Assignment 1

Learning with Graphs (CS 768)

1 Logistics

Date of release: **September 5th 2020**
Submission due: September 19th, 2020, 22:00

This assignment is to be submitted individually by each student. It will be graded for a total of 40 marks, the relative weightage is shown in the next section.
Your assignment submission should include your code and report. The report should be a PDF reporting all the required numbers and comments. The code can be written in any language but you should include 'run.py' or 'run.sh' in your home folder which takes as an argument the path to the dataset file containing edge list. When run, it should write to STDOUT the MAP, MRR values when using different measures for **'test_fraction=0.8'**.

Example output:

```
$ python run.py dataset/facebook.txt
Using test fraction=0.8

Adamic Adar
MAP=XXX
MRR=XXX

Common Neighbor
MAP=XXX
MRR=XXX

Preferential Attachment
MAP=XXX
MRR=XXX

Katz
MAP=XXX
MRR=XXX

Training logistic regression
train status...
MAP=XXX
MRR=XXX
```

Your submission file should be zipped and named rollno.zip.

2 Tasks

In this assignment we will evaluate different measures for predicting edges in a Social Graph.

1. Download *facebook.txt* from this link.
2. Split the edge list into train and test by X% and (100-X)% with X=60, 70, 80, 90. Download the sample code from this github link and use it to load the dataset file and make splits. **Note: This step is not optional. Since we want everyone to work on the same data, you should use this code to make the train-test split.**
3. Use Adamic Adar, Common Neighbor, Preferential Attachment and Katz measure for predicting the missing edges. Report MAP (Mean Average Precision) and MRR (Mean Reciprocal Rank) for each split and measure. [32x0.5=16]
4. Comment on which edge-prediction measure does the best and why. [4]
5. Comment on which evaluation measure is well suited according to you. [4]
6. Do you see consistent results when using MAP or MRR, i.e. is the same prediction method remains the best or worst irrespective of whether you use MAP, MRR? If not, why? [4]
7. Train a logistic regression model that predicts the edge score using the four measures in point 2 as features. Report MAP and MRR again on the test set for each fraction. You can use a sample of the train data as validation for any hyper-parameter tuning, you should not use test set for any kind of optimization. [8x1=8]
8. Does this linear predictor model fare better than any of the individual predictors? Comment. [4]