**SNA HW1 - Link prediction M10615103**

1. **The Introduction, concept or algorithms of your methods.**

這次的作業是Link prediction，我透過period1中的現存網路資訊，去計算node之間的feature值；有了feature來源之後，再用period2 中的新增edge當作正 train data, 此時的無edge處當作負train data，獲得正負label之後去做訓練model；訓練結束之後，即可用該model去預測test data是否有edge。

演算法部分：

**資料前處理**

1. 首先先把period1中的資訊(有連線的edge)存起來，使用networkx in python的架構儲存資料。
2. 接著用內建的Jaccard/Preferencial attachment 函式，計算出兩兩node之間的該兩個feature值，存在各別的adjacent matrix中。

**訓練model**

1. 處理完period1 中之資訊後，開始用period2中的node資訊，從adjacent matrix提取出Jaccard/Preferencial attachment 值，並label為 1。
2. 負採樣部分，從period1 + period2 的graph中，找到沒有連接的edge，提取出該edge之Jaccard/Preferencial attachment 值，並label為 0。
3. 接著就把二維的資料點，當作train data，還有相對應的label 作為 score 送入 Logistic regression的model中。

**預測testData**

1. 讀取testData中的node資訊，提取出相對應的feature值，接著將該數對使用model的predict函式來取得結果。
2. **The Difficulties encountered that you have facing in data analysis and prediction, you also emphasize the methods that you try to resolve and its results.**

遇到的問題有以下幾點：

* 資料的id非規律：一開始想直接用id的值去取得feature矩陣中的值，才發現不太對，因為取得feature矩陣要使用到的是矩陣的index(0~….)，index對應到的不是id，而是他在矩陣中的次序位置。我的處理方式，是利用node\_information的資訊，把每一篇的id，依序轉換成0~…的格式。
* 資料點未知：無論是period2 或是 test\_data都有出現未曾出現的資料點，一開始我有點困惑，因為一般理論的模型討論，我沒有看過討論新的點的case。想了一陣子，就把這些data當作feature值為0。
* 不熟悉工具、資料結構：一開始的資料結構有點不確定怎麼做比較好，透過參考別人怎麼實作adjacency matrix 以及資料處理的方法，有比較釐清了。
* 取負採樣的數量：有聽說本次實驗，若把全部的負data都採用，會有過多的負data，所以我這次的實驗，負採樣的數量有限制為與正data的數量一樣多。

1. **The conclusion and important point, that you’ve learn from this homework.**

資料的取樣很重要，過多的負data，會導致整個model經常預測為負的結果。對於不同的network，會有各自適當的feature，這次我取的feature是討論兩個點之間的共同鄰居關係(Jaccard and Preferencial attachment)，又因為本次的network比較稀疏(sparse)，所以預測的結果不是很好。

透過教授補充的資訊得知，日後若要處理imbalanced data issue，可以透過複製positive instances (即"oversample" positive instances)，也可以"undersample" negative instances。甚至還可以創造新的positive instances (而非單純複製)，其中一個經典的做法叫做 SMOTE (Synthetic Minority Oversampling Technique)，但由於時間關係，並未在此次作業中實作。

1. **The source code execution instructions, that including with** **programing language platform,** **library, and environment to execute your code.**

**Source code execution instructions please follow the readme.md in github.**

* 1. programing language platform : python
  2. library : networkx, numpy, pandas, matplotlib, sklearn.linear\_model LogisticRegression
  3. environment : jupyter notebook (any kernel with python3)