

# RMBI4310/COMP4332 Project 2 Report

Group 35: Jamie, Lo Sze Yuen (20619883), NANDA, Nikhil (20491384), LAI, Wai Chun (20510702), LIN Chi Wing ()

## Abstract

This project aims to solve the link prediction problem on a subset of the Yelp dataset. The data consists of user\_id and friends which correspond to a directed graph, whose edges serve as the labels. Two different random walk based embedding algorithms are compared - 1) DeepWalk and 2) node2vec.

## 1) DeepWalk

The DeepWalk approach takes fixed-length, unbiased random walks starting from each node. In order to find the values for parameters that maximize the AUC - ROC score on the given dataset, grid search is performed on 3 parameters namely - node dimension (node\_dim), number of random walks (num\_walks) and length of each walk (walk\_length)

### node\_dim

Grid search has been performed on various values for node\_dim including 5, 10, 20, 30 and 40. The Figure 1 heatmap visualizes the validation AUC - ROC score corresponding to various node\_dim values. It can be seen that node\_dim value of 10 results in the highest AUC - ROC score of 92.67%.

### num\_walks & walk\_length

Grid search has also been performed on various values for num\_walks including 5, 10, 20 & 40 and for walk\_length including 10, 20 & 40. The Figure 2 heatmap visualizes the validation AUC - ROC score corresponding to various node\_dim and walk\_length values. It can be seen that num\_walks value of 10 and walk\_length value of 20 results in the highest AUC - ROC score of 93.18%.

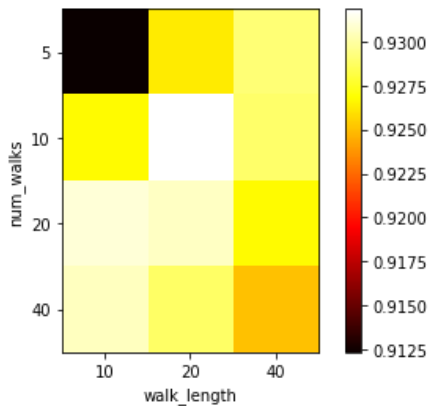


Figure 1: Grid Search on node\_dim

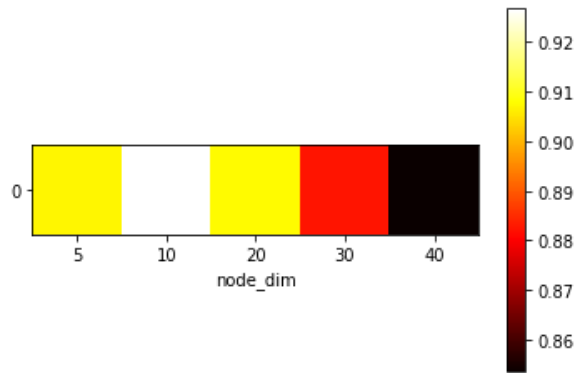


Figure 2: Grid Search on walk\_length and num\_walks

## Final DeepWalk Model

node\_dim = 10, num\_walks = 10 and walk\_length = 20 achieves validation accuracy of 93.17%

## 2) node2vec

The node2vec approach takes flexible, biased random walks that can trade off between local and global views of the directed graph which is controlled by two parameters -  $p$  and  $q$ . In order to find the values for parameters that maximize the AUC - ROC score on the given dataset, grid search is performed on 5 parameters namely - node dimension (node\_dim), number of random walks (num\_walks), length of each walk (walk\_length), return parameter ( $p$ ) and In-out parameter ( $q$ )

### node\_dim

Grid search has been performed on various values for node\_dim including 5, 10, 20, 30 and 40. The Figure 3 heatmap visualizes the validation AUC - ROC score corresponding to various node\_dim values. It can be seen that node\_dim value of 10 results in the highest AUC - ROC score of 92.57%.

