COMP4332 Project 2 - Social Network Mining (Group 14)

Members: CHEN, Jiawei (20763842), Leung King Suen (20770625), Kwong Ka Lok (20772439)

Introduction

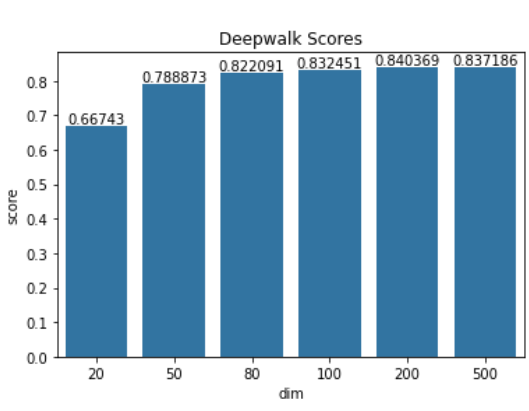
For this project, we followed the given general pipeline and adopted a 3-stage parameter tuning strategy for both deepwalk and node2vec model to achieve a higher AUC-ROC score on the validation set.

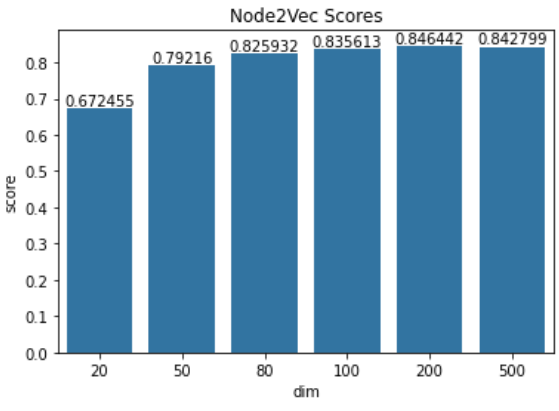
Stage-1 Tuning on dimension of node embedding

Parameter Setting

	Deepwalk	Node2vec
node_dim	[20, 50, 80, 100, 200, 500]	
num_walks	10	
walk_length	10	
p	0.5	
q	0.5	

Results





The AUC-ROC score of deepwalk and node2vec model reached the highest score of **84.03%** and **84.64%** respectively when **node_dim=200**. With a higher node_dim of 500, it shows no

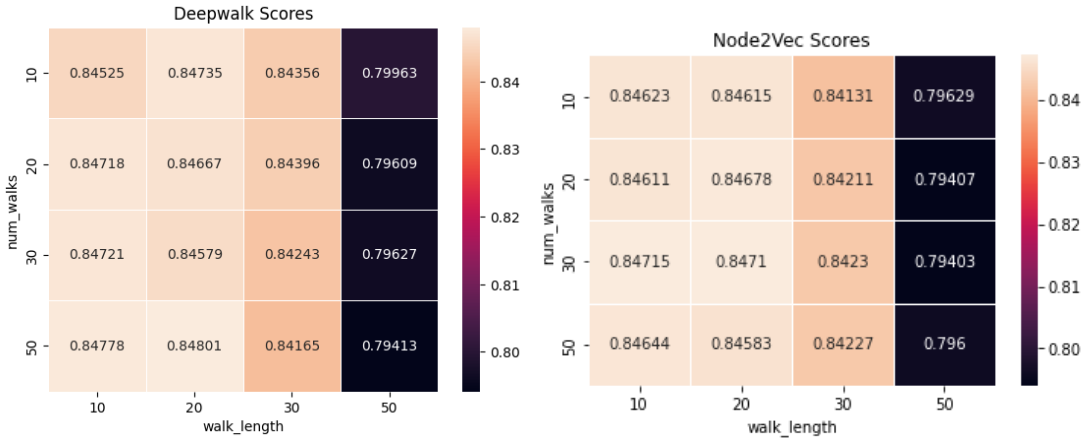
improvement in performance and doubled the training time.

