Assignment 1 **Social Media Analytics CS G519** **Max Marks: 15**
Due date: 19th November 2021

Community Detection - Empirical Study (10%)

In this problem we will compare the different community detection algorithms studied in class for multiple real world social networks.

1. Download and **read** the following graphs:
   • Karate Club network (http://www-personal.umich.edu/~mejn/netdata/karate.zip)
   • Dolphins social network (http://www- personal.umich.edu/~mejn/netdata/dolphins.zip)
   • Jazz musicians network (http://deim.urv.cat/~alexandre.arenas/data/xarxes/jazz.zip)

2. **Statistics**
   Compute the following statistics describing the datasets:
   • number of nodes n,
   • number of edges m
   • average path lengths $\bar{d}$
   • average clustering coefficient C
   Present your results in a table with the datasets as rows and the statistics as columns.

3. **Implementation**
   (a) Write a program that computes betweenness-based clustering using the Girvan-Newman algorithm).
   (b) Write a program that computes modularity-based clustering (modularity maximization).
   (c) Write a program that computes spectral clustering (using the graph Laplacian).

4. **Comparison**
   Run all three methods on the three datasets (KARATE, DOLPHINS, JAZZ) and report the following for all methods and datasets:
   • number of clusters found
   • modularity score for this clustering
   • run time of the algorithm (HINT: to get this number re-execute your computation 5-10 times and take the mode runtime).

Report
Submit assignment report along with your code. The report should include results of the above questions and the following:

Write your observations according to these findings that which algorithm performs the best w.r.t. both efficiency and quality.

Justify the quality of your results by visualizing the clusters. Use Gephi (https://gephi.org) – an open source graph visualization and analysis tool – and try different layouts and experiment with node and edge coloring, sizes, etc. Add a careful description on what your visualization shows.

BONUS

The representative network is a network that uses one node per cluster (the node size should represent the community size) and edges between clusters with weights reflecting the number of cross-community edges between nodes in the respective clusters. Create and visualize the representative network for one community detection method for each dataset. Submit your program (creating the representative networks and the plot)

Instructions

- Write name of group members (max 2 members) in your report.
- Code and report should be properly named as q1.py, q2.py,.. and bonus.py.
- Comment your code to receive maximum credit.
- Submit your code through Github. Add the repository name to your report.
- Reports should be submitted through Nalanda.
- Late submissions will attract penalty.