Project 2: Identify Influential Actors in a Network

1) Download the "congress-Twitter" data from this **site**.
2) Apply centrality measures learned in class to identify the top five most influential nodes in the network. In particular, complete each of the following requirements:

- Overall Network Understanding: Analyze the network's order, size, density, strong connectivity, and so on.
- Degree Distribution: Plot histograms to visualize the distribution of the nodes' degrees (total, in-degree, and out-degree). You may choose to plot it using one of the approaches that we mentioned in class.
- PageRank: This algorithm identifies influential nodes based on the idea that a node is important if it is linked to by other important nodes. It considers the quality of incoming links.
- Hub and Authority Scores: Hubs are nodes that point to many authorities, while authorities are nodes that are pointed to by many hubs. This approach identifies nodes that are good at providing information (authorities) and those that are good at linking to informative nodes (hubs).
- Closeness Centrality: This measure reflects how close a node is to all other nodes in the network. It can be calculated using closeness or harmonic centrality. Closeness centrality is the reciprocal of the average distance to all other nodes, while harmonic centrality is the sum of the reciprocals of the distances to all other nodes. (If your network is strongly connected, then use closeness; otherwise, use harmonic centrality.)
- Betweenness Centrality: This metric identifies nodes that lie on many shortest paths between other nodes. Nodes with high betweenness centrality act as bridges in the network.

Here are some factors to consider when making your selection:

- **High PageRank:** Nodes with high PageRank scores are likely to be influential as they are linked to by other important nodes.
- **High Authority Score:** These nodes are considered experts or authorities in the network as they are referenced by many other nodes.
- **High Closeness Centrality:** Nodes with high closeness centrality are well-connected and can efficiently reach other nodes in the network.
- **High Betweenness Centrality:** These nodes act as bridges in the network and control the flow of information between different communities.

3) Justification

Write a paragraph (at least 200 words) justifying your selection of the top five influential actors based on the centrality measures. Explain how the specific values of each measure contribute to their influence within the network.

The data in the file "congress.edgelist" is not format ready. The following shows how to reformat it into an edge-list with the right format.

```
# Set working directory and load data
setwd("/Users/szs0398/Library/CloudStorage/OneDrive-AuburnUniversity/Teaching/25S_5740-6740/R-files")
D <- read.csv("congress_network/congress.edgelist")

# Verify data structure
print("Data structure:")
```

```r
str(D)
print("First few rows:")
head(D)

# Load the library
library(igraph)
library(stringr)

# Format the data
edges_df <- data.frame(
  from = as.numeric(sub("^(\\d+).*", "\\1", D$X0.4...weight...0.002105263157894737.)),
  to = as.numeric(sub("^\\d+\\s+(\\d+).*", "\\1", D$X0.4...weight...0.002105263157894737.)),
  weight = as.numeric(sub(".*'weight': ([0-9.]+).*", "\\1", D$X0.4...weight...0.002105263157894737.))
)

# Create the graph
g <- graph_from_data_frame(edges_df, directed = TRUE)



# Create the graph
g <- graph_from_data_frame(edges_df, directed = TRUE)
```