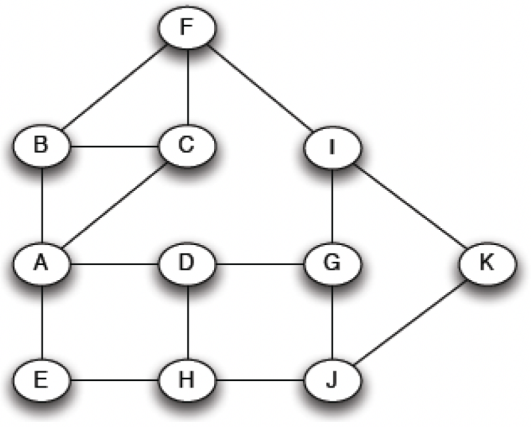
THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4641: Social Information Network Analysis and Engineering
Spring 2023 Assignment 2
Due time and date: 11:59pm, Apr 12 (Wed), 2023.

IMPORTANT NOTES

- **Your grade will be based on the correctness and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **ZERO-Tolerance on Plagiarism: All involved parties will get zero mark.**

Q1: Girvan-Neuman Algorithm

Given the following graph. consider the edge betweenness contributions for paths starting from node C. You are required to proceed as in the lecture notes. You can put all the answers on one single diagram.



1. draw the BFS result for paths starting from node C.
2. obtain the number of shortest paths from C to each node. show this number next to each node.
3. obtain the betweeness contributions for each path. show this number next to each edge.

Q2: Node Embedding and Clustering

In this question, we use the Deep Graph Library (DGL) on the Zachary karate club dataset[1]. The graph neural network used is called GraphSAGE ("Inductive Representation Learning on Large Graphs", William L. Hamilton, Rex Ying, Jure Leskovec. NIPS 2017), which is implemented as dgl.nn.SAGEConv in DGL (see `https://docs.dgl.ai/en/0.8.x/generated/dgl.nn.pytorch.conv.SAGEConv.html` for details).

The GraphSAGE (with mean aggregator) uses the following update scheme:

$$\mathbf{h}_v^k = \sigma\left(\left[\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}, \mathbf{B}_k \mathbf{h}_v^{k-1}\right]\right),\ k = 1, 2, \ldots, K,$$

where $[\cdot,\cdot]$ denotes concatenation. In other words, instead of adding $\mathbf{W}_k \sum_{u\in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}$ and $\mathbf{B}_k \mathbf{h}_v^{k-1}$ together as in the lecture notes, GraphSAGE concatenates them together. Partial code has been provided in `https://colab.research.google.com/drive/1NP4dZmXk3tHtbRFC28u6MMtfQe1bj6jl?usp=sharing`.

[1]`http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm`

1. Build the GraphSAGE model with two GNN layers, and ReLU as activation function at the first layer. To train the node embedding, use the adjacency based similarity function in the loss (page 13 on nrltutorial-part1.pdf).

2. Draw the node embedding using Matplotlib (similar to the output on page 3 (right) of nrltutorial-part1.pdf).

3. Use the k-means clustering algorithm sklearn.cluster.KMeans (details in `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`) to partition the graph into two subgraphs. Show the nodes in the embedding space in red and blue.

4. Repeat the 3 steps above, but with 3 GNN layers and 2-hop similarity function (replacing the adjacency based similarity function in step 1 with 2-hop neighbors) in the loss (page 18 of nrltutorial-part2.pdf).

Q3: Node Classification

In this question, we use DGL on the Cora dataset[2]. The task is to train a model to predict the ground truth category of each node. Some partial codes have been provided in `https://colab.research.google.com/drive/1NP4dZmXk3tHtbRFC28u6MMtfQe1bj6jl?usp=sharing`.

1. Build the model based on GraphSAGE model with two GNN layers, and ReLU as activation function at the first layer. Since this is a classification task, we add a linear layer on top as the classifier. The loss function is the cross-entropy loss. Report the testing accuracy.

Submission Guidelines Please submit one Python notebook (A1.ipynb) and a report (report.pdf) for your results and conclusions. The submitted folder should be zip all the files into A1_awangab_12345678 (replace awangab with your ust account and 12345678 your student id). Please submit the assignment by uploading the compressed file to Canvas. Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. Plagiarism will lead to zero point on this assignment.

[2]https://graphsandnetworks.com/the-cora-dataset/