DATA 620 Week 4 Assignment

Centrality Measures

Shyam BV

Question:

Centrality measures can be used to predict (positive or negative) outcomes for a node.

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.

Solution:

Dataset:

For this assignment, I have used the world trade dataset which is provided in the link http://vlado.fmf.uni-lj.si/pub/networks/data/esna/metalwt.htm. The format of this file is in pajek(.paj). Here the data is related to the world trade in miscellaneous manufactures of metal in 1994. Dataset has the trade volume between different countries.

From this dataset, we will be finding out the country with higher centrality. Also explain about the importance and degree of trade by that country.

Technology stack:

Technology which will be used for performing this network analysis will be of following.

1. Python language
2. Packages - Networkx, matplotlib, pandas and numpy
3. Visualization or Processing(optional) - Neo4j

Data cleansing:

1. As the data is in pajek format, we need to cleanup and load as a graph in python. There is a parser (nx.read_pajek) which is available in networkx package. We can use that parser and create a graph object for the dataset.
2. By reading the dataset, it has a country field but it does not have region parameter. If required we can join with https://github.com/lukes/ISO-3166-Countries-with-Regional-

Codes/blob/master/all/all.csv and get the region of it. For that graph needs to be converted to dictionary and then to pandas, join the data and convert back to graph.

3. Once we have the final graph, we can use it for all the future analysis

Exploratory Analysis:

As an exploratory analysis, we can plot the countries which have only one degree (one country as trade partner). Also countries which has highest degree (country which has deals with multiple countries).

Calculating centrality metrics:

Once we plot all the countries in a graph, we can find out different centrality measures.

1. Degree centrality
2. Closeness centrality
3. Betweenness centrality
4. Eigenvector centrality

Degree centrality:

This metric is to find out the number of connections or trade a country makes with all the different countries which is divided by all the total number of countries. This will should the countries which has most trades with different countries. This can be calculated by the function net.degree_centrality(G) in networkx package.

Closeness centrality:

This type of centrality metric provides the closeness or shortest path to all the nodes from ego (or particular country). This can be calculated by the function net.closeness_centrality(G).

Betweenness centrality:

This metric is more important for finding the important country which links other countries. By finding this we can find the trade partners which links different regions. Also the countries which buys and sells to other countries.

Eigenvector centrality:

This centrality is the iterative approach of degree centrality to find out the Gray Cardinal in the complete trade graph. It can also be found by net.eigenvector_centrality(G) method.

Summary:

Once we calculate all the centrality, we can create a final combined data frame which consists of countries, region and all four centralities. Final top 10 countries are the top important countries in the entire trade group. This analysis is helpful to find the countries which has lot of trading partners and has huge leverage in this type of trade.