# Project 2 Social Network Mining Project Report

Group: 4-Hot-Vectors
Member: Lin Shuya, Yao Xinjie, Fan Yuwen, Chen Ziyi

**Final Result**

Node to vector with num_walks=15, walk_length=15, p=0.5, q=0.5. The validation auc score is 0.93468.

**Algorithms**

- **Network Embedding**
  - Both *Deep Walk* and *Node to Vector (Node2Vec)* are used as two different network embedding. They are tuned separately in order to get the best result from each of them.
- **Generation of false edges**
  - False edges are generated from the training network and validation network. The number of false edges is approximately equal to the number of true edges in the validation network, in order to generate a reliable performance result.
    - number of false edges =20738, number of true edges = 19268
    - number of false edges ≈ number of true edges
    - number of false edges + number of true edges = 40000

**Analysis of Hyper-parameters Tuned**

- **Numer of Walks (num_walks) and Walk Length**
  - num_walks: how many random walks are generated.
  - walk_length: the maximum length of each random walk.
  - p=0.5, q=0.5 are set to get the best combination of num_walks & walk_length.
- **Preference for Depth-First Search and Breadth-First Search (p and q)**
  - Basis: The unnormalized probability of moving to a same-distance node is 1.
  - Return parameter p: $\frac{1}{p}$ is the unnormalized probability of returning back.
  - In-out parameter q: $\frac{1}{q}$ is the unnormalized probability of moving further.
  - The best combination of num_walks & walk_length is set to tune p & q. If there are multiple best combination, the one with smallest num_walks & walk_length is used to save time.

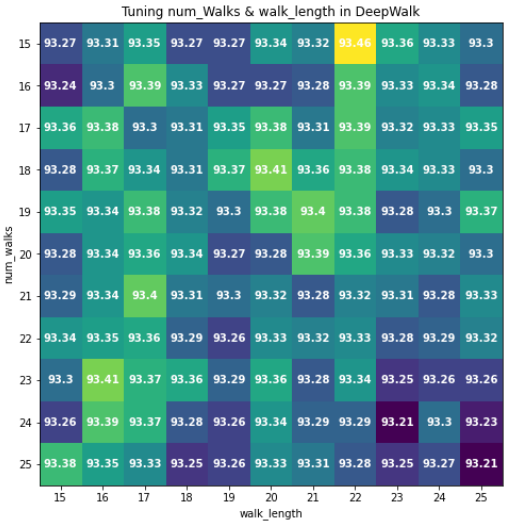
**Heatmaps**

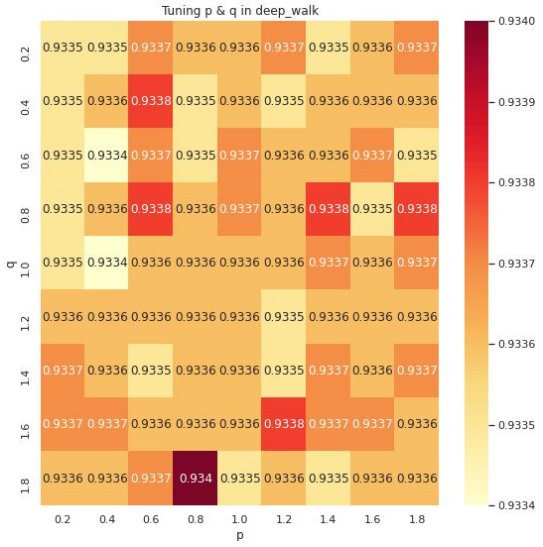
- Heatmaps are generated to show the trend of auc scores and narrow the range of choosing parameters. Since the relationship between parameters and scores are not absolute, the stable combinations around the ‘best scores’ may also be tried and used in the final result.

- **Deep Walk:**
  - **num_walks & walk_length:** After an initial broad grid search for num_walks and walk_length ranging from 10 to 100 (with a step of 5), we found that better results occur in the range from around 15 to 25. Hence, we narrowed down the grid search to be in this range and got multiple good results higher

than the 100% baseline. The best validation performance was 93.46 when num_walks is 15 and walk_length is 22, as shown in the following heat map.

- **p & q:** With num_walks as 15, walk_length as 22, we observe multiple optimum spots exist for p&q value selection. Especially when p,q is varying at the range of 0.5 or q is a bit larger than p, the results are highest among all other combinations in the range of 0.2-2 with step size as 0.2.




- **Node2Vec:**
  - **num_walks & walk_length**: Better results appear when both num_walks & walk_length are between 15-25 and when their difference is small.
  - **p & q**: The below heatmap uses walk length = 15, number of walk = 15, this combination is the stable pair of the best combinations from the above graph. Observation is that the better results have high q compared with low p. It means depth first search preference results in a higher accuracy.