Inferring the Nasdaq 100 Network of Correlated Social Media Chatter

1. INTRODUCTION

Daily social media chatter might foresee future performance ahead of firms' reports on quarterly sales and cash flows. Some firms may have closely related patterns in the amount of social media chatter about them over time. The utilization of relational features intrinsic to a network can draw meaningful inferences from observed network data, which can be useful for link predictions between firms. In this advanced lab, we build and describe the undirected networks that link firms based upon the (partial) correlations in their Twitter activity. Twitter spike events may lead to sudden spikes in trading volumes, a common measure of market activity levels, in the stock for each firm.

In this network, two firms are linked if there is a statistically significant correlation in the daily number of Twitter messages that mention them. In this lab, the dataset used shows the number of Twitter messages, collected each day during financial trading hours, that mention each of the Nasdaq 100 firms in our dataset. We consider 92 firms listed in the Nasdaq 100, during 198 days of trading in the period from June 21, 2012 to September 18, 2013, as described in the paper by Tafti, Zotti and Jank (2015).

2. ASSOCIATION NETWORK INFERENCE

In Fig. 1 is shown a heatmap visualization of the Nasdaq 100 twitter chatter data. The firms (columns) have been ordered according to a hierarchical clustering of their corresponding vectors of twitter chatter levels. The activity levels of Nasdaq 100 firm pairs often can be expected to vary together. We see some visual evidence of such associations in the figure, where certain firms show similar twitter chatter activity levels across certain subsets of trading days. This fact suggests the value of constructing an association network from this data. Here we focus on two common and popular linear measures of association, correlation and partial correlation, and the corresponding methods for inferring an association network based upon them.

Now, to determine statistically significant links, we begin by finding the empirical correlations across all firm pairs in the twitter chatter data using *cor* function in R. Fisher's transformation is then applied to these correlation values to approximate the bivariate distributions and to determine the confidence intervals that are used to obtain *p*-values. Next, we apply a Benjamini-Hochberg adjustment, wherein *p*-values are adjusted through control of the false discovery rate. Comparing these *p*-values to a standard 0.05 significance level, we find a total of 180 statistically significant edge pairs between the Nasdaq 100 firms. Fig. 2 shows the density, CDF and false discovery rate of the empirical correlation coefficients for the twitter chatter data.

It is desirable to construct a graph where the inferred edges are more reflective of direct influence among vertices, rather than indirect influence, thus determining partial correlation becomes relevant. Subsequently, empirical partial correlation coefficients are computed to which Fisher transformation is applied, and again using Benjamini-Hochberg method to adjust the false discovery rate. Applying a nominal threshold of 0.05, we now get a total of 48 edges that are statistically significant between 56 firms. In other words, these edges are connections between firms for which there is a statistically significant correlation in the number of twitter messages between them.

