Evan Pliska

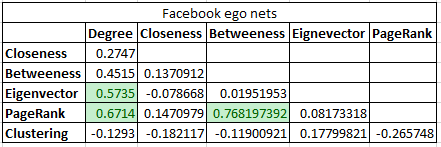
CSE 5245

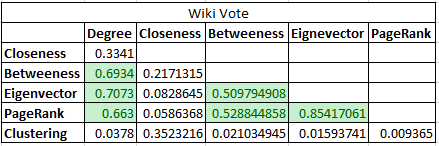
Lab 1 Results Analysis

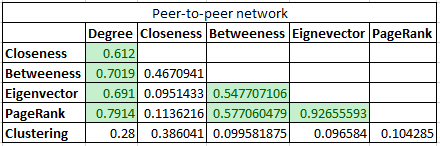
For this lab, I was asked to compute several measures of centrality (degree, closeness, betweenness, eigenvector, PageRank, and clustering) for each node in 4 given networks. The networks represented of a group of Wikipedia editors voting for on another for moderator approval, groups of quantum researchers who had worked together on research papers, a peer-to-peer file sharing network, and an ego-net of users of the Facebook app.

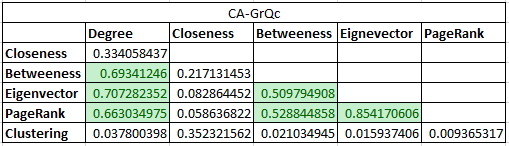
The first thing to note about computing the centrality measures for each node of these networks is the computation time. Degree centrality for each graph was computed almost immediately, whereas measures such as closeness centrality and especially betweenness centrality took much longer. This is to be expected since calculating the betweenness centrality for every node in an entire graph is the only represented algorithm that takes O(n^3) time, whereas the calculating the degree centrality for every node only takes O(n) time. The betweenness centrality for every node of each of these networks took between 5-10 minutes for graphs raging in size between 4039 and 7115 nodes.

The correlation coefficient for each pair of centrality measures for each network is given below in a series of tables. The pairs of measures who are strongly correlated (whose correlation coefficient is above 0.5) are highlighted in green.









In each of these networks, there is a strong correlation between Degree and Eigenvector Centrality, Degree and PageRank Centrality, and Betweenness and PageRank Centrality. In most of these graphs, Eigenvector centrality is very strongly correlated (correlation coefficient of 0.85 or higher) with PageRank Centrality. In general, PageRank and Degree Centrality seem to be more often more correlated with other centrality measures. Closeness and clustering Centrality were not strongly correlated with any other centrality measure.

To identify certain nodes as more important/central, we will seek to maximize as many centrality measures as possible. And to do this, we will sort the nodes of a graph by a centrality measure that is highly correlated with another measure, and then seek out nodes that have higher values of measure that are not correlated with any of the other measurements.

In the p2p network, we will first sort the nodes based on their Eigenvector Centrality, since its correlation coefficient with PageRank is 0.93, which is the highest of any pair of measure on any of these networks, and it is also strongly correlated with degree centrality. For this network, eigenvector centrality also has the weakest correlation with Closeness Centrality and the Clustering Coefficient (both the only values less than 0.1). So, once we have the nodes sorted by Eigenvector centrality, we will search for nodes with relatively high values for Closeness Centrality and Clustering. Such nodes for the p2p network include 7, 124, 145, and 856.

This method of sorting on eigenvector centrality and searching for nodes with high values of Closeness Centrality and Clustering Coefficient seems to be a good way to identify important nodes in a network given the available information. Obviously, if we had more measures of centrality then one might emerge as more useful; or if we had more or larger data sets to work with then we could say with more confidence which measures were more useful. Also, we have virtually no contextual information on our data, so we have no way to make a sanity check to ensure that the selected nodes are intuitively more central to the network. All that being said, selecting Eigenvector centrality as our primary centrality measure seems to be a good decision since it is usually correlated with Degree Centrality, Betweenness Centrality, and PageRank. This means that when we sort on Eigenvector Centrality, we are also taking into account (to some degree) 4 other measure of Centrality. It also provides the added bonus of not necessarily requiring us to compute betweenness centrality, which may not be feasible to compute for an entire graph if it significantly larger that the ones used in this lab. And again, Eigenvector Centrality tends to be very weakly correlated with Closeness Centrality and Clustering Coefficient, so if we find nodes whose values are high for these measures, it may be more of a ‘coincidence’ than finding nodes with high values for measure that are highly correlated. Essentially, this method is relatively simple, can be efficient, and tends to capture nodes that are considered central by all 6 of the selected measures.