CS 5409: Applied Social Network Analysis **Fall 2024**

Assignment 3

Due 11:59 PM November 12, 2024

General Instructions [Read Carefully]:

- This assignment is to be completed individually (no collaboration is allowed). This is **NOT** a group assignment. You can discuss the problems among your peers, but you should attempt to solve the problems by yourself.

- Please share a **Google Colab notebook** containing all the codes. Note that the notebook should be self-sufficient. You can add one extra folder "results" where you will put all the output file. Do not upload the data but must specify the path clearly so that the results can be reproduced.

Questions:

1. Implement the HITS and Pagerank algorithms. Run these algorithms on the Twitter interaction network for the 117th United States Congress (data is available at https://snap.stanford.edu/data/congress_network.zip . Please use the congress.edgelist file for the network. Disregard the edge weights.)

   a) Plot the distribution of Hub and Authority scores of the nodes obtained from HITS algorithm
   b) Compute the pearson's correlation coefficient between the hub scores computed by the HITS algorithm and the pagerank scores
   c) Compute the pearson's correlation coefficient between the authority scores computed by the HITS algorithm and the pagerank scores.

   [20 points]

2. Implement the Girvan-Newman community detection algorithm[1] and find the community structure of the social network of LastFM (dataset available at: https://snap.stanford.edu/data/lastfm_asia.zip . Please use the lastfm_asia_edges.csv to obtain the edgelist of the network.)

   [1]Girvan and Newman 2002, "Community structure in social and biological networks". (https://www.pnas.org/doi/epdf/10.1073/pnas.122653799)

   You need to provide a) number of communities obtained b) distribution of the community sizes.

Also visualize the community structure (separate color for separate community) if you run the algorithm on Zachary's karate club network (available in networkx library). You need to also provide the hierarchical tree after complete pass of the algorithm.

[10+5 = 15 points]

3. Apply the node2vec (https://snap.stanford.edu/node2vec/) algorithm on the following dataset (https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz - read the "README" file carefully. You can build the network from cora.cites file where each line represents <ID of cited paper> <ID of citing paper>. For the class labels, you need to consider cora.content file) to generate the node embeddings of various nodes. Then use logistic regression classifier on the obtained node embeddings as feature. Report the Macro-F1 score and Accuracy. Use 80:20 split of the data for training and testing and default parameter settings as d = 128, r = 10, l = 80, k = 10, p = 1 and q = 1 [p - in-out parameter, q - return parameter, r - number of walks per node, l - walk length, k - neighborhood size, d - dimension of the feature vector]

   Plot how Macro-F1 and Accuracy will change with varying parameter values in the node2vec algorithm. You are required to provide the following plots

   a) Macro-F1 (Accuracy) vs $\log_2 p$ [p = 0.25, 0.5, 1, 2, 4]

   b) Macro-F1 (Accuracy) vs $\log_2 q$ [q = 0.25, 0.5, 1, 2, 4]

   c) Macro-F1 (Accuracy) vs $\log_2 d$ [d = 16, 32, 64, 128, 256]

   d) Macro-F1 (Accuracy) vs r [r = 6, 8, 10, 12, 14, 16, 18, 20, 24]

   e) Macro-F1 (Accuracy) vs l [l = 40, 50, 60, 70, 80, 90, 100]

   f) Macro-F1 (Accuracy) vs k [k = 8, 10, 12, 14, 16, 20, 24, 30]

   [15 points]