INFS7450 Social Media Analytics

Project 2 – Link Prediction

Semester 1, 2025

Marks:	15 marks (15%)
Submission Due:	27 May 2025 16:00 (Brisbane Time)
Deliverables:	See deliverables part
How to submit:	Electronic submission via Blackboard

Goal: This project aims to design and implement an effective algorithm to predict missing links. The implemented algorithm will be evaluated on a real-life graph dataset, and its performance will be reported. Students are required to finish this project individually.

Dataset: The dataset for this project comprises three primary files: "trainingset.csv", "testset.csv", and an example submission file "SubmissionExample.csv".

- **trainingset.csv:** This file represents an undirected graph's edge list. Each row corresponds to an edge between two nodes in the graph, indicating the presence of a relationship between them.
- **testset.csv:** This file includes 100 positive edges (actual links between nodes) and 900 negative edges (no actual link between nodes). The challenge for the students is to find the top-100 edges that are most likely to exist.
- **SubmissionExample.csv:** This file provides an example format for submission. It consists of 100 rows, reflecting the top-100 edges predicted by the student's model.

The dataset is available from UQ blackboard. See /Assessment/INFS7450 Project Two.

Tasks:

1. Predict the missing links for the given graph. (**15 marks**)
   You are required to develop a link prediction model using the "**trainingset.csv**" file. The model should then be applied to the "**testset.csv**" file to predict the top-100 most likely edges. The predictions must be submitted in a CSV file, formatted according to the "**SubmissionExample.csv**"

Requirements:

1. There is no restriction on algorithms, packages.
2. **Failure to properly cite AI contributions is a form of academic misconduct.**
3. You are not allowed to look at the code of any other student. **All submitted codes and reports will be subject to electronic plagiarism.**

Programming Languages:

1. Python and NetworkX are recommended. However, you have your own choices of preferred programming languages including, but not limited to, Python, MATLAB, Java, C, C++, etc.

Deliverables (!!VERY IMPORTANT):

Your submission should be compressed into a zip file named after your student's eight-digit ID number (e.g., 12345678.zip), organized as follows:

```
12345678.zip/
├─── Data/
│    ├─── trainingset.csv
│    └─── testset.csv
├─── Code/
│    ├─── main.py
│    └─── utilities.py     # This file contains additional functions
├─── 12345678.csv # This is your prediction results file, formatted according to SubmissionExample.csv
└─── 12345678.pdf       # This is your report file
```

Submitting files in an incorrect format can lead to deductions in your overall project score. It's crucial to adhere to the specified submission guidelines, including file types, naming conventions, and organizational structure.

Marking criteria (Total marks: 15):

- 15 marks = 4 marks (code) + 7 marks (results) + 4 marks (report)
- Your results should be reproducible and your codes should be readable. If your codes cannot be executed or generate the results as reported, the corresponding marks for the code and results will be deducted.
- Accuracy is used as the evaluation metric for your submission as follows:

Accuracy = The number of correctly predicted edges/100.

Accuracy is calculated based on your submitted results compared against the ground truth.

- **Time Limitation:** The prediction running time of your code must be less than an hour. If you have employed a trainable model, such as a deep neural network, please ensure to include the trained parameters/weights within your code submission. These parameters are critical for reproducing your results and for evaluating the performance of your model. Otherwise, the score for your code part will be 0.

Kaggle:

We will be deploying our projects on the Kaggle platform.
You can join the Kaggle competition via the below links:
https://www.kaggle.com/t/e32f03c8fe8648cfb520cfdd8432dcea

Please note that the scores on this platform will not be considered in the final grading. The final score will still be determined by the results you upload to Blackboard.