Visualizing Networks – Enron Links

Crystal Mosley

Northwestern University

Visualizing Networks – Enron Links

Summary—

After reading in all 367,662 email links, only 2 links were from the executive to him/herself.

Node 1 had a large network of communication traffic, but seems as though he/she were sending normal communications out to the company as a company-wide email message. Nodes 50, 49, 7, 53, 54, and 56 all have more incoming and outgoing email traffic, which would seem suspicious enough for investigating in further detail. Node 76 and 136 from the *top enron graph with the core* both have the most networks going through them; therefore, they would seem suspicious enough for further investigation as well.

Research design, measurement, statistical methods, and predictive models employed—

The Enron email links contained nearly 400,000 networks; therefore, this data set was chosen since it had the potential for many nodes and networks to visualize and experiment with. To ensure the data was clean, only non-zero nodes were utilized, and only nodes that were not self-referring were used. Different measurements were taken to experiment with the data:

- Only a subset of nodes that Node 1 communicated with

- Degree of centrality

- The most active nodes

- A subset of the top 50 executives based on node activity

- The two largest cliques of nodes – each clique at 13 nodes/executives

Layout models chosen to experiment with were: Fruchterman−Reingold, Kamada−Kawai, Circle, and Reingold−Tilford. Of these layouts, Fruchterman-Reingold and Kamada-Kawai layouts were almost the same display of nodes with the exception F-R being in more of a tight cluster. The Circle layout is completely useless due to the nodes being pushed to the edge and the network within the circle being a tight spiderweb—it's not useful for the viewer with this many nodes and networks because it does not show any detail as to which nodes these outside nodes were networking with. The best layout of these 4 is

Reingold-Tilford. This type of layout shows the viewer, in this Enron case, that node 76 had a high networking connection to a large subset of nodes—it is easily understandable and doesn't have a complex feel.

Overview of programming work—

The programming work that was utilized came from the "Jump Start Code" that was provided via assignment_3_jump_start_v002.R file. To get the code to work, the igraph, network, and intergraph libraries were utilized. Multiple model layouts were experimented with, but only Kamada-Kawai and Reingold-Tilford layouts were the most helpful for viewing the overall tops nodes with the most networks, as well as a more scaled down investigation for Node 1 and Node 76.