Stage-2 Tuning on number of walks, walk length

Parameter Setting

	Deepwalk	Node2vec
node_dim	200	
num_walks	[10, 20, 30, 50]	
walk_length	[10, 20, 30, 50]	
p	0.5	
q	0.5	

Results



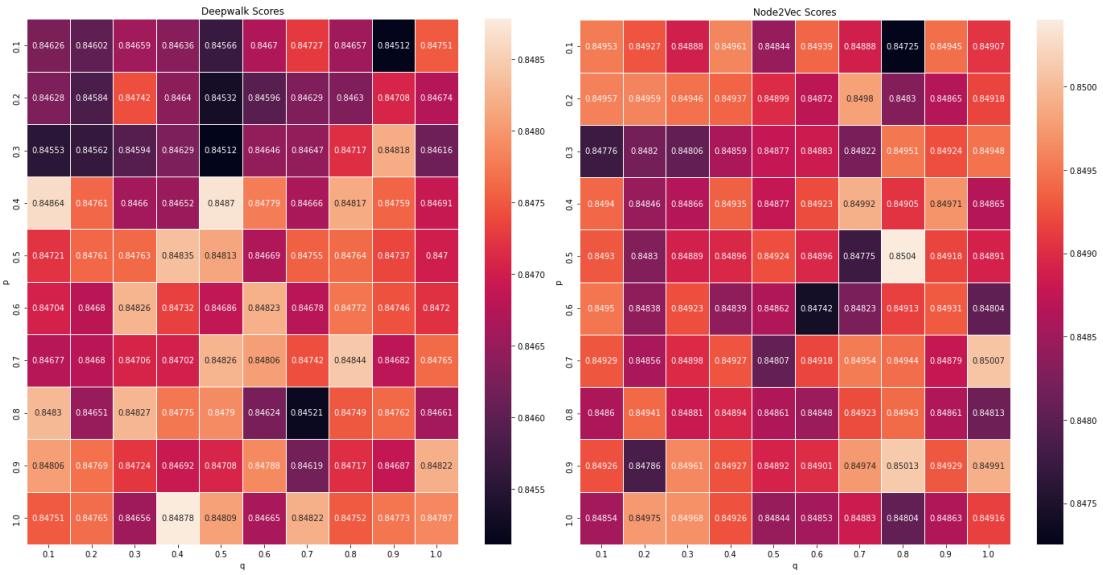
The above two heatmaps showed comparable results for a walk length of 10 and 20. For the different number of walks performed, the results were rather similar. The deepwalk model reached the highest score of **84.80%** with **num_walks=50** and **walk_length=20**; while the node2vec model reached the highest score of **84.71%** with **num_walks=30** and **walk_length=10**. Compared to the previous stage, the models' performances increased by **0.77%** and **0.07%** respectively. We noticed a small decrease in scores on 30 walk length and noticeable decrease with 50 walk length. This hinted that the network is less complex and suited for 10-20 walk lengths. A longer walk length is more favored to networks with more complex patterns however overfitted on simpler networks. The model's training time increased linearly with num_walks and walk_length.

Stage-3 Tuning on p, q

Parameter Setting

	Deepwalk	Node2vec
node_dim	200	
num_walks	50	30
walk_length	20	10
p	[0.1 – 1.0]	
q	[0.1 – 1.0]	

Results



The deepwalk model reached the highest score of **84.878%** with **p=1.0** and **q=0.4** while the node2vec model reached the highest score of **85.04%** with **p=0.5** and **q=0.8**. Compared to the previous stage, the models' performances increased by **0.078%** and **0.27%** respectively. The patterns are mostly random. A more pronounced black-region was only at low p value (0.1-0.3) in the deepwalk model, this indicated that for deepwalk model on this network, a bfs-like random walk performed a bit worse. However, it is worth noting that the maximum difference in performance is only 0.4%.

Model Performance after Parameter tuning

	Deepwalk	Node2vec
node_dim	200	200
num_walks	50	30
walk_length	20	10
p	1.0	0.5
q	0.4	0.8
AUC-ROC score (validation)	84.878%	85.04%

As the final node2vec model has a higher score on the validation set, we choose it as our final model to generate the predicted scores on the test set. We concluded that the most prominent parameter to the model score is node_dim parameter. As for other parameters, they did not affect us much. Apart from walk length, where a larger value may lead to overfitting.