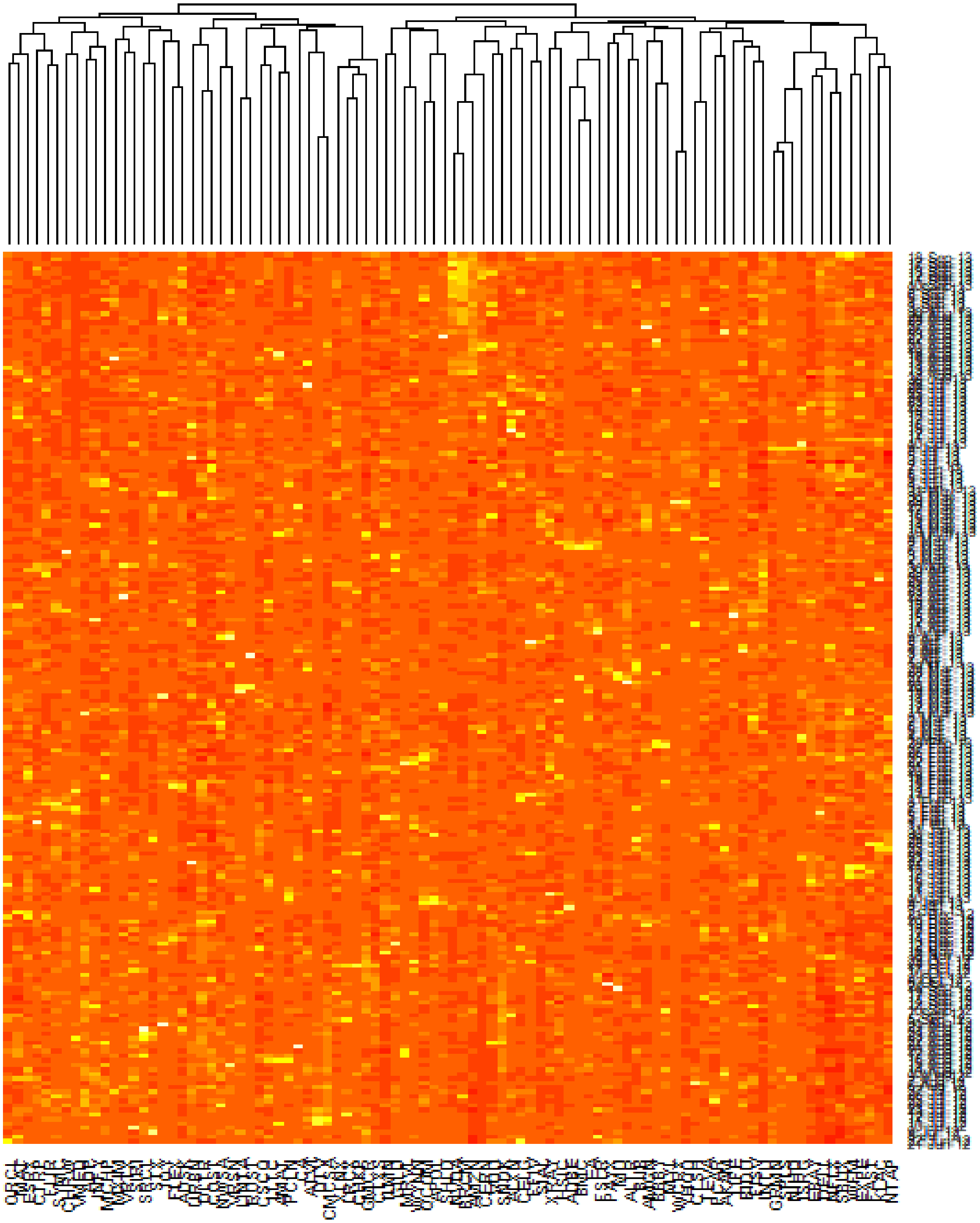
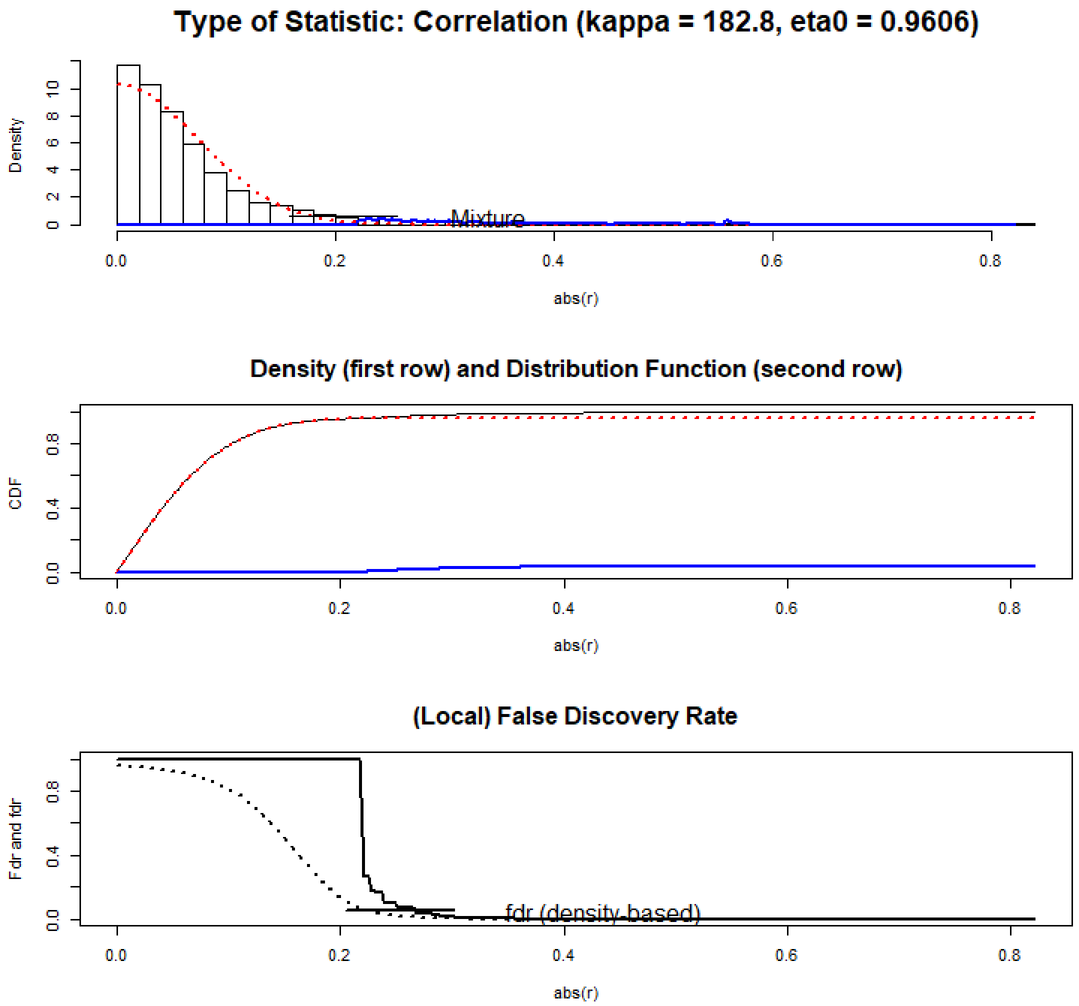


