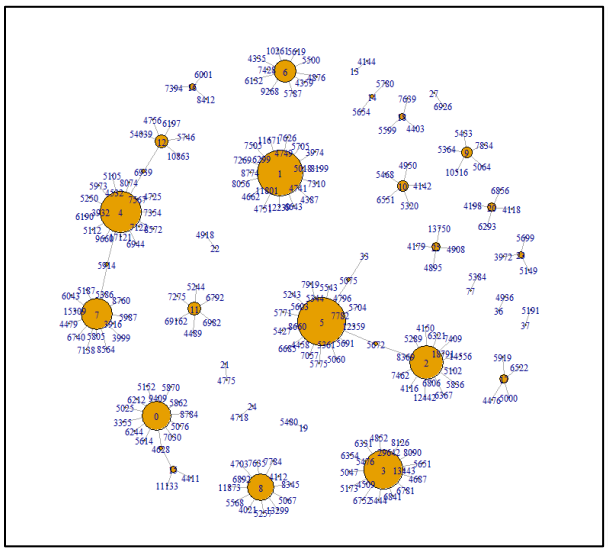


Fig 1: Original Network

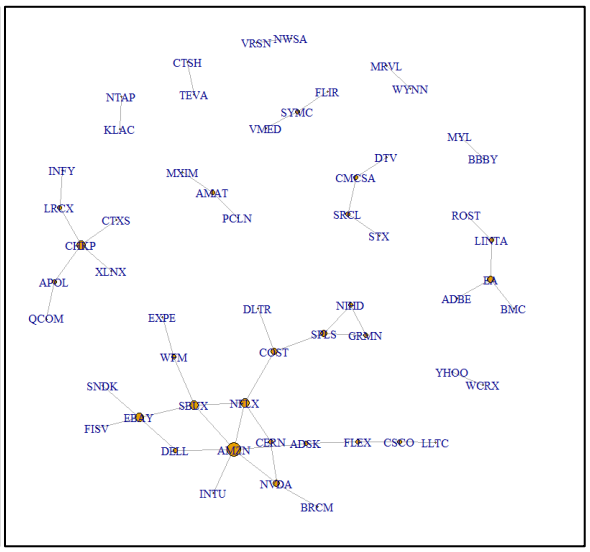


Fig 2: Network with statistically significant edges having p<0.05

The network on the right was created using Benjamini-Hochberg adjustment and we can observe that there were lots of relationships between the firms (as described in Twitter) which were not significant enough and thus got removed, when compared to the original network. Thus, Benjamini-Hochberg adjustment allows to check the false discovery rate and enables us to deduce findings from only the important links/ connections which are statistically significant. This adjustment reduced the number of edges to mere 48 significant edges, which could be used for inferring any result.

Do the edges appear to represent competitive relationships, cooperative relationships, or some other type of connection?

Yes, the edges in this network are representing some sort of relationships among the nodes (NASDAQ firms). The type of these relationships varies from being a competitive to cooperative one, making the network has a mix of all type of connections between the firms. To elaborate on few of the very predominant relationships, I would like to pick up the biggest connected firm in our network i.e. Amazon (AMZN), as observed in the network with statistically significant edges with $p<0.05$. It is connected to some other big firms like Dell, Intuit, Netflix, etc. where it seems to have a competitive relationship with Dell whereas a cooperative one with Intuit.

1. Amazon's Relationship with Dell: I believe that Amazon would have a competitive relationship with Dell. Now someone might believe that they should be partners as Dell is selling its laptop on Amazon and thus, they both are getting business from each other but then we cannot neglect Dell's adaptation of cloud based services. Back in 2011, Dell entered public cloud space with VMware, investing $1B in cloud computing. Then, its partnership program with EMC in 2017, further established its hold in cloud computing arena. These moves make Dell a potential threat to Amazon since they are market leaders in providing cloud computing services.

2. Amazon's Relationship with Intuit: Intuit is a financial software company which produces relevant software for its customers. It relies heavily on Amazon for selling these products and for AWS services also. To quote from the link (provided below in the Citations), on November 2017, Amazon announced that Intuit selected AWS as its standard for machine learning and artificial intelligence workloads. Intuit says that "We started our journey to AWS in 2013, and AWS has been essential to our transformation in the cloud. We extended our relationship with AWS to enhance our flagship products and services, including QuickBooks, Mint, and TurboTax, and accelerate our efforts to apply artificial intelligence within our business". This clearly showcases that Intuit is having a dependent relationship with Amazon.

