Project 2

1. Link prediction on real large-scale social media data.

2. Train, Dev, and Test data are provided.

3. Similar to our tutorial, AUC-ROC is used as the evaluation metric.

Project 2

We provide:

Train data (‘train.csv’): positive edges

Dev data (‘dev.csv’): positive/negative edges with their labels.

Test data (‘test.csv’): positive/negative edges without their labels.

Marking threshold based on the dev data will be released next tutorial.
(Similar to project 1, this is just for your reference.)

You need to submit:

(1) The ‘test.csv’ file with the ‘score’ column filled.
(2) A zip file that contains all the code for us to duplicate your result.
(3) A brief report.

PS: you should fill in the plausibility scores for all the candidate node pairs in the ‘score’ column of the `test.csv’ file. (higher score indicates that an edge is more likely to exist between these two nodes).

Project 2

Other information:

1. Due time: Apr. 7th 23:59 HK time

2. You are welcome to use any methods to make the prediction, but you should at least try the embedding methods, which typically achieve better performance than the traditional methods.

3. The methods taught in the class/tutorial are enough for you to get the full marks.

4. Late submission policy is same as project 1.

5. Your final grade is automatically calculated based on the AUC-ROC score of your submission file with the gold mention labels. So please do not change the order of test node pairs in the `test.csv’.

6. At current stage, the direction of edges is not considered for this project, which means that the reverse of a positive edge is not considered as a valid negative edge, but you should keep that in mind in the future.

7. Like the tutorial, you may meet unseen nodes in the prediction, you should make proper prediction for them. (random guess or 0? Depends on you)