Fig. 1. Heatmap visualization of twitter chatter mentioning 92 Nasdaq 100 firms (rows) during 198 days of trading (columns)


Type of Statistic: Correlation (kappa = 182.8, eta0 = 0.9606)

Density (first row) and Distribution Function (second row)

(Local) False Discovery Rate

Fig. 2. Analysis of the empirical correlation coefficients for the twitter chatter data. Top: Estimated components f_0 and f_1. Middle: Estimated density f. Bottom: False discovery rate

The network graph of statistically significant relationships obtained above is shown in Fig. 3. This network graph shows only the nodes of degree greater than zero, i.e. only the nodes that are connected to at least one other node. The layout is obtained by applying Fruchterman and Reingold layout algorithm, which is one of the most used force-directed layout algorithms. The nodes in the graph are labeled with the corresponding stock ticker symbol. Further, nodes are colored based on their vertex degree. Also, the degree distribution is shown in Table I.

We can see that the firm AMZN (Amazon.com, Inc.) has the highest degree and it is connected to 6 other firms. It is connected to firms ADSK, DELL, INTU, NFLX, NVDA and SBUX. This result suggests that that a spike in chatter on Twitter about AMZN may signal an impending surge in trading activity of the firm’s stock. This surge may be similar for other firms which are predicted to be linked with AMZN in this network. This could be representative of a competitive relationship between the linked firms. In the media segment, AMZN competes with NFLX (Netflix), thus a link between them in this network makes sense. We can expect a similar level of information diffusion for these two firms. Further, firms with lesser degree in the network may not show a surge in trading activity which is comparable to many other firms in the network.

The viola plot in Fig. 4 shows that vertex pairs that are incident to each other (i.e., edge) have slightly fewer neighbors compared to the vertex pairs that are not incident to each other (i.e., no edge). This is evident as the number of statistically significant relationships are less.