How would you describe the meaning of the edges between nodes in the resulting network?

As discussed in the above question, edges signify a certain type of relationships between the nodes (firms) like competitive or dependent. We studied the network with statistically significant edges having $p<0.05$ above. But if we look at another statistically significant edges' network with $p<0.01$, we can see that number of significant edges further reduced to 32.

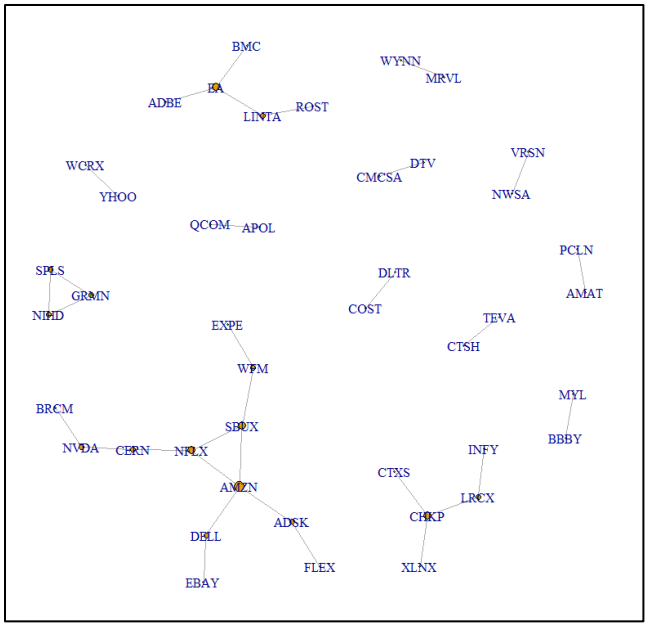


Fig 3: Network with statistically significant edges having p<0.01

This showcases that the edges also stand for the strength/ significance of the connection/ relationship between 2 firms. What I mean here is that, if we further stricken the significant criteria to p<0.01, few more weaker connections (as compared to the case with p<0.05) between the firms are getting disappeared. Thus, we can conclude that each edge stands for 2 things:

- Nature of the relationship between 2 firms
- Strength/ Significance of the relationship between 2 firms

For further insight, consider comparing your results to what you would get with alternative thresholds of statistical significance (i.e. p <0.01)

Comparing the results of alternative thresholds of statistical significance (Fig 2 and Fig 3):

- Statistically significant links with p<0.05 = 48 (Fig 2)
- Statistically significant links with p<0.01 = 32 (Fig 3)

Let's compare few of the links which went missing when the significance level was tightened to p<0.01, compared to p<0.05:

- Link between Netflix (NFLX) and Costco (COST): Costco Wholesale is reportedly considering offering free streaming video to its executive level members, as mentioned in the link under citations. This could have led to Costco having a certain level of significant relation with Netflix, it being one of the biggest streaming channels. But, at the same time, company also reports that its streaming video efforts might get a bit uncertain, given their financials. Although, it has started recovering but even then, it is below the key level (its recent monthly high). This could be a possible reason that when we stricken the significance criteria, the relation between Netflix and Costco got removed, showcasing that they have a significant relationship, but the level of that significance is not very high

- Link between Amazon (AMZN) and Intuit (INTU): I mentioned above how Intuit has a dependent relationship with Amazon, especially relying on them for AWS. But as we all know that Intuit is a huge company in itself, focusing on developing and selling financial, accounting and tax preparation software and related services for small businesses, accountants and individuals. Therefore, it does have significant relationship with Amazon but not of very high significance level as it has its own huge line of business which is not dependent on or related to Amazon. Therefore, when the significance level is increased to $p<0.01$, the relation between Amazon and Intuit seems to fade away

Similarly, there are other connections also which were of comparatively low significance level and thus got removed when we increased the significance level to $p<0.01$ compared to $p<0.05$.

Citations:

- https://www.zdnet.com/article/dell-enters-public-cloud-space-with-vmware/
- https://searchitchannel.techtarget.com/definition/Dell-PartnerDirect
- https://www.amazon.com/Intuit-Software/b?ie=UTF8&node=497488
- https://www.businesswire.com/news/home/20171127005245/en/Intuit-Selects-AWS-Machine-Learning-Artificial-Intelligence
- https://www.investors.com/news/costco-mulls-free-streaming-video-top-level-members/