Dataset Size and Organization:

There are three datasets in this stage: cora, citeseer and pubmed, each of which represents one directed graph. All these datasets are composed of two files: node file, and link file.

Cora
In pubmed dataset, the node file has 2,708 lines, each line denoting one node. For the nodes in cora, each line has 1,435 elements, which are organized as follows:

<node_id> <1433 node_features> <node_class_label>

The first element denotes the node index. (Each node has one unique index, which will also be used in the link file denoting the nodes). The last element (a string) denotes the node label, and the mid 1433 elements denote the features of the node.

The link file has 5,429 lines, each line denoting one directed link (A B, i.e., it denotes B pointing to A).

Citeseer
In the citeseer dataset, the node file has 3,312 lines, each line has 3,705 elements denoting one node (<node id> <3,703 node features> <node label>). The link file has 4,715 lines, each line denoting one directed link (A B, i.e., it denotes B pointing to A).

Pubmed
In pubmed dataset, the node file has 19,717 lines, each line has 502 elements denoting one node (<node id> <500 node features> <node label>). The link file has 44,324 lines, each line denoting one directed link (A B, i.e., it denotes B pointing to A).

Training Testing Set Partition:

Prior to data partitioning, please open (or print out) the data files to take a look at the data first. Please pay attention to the node index representation, node feature representations, element separation, and node label representations.

For cora, it has 7 different classes (Case_Based, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning, Rule_Learning, Theory). Please randomly sample a training set with 140 nodes (20 node instances per class), and evaluate the learned model on a randomly sampled testing set with 1050 nodes (150 node instances per class).

For citeseer, it has 6 different classes (AI, Agents, DB, HCI, IR, ML). Please randomly sample a training set with 120 nodes (20 node instances per class), and evaluate the learned model on a randomly sampled testing set with 1200 nodes (200 node instances per class).

For pubmed, it has 3 different classes (0, 1, 2). Please randomly sample a training set with 60 nodes (20 node instances per class), and evaluate the learned model on a randomly sampled testing set with 600 nodes (200 node instances per class).

Task To Be Done:

Please train a GCN on each of these datasets, and evaluate the learned model performance on the testing set. Please include your model learning performance in the report.