TABLE I. Degree Distribution

Degree	1	2	3	4	6
# Firms	32	15	4	4	1

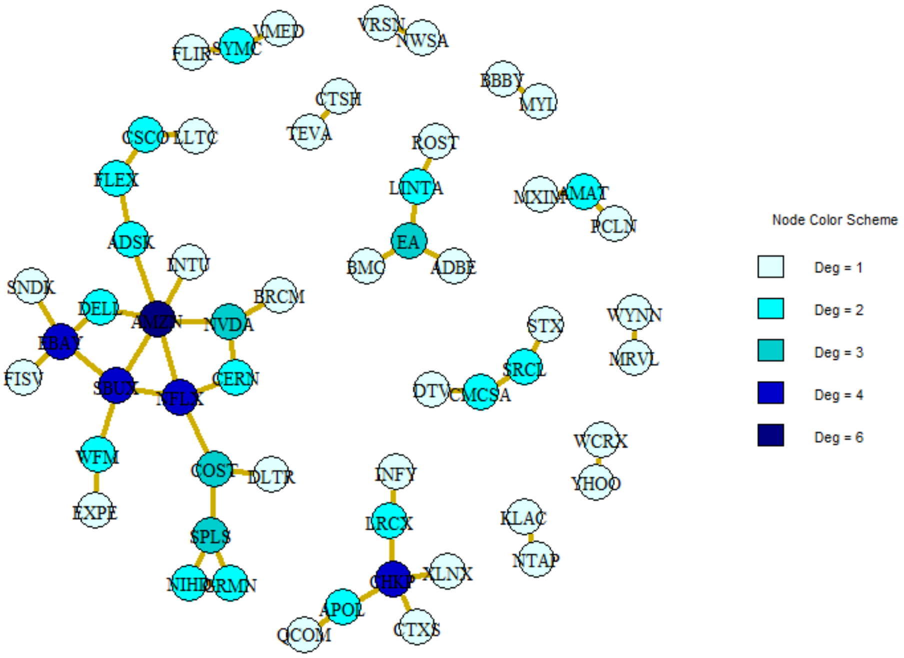


Fig. 3. Network of relationships among 92 Nasdaq 100 firms (using BH with $p < 0.05$ for partial correlation)

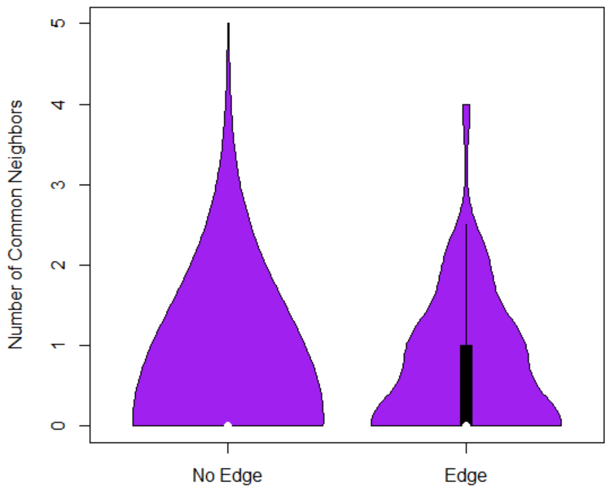


Fig. 4. Comparison of the number of common neighbors score statistic in the Nasdaq 100 network (using BH with $p < 0.05$ for partial corr.), grouped according to whether an edge is present between a vertex pair, for all vertex pairs

For further insight, comparing results with alternative thresholds of statistical significance (i.e. *p*-value < 0.01). We now get a total of 32 edges between 43 firms that are statistically significant, which is a subset of edges that were obtained in the original link prediction problem. The degree distribution is shown in Table II. The corresponding network graph and viola plot is shown in Fig. 5 and Fig. 6, respectively. Again, the firm AMZN has the most links but the degree has reduced to 4. It is now connected only to the firms ADSK, DELL, NFLX and SBUX.