### num\_walks & walk\_length

Grid search has also been performed on various values for num\_walks including 5, 10, 20 & 40 and for walk\_length including 10, 20 & 40. The Figure 4 heatmap visualizes the validation AUC - ROC score corresponding to various node\_dim and walk\_length values. It can be seen that num\_walks value of 10 and walk\_length value of 20 results in the highest AUC - ROC score of 93.06%.

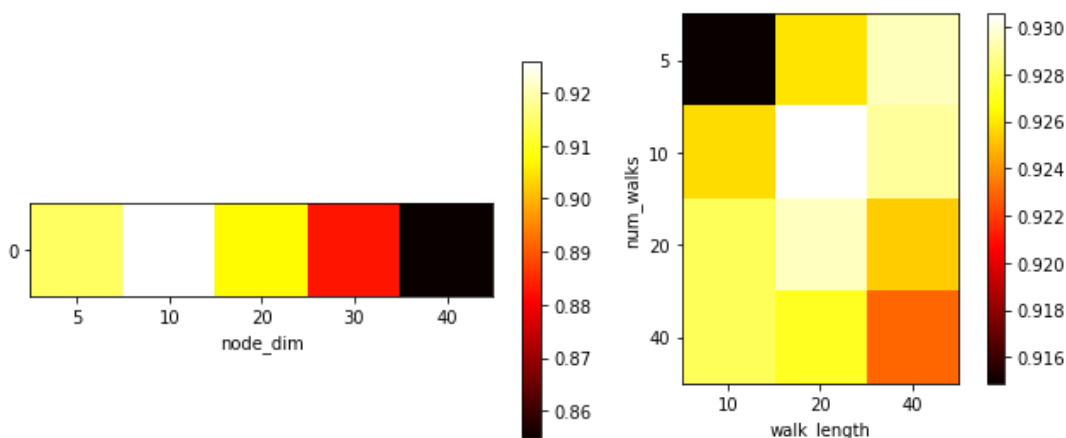


Figure 3: Grid Search on node\_dim

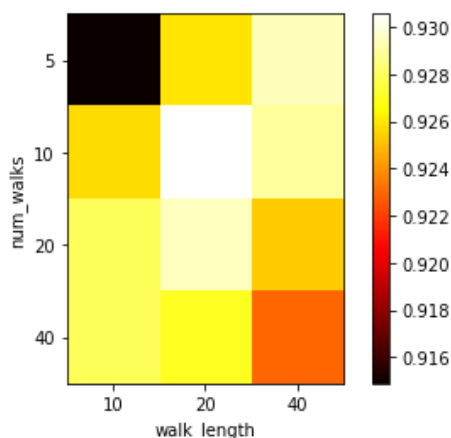


Figure 4: Grid Search on num\_walks and walk\_length

### p and q

Lastly, grid search has been performed on various values for  $p$  including 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5 and for  $q$  including 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8. From the Figure 5 heatmap, it can be seen that  $p$  value of 0.4 and  $q$  value of 1.8 results in the highest AUC - ROC score of 92.95%.

### Final node2vec Model

node\_dim = 10, num\_walks = 10, walk\_length = 20,  $p$  = 0.4 and  $q$  = 2.0 achieves validation accuracy of 93.18%

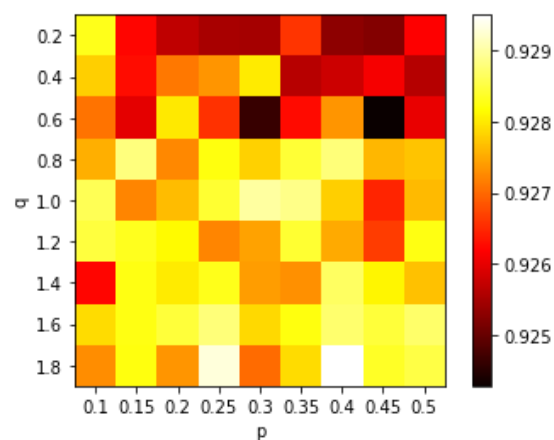


Figure 5: Grid Search on  $p$  and  $q$