**ASSIGNMENT 2/PROJECT 1**

**Brief on the data set:**

A network of books about US politics was published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co purchasing of books by the same buyers. The network was compiled by V. Krebs and is unpublished, but can found on Krebs' [web site](http://www.orgnet.com/).

Link: <http://www-personal.umich.edu/~mejn/netdata/polbooks.zip>

Nodes represent books about US politics sold by the online bookseller Amazon.com. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon. Nodes have been given values "l", "n", or "c" to indicate whether they are "liberal", "neutral", or "conservative".

**Approach:**

1. First step in this assignment was to create data files required for network analysis.

Data clean up and file formatting was done using R script and worksheet

Steps used for this purpose:

Load data file from Source 🡪 create raw data files in csv format 🡪 Use R script to get data in right format 🡪 create two csv files (1. Node 2. Edges) for data analysis



The following two files were created:



**Plan for Analysis:**

1. Each Node has one categorical variable book category which indicates the type of book

From this network analysis, we can learn which books are bought by the same buyer. Also, we can test whether such books have any relationship to the categories.

Following analysis will be performed in this data set:

1. Measurement of degree centrality (the number of connections to a node) and eigenvector centrality (measure of an influence of a node).

Sanjive>> (My take on Measurement of centrality)

* Degree centrality of top 10 books based on the highest centrality value and analyze the range of degree
* The above method will help us identify the gap between highest and lowest degree and to cover the gap we need to go to lower nodes and find the core of the Books dataset in the network. This can be done by trimming the dataset to 5 degree and derive closeness centrality between books
* Now, we also need to find out the books which are bridges and liked by all three audience category "liberal", "neutral", or "conservative". This can be derived by between centrality.

Once we have all the three centrality measure we can compare them across categorical groups.

* However, the above methodology still leave a room to and eigenvector centrality.

1. Comparison of centrality measures across categorical groups. We can perform a t-test to determine if the mean degree centrality of books from one category is higher/lower than the mean degree centrality of books from another category.

One outcome that is more likely to be expected is that individuals who buy books of one category are likely to buy books from that same category. It is not evident that a popular outcome would be an individual buying equal amounts of books throughout all three categories. However, this outcome will probably not be reflected in degree centrality. The likelihood of buying books from a similar category will not explain for example the difference in means of conservative books versus liberal books. What can be predicted, however, is the likelihood of which category will sell more books, since once a book is sold, the likelihood of another being bought increases.