TABLE II. Degree Distribution

Degree	1	2	3	4
# Firms	28	10	4	1

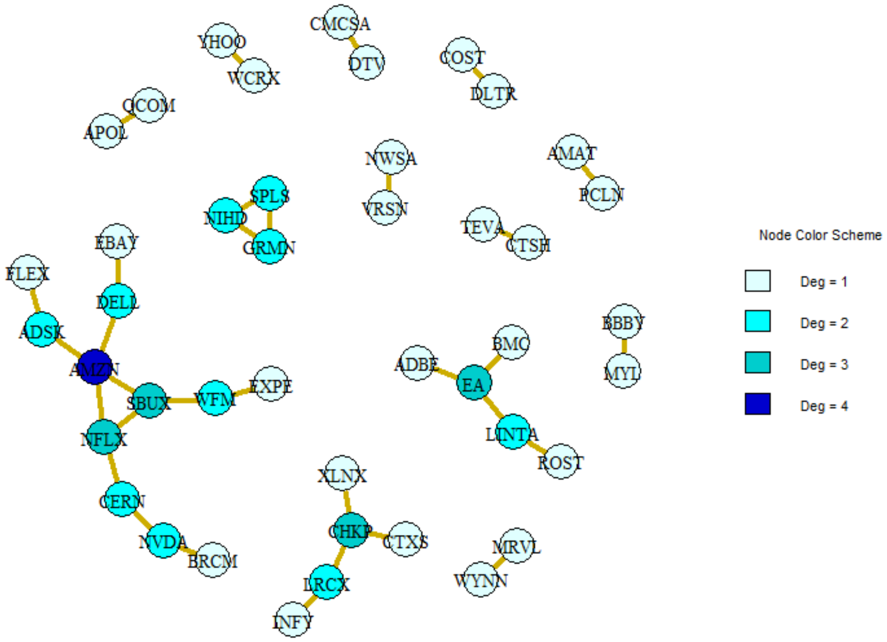


Fig. 5. Network of relationships among 92 Nasdaq 100 firms (using BH with *p* < 0.01 for partial correlation)

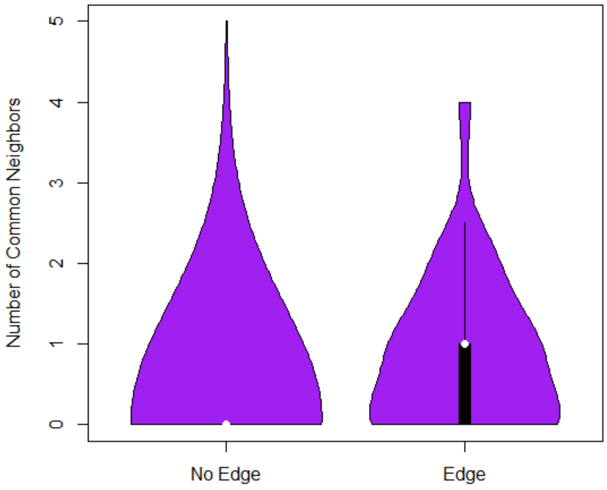


Fig. 6. Comparison of the number of common neighbors score statistic in the Nasdaq 100 network (using BH with *p* < 0.01 for partial corr.), grouped according to whether an edge is present between a vertex pair, for all vertex pairs

Next, we compare results of original link prediction problem with an alternative method for constructing the edges, such as using the overall correlation rather than partial correlations. We use Benjamini-Hochberg adjustment to control for the false discovery rate and a threshold of p-value < 0.05 to identify statistically significant overall correlations. We now get a total of 180 edges between 89 firms that are statistically significant. The degree distribution is shown in Table III. The corresponding network graph and viola plot is shown in Fig. 7 and Fig. 8, respectively. The firm, EXPE (Expedia Group Inc), now has the most links with degree of 13. It is connected to the firms ATVI, DELL, INTU, LLTC, MYL, NTAP, PAYX, ROST, SBUX, SNDK, WCRX, XRAY and YHOO.

