

HW2 - Community Detection

說明：

這次的作業會提供 317,080 個 node 和 1,049,866 個 undirected edge。同學需自行預測 test.csv 中的 node pair 是否在同一個 community 中，「是」則為 1，「否」則為 0。我們會根據預測的結果給予分數，評分的標準為 Categorization Accuracy。你可以使用任何的方法預測，可使用 library 實作這次的作業。作業要上傳到 Kaggle 評分，並繳交程式檔和說明文件到 Moodle。說明文件中請簡述演算法流程，並說明如何執行你的程式。

檔案說明：

- train.csv 是用來訓練的資料，'Node1'和'Node2'代表 edge 兩端的節點。
- test.csv 是同學需要回答的問題，'Id'欄位代表 index，'Node1'和'Node2'代表問題中的 node pair。同學需判斷每個 node pair 是否在同一個 community 中。
- sampleSubmission.csv 用以展示上傳到 Kaggle 的檔案的格式，'Id'從 0 開始依序排到 999，Category 就是預測結果(0 或 1)。

參考方法：

1. 根據 train.csv 中的 node pairs 建出圖 (igraph, networkx)。
2. 用 community detection algorithm 分出每個 community。
3. 回答 test.csv 中的每一個 node pair 是否在同一個 community 中。
4. 注意上傳的檔案只需要 node pair 的 id 和 prediction 結果。

Kaggle link: <https://www.kaggle.com/t/3da645f5feb8463cb89731f8d2ca0217>

HW2 - Community Detection

Description:

In this assignment we provide 317,080 nodes and 1,049,866 undirected edges. You need to determine whether each node pair in "test.csv" is in the same community, "Yes" is 1, "No" is 0. This assignment will be graded based on Categorization Accuracy. You can use any prediction methods, including library function. You need to upload your prediction to Kaggle for evaluation. Please also submit your program files and a report to Moodle. In your report, please briefly describe the algorithm you use and instructions on how to execute your program.

File descriptions:

- train.csv is the information used for training. 'Node1' and 'Node2' represent two ends of edge
- test.csv is the question you need to answer. The 'Id' field represents the index number, and 'Node1' and 'Node2' are the node pair in question. You need to determine whether each node pair is in the same community.
- sampleSubmission.csv shows you the file format needed to upload to Kaggle. 'Id' starts from 0 to 999. 'Category' is the prediction result (0 or 1).

Example Method:

1. Build the graph based on node pairs in train.csv (igraph, networkx).
2. Use a community detection algorithm to identify each community.
3. Answer whether each node pair in test.csv is in the same community.
4. The file you upload to Kaggle only need ids and prediction results of node pairs.

Kaggle link: <https://www.kaggle.com/t/3da645f5feb8463cb89731f8d2ca0217>