IDS 564: Dr. Ali Tafti, Spring 2016

Advanced Lab 5, Link Prediction: Inferring the Nasdaq 100 network of correlated social media chatter

Some firms may have closely related patterns in the amount of social media chatter about them over time. In this advanced lab, you will build and describe the undirected networks that link firms based upon the (partial) correlations in their Twitter activity. In this network, two firms are linked if there is a statistically significant correlation in the daily number of Twitter messages that mention them.

As described in the paper by Tafti, Zotti and Jank (2015) (see below for the link to the paper), we retrieved Twitter messages over a time period of slightly more than one year. The attached dataset shows the number of Twitter messages—collected each day during financial trading hours—that mention each of the Nasdaq 100 firms in our dataset. Please note that since we used a specific keyword for our filter, this is not a comprehensive count of mentions on Twitter. However, the daily variation in activity and their correlations over time might suggest how temporal patterns of sentiment about firms may be connected with one another.

Please display the final network graphs, and discuss what those relationships suggest. Please use the methods from Lab 5. In your network graphs, show only the nodes of degree greater than zero; i.e. only nodes that are connected to at least one other node. Label the nodes in the graph with the corresponding stock ticker symbol. Also, resize the nodes for better visibility.

To determine statistically significant links, please do the following:

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.

What does the resulting network reveal about the relationship between the firms as they are described in Twitter? Do the edges appear to represent competitive relationships, cooperative relationships, or some other type of connection? How would you describe the meaning of the edges between nodes in the resulting network? Consider any characteristics of the firms that you may already know about or find out through an Internet search based on their ticker symbol, which is given in the column header of the dataset. For further insight, consider comparing your results to what you would get with

alternative thresholds of statistical significance (i.e. $p < 0.01$), or by trying an alternative method for constructing the edges, such as using the overall correlation rather than partial correlations.

Please download the related research paper from the PLoS One journal:

Tafti, Ali, Ryan Zotti, and Wolfgang Jank, "Real-Time Diffusion of Information on Twitter and the Financial Markets," *PLoS ONE* 11(8) 2016: e0159226.
Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159226