TABLE III. Degree Distribution

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

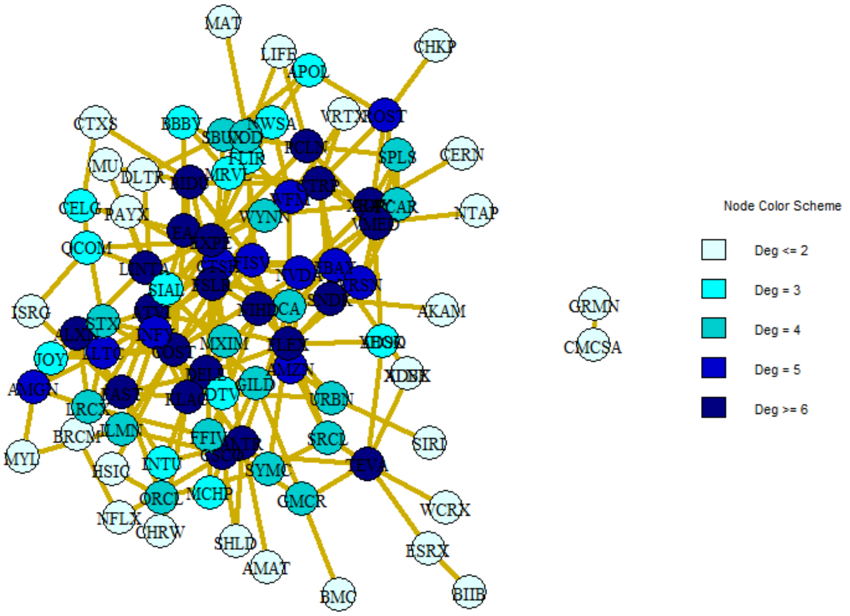


Fig. 7. Network of relationships among 92 Nasdaq 100 firms (using BH with $p < 0.05$ for overall correlation)

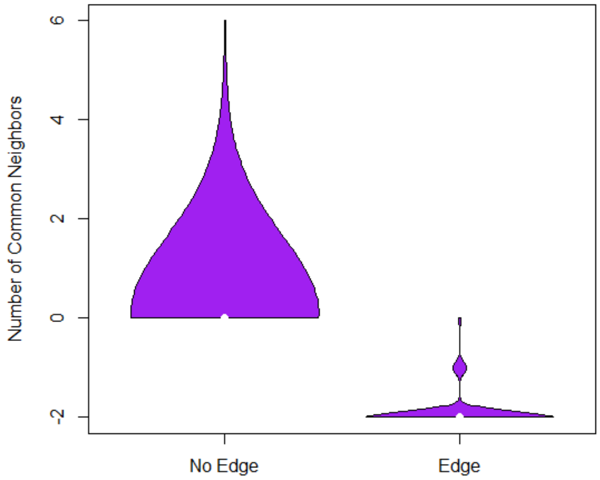


Fig. 8. Comparison of the number of common neighbors score statistic in the Nasdaq 100 network (using BH with $p < 0.05$ for overall corr.), grouped according to whether an edge is present between a vertex pair, for all vertex pairs

Alternatively, using *fdrtool* library in R to adjust the false discovery rate with threshold of *p*-value < 0.05 to identify statistically significant partial correlations. This yields almost the same number of edges as the original link prediction problem. We get a total of 47 edges between 56 firms that are statistically significant, which are in fact the same edges as obtained in the original link prediction problem, thus suggesting a certain robustness of our results. Only the edge relation between AMZN and NVDA wasn't predicted as a significant link. The degree distribution is shown in Table IV. The corresponding network graph and viola plot is shown in Fig. 9 and Fig. 10, respectively. Again, the firm AMZN has the most links but the degree has reduced to 5. It is now connected to the firms ADSK, DELL, INTU, NFLX and SBUX.

