COMP 4332 / RMBI 4310
Big Data Mining (Spring 2021)

