Stephen Phillips Project 1

Dataset:

The dataset that I've chosen is from https://snap.stanford.edu/data/gemsec-Deezer.html. This dataset contains information collected from the music streaming service Deezer. The dataset contains friendships between users located in Europe. Nodes are the users and edges are the friendships. The genre preferences of the users are also available as metadata in the network as lists of genres. 84 distinct genres are available.

Loading Data:

This data set contains 3 different CSVs, with 1 for each country that the data was collected from. The three countries are Hungary, Croatia, and Romania and the CSVs contain the edges for the network. In addition to this, there is a JSON file for each country which contains the musical taste metadata for each node. The format for this JSON is a dictionary with the node number as the key and a list of the genres for the item. Loading this data should be a simple task, as the edges can be loaded in as a dataframe and the metadata is in a format that would be easy to iterate over to add to each of the nodes.

Analysis:

I think that there are many interesting ways that this data could be analyzed. A few of them:

- look at individual genres and see which one has the most interconnected nodes
- look at relationships between the amount of genres that someone listens to, and their amount of connections
- determine if people that are friends have the same musical taste

Based on these three ideas, we could make comments on how different genres contain different types of people. It is possible that genres which are more marketed to outcasts (perhaps stuff like metal music) will have less connections and less centrality.