TABLE IV. Degree Distribution

Degree	1	2	3	4	5
# Firms	32	16	3	4	1

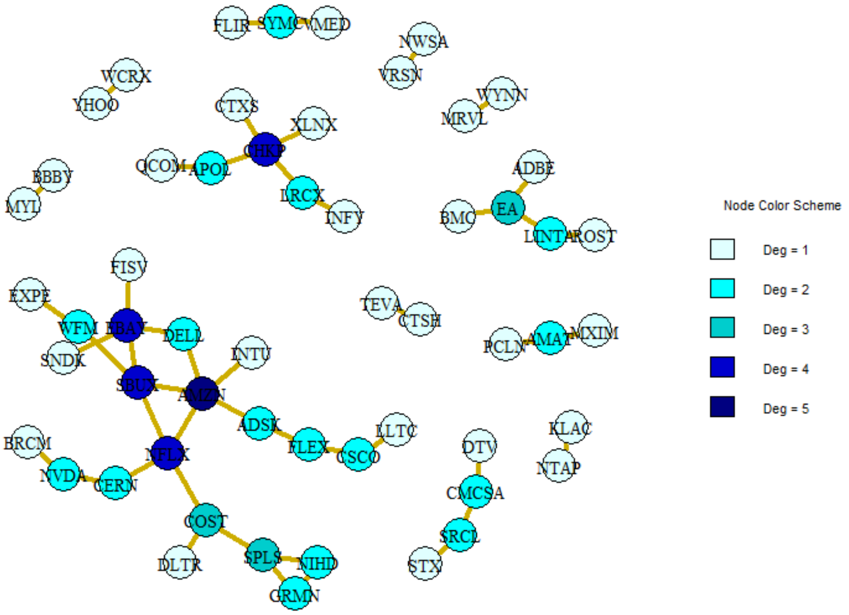


Fig. 9. Network of relationships among 92 Nasdaq 100 firms (using FDR tool with *p* < 0.05 for partial correlation)

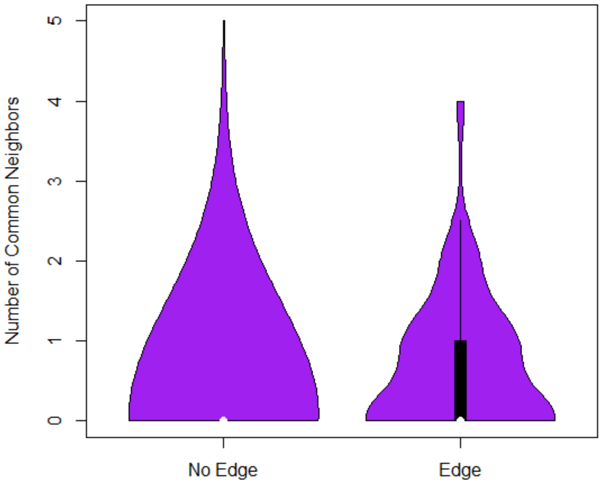


Fig. 10. Comparison of the number of common neighbors score statistic in the Nasdaq 100 network (using FDR tool library with *p* < 0.05 for partial corr.), grouped according to whether an edge is present between a vertex pair

Finally, we use the R package *huge*, ‘high-dimensional undirected graph estimation’, to arrive at a network graph. The main function *huge* generates an initial set of estimates, following several pre-processing steps that seek to transform the data to have marginal distributions close to normal and to stabilize the overall estimation problem. We now get a total of 109 edges between 77 firms that are statistically significant, which is a substantially higher number of edges than that obtained in the original link prediction problem. The degree distribution is shown in Table V. The corresponding network graph and viola plot is shown in Fig. 11 and Fig. 12, respectively. Again, the firm AMZN has the most links but the degree has increased to 10. It is now connected to the firms ADSK, BRCM, CERN, DELL, INTU, NFLX, NVDA, SBUX, VOD and YHOO.

