**Assignment 4:** Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes. As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.). In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

There is an interesting data from snap.stanform.edu/data on Amazon’s movie review data. The dataset holds useful data such as product review, userID, helpful review. The data would be uploaded in Python and review data will be used to create nodes and edges. User nodes will consist of User id, user helpfulness and review score and Movie nodes will consist of Movie ID, movie name. Edges will be both inbound and outbound on user reviews, be User **reviews** and Movie—**reviewed by** Users. Edges will be both inbound (Users->Movie) and outbound (Movie->Users). The python program will calculate eigenvector centrality and betweenness centrality. Eigenvector would calculate if highly reviewed movies are connected to other highly reviewed and if they are more important in the network. Betweenness centrality would calculate shortest bridge to find a nodes importance due based on its connectivity.