TABLE V. Degree Distribution

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

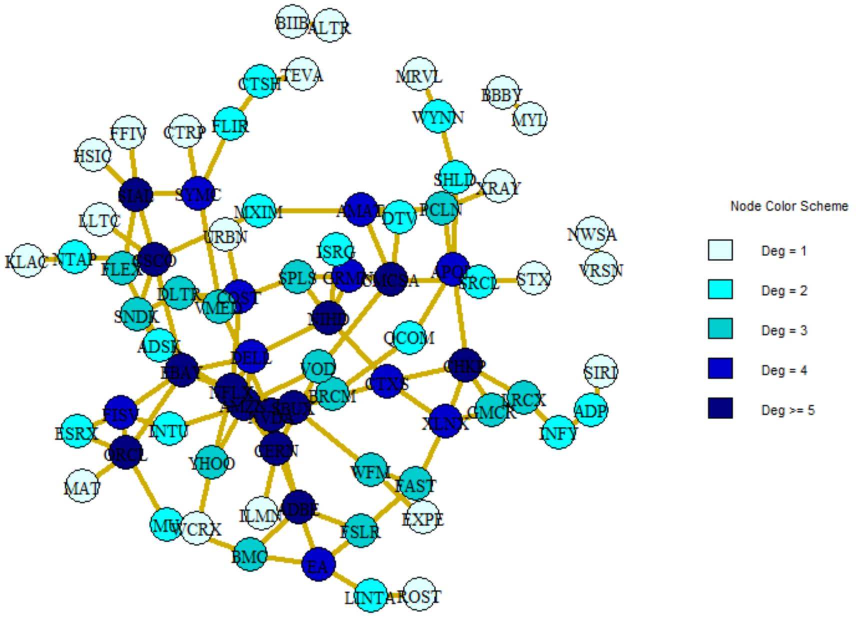


Fig. 11. Network of relationships among 92 Nasdaq 100 firms (using HUGE)

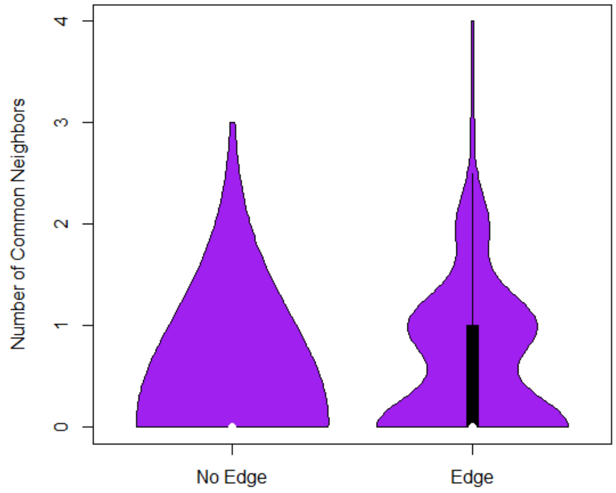


Fig. 12. Comparison of the number of common neighbors score statistic in the Nasdaq 100 network (using HUGE), grouped according to whether an edge is present between a vertex pair, for all vertex pairs

3. CONCLUSION

This analysis helps to learn about the impending state of financial markets and the similarities between companies through link predictions. The firms are linked if there is a statistically significant correlation in the daily number of Twitter messages that mention them. This not only presents a chance to harness valuable material for participants in financial markets, but also provides better understanding into the types of information that spread on large-scale social networks such as Twitter. This also helps us find about the pattern of information diffusion.

It is usually seen that the information about the firms spreads quickly on Twitter, but it takes a bit of time for financial traders to process and act upon that information. As such, signals propagating in Twitter may be useful to traders seeking to exploit small delays in the diffusion of news and the relatively slow responsiveness of the markets.

Link prediction between firms helps to recognize which firms are trending on social media. The statistically significant correlations may help understand the competitive relationships between firms. They may have things in common such as, their share prices might increase or decrease very frequently and so are more tweeted about; they are famous companies which are tweeted about more. All these factors have significant importance in inferring association networks.

REFERENCES

Tafti, A., Zotti, R., & Jank, W. (2016). Real-time diffusion of information on Twitter and the financial markets. *PloS one*, *11*(8), e0159226.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65). New York: Springer.

Li, Q., Zhou, B., & Liu, Q. (2016, July). Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* (pp. 359-364). IEEE.

Zuo, Y., Kajikawa, Y., & Mori, J. (2016). Extraction of business relationships in supply networks using statistical learning theory. *Heliyon*, *2*(6), e00123.

Filson, D. (2004). The Impact of E-Commerce Strategies on Firm Value: Lessons from Amazon. com and Its Early Competitors. *The Journal of Business*, *77*(S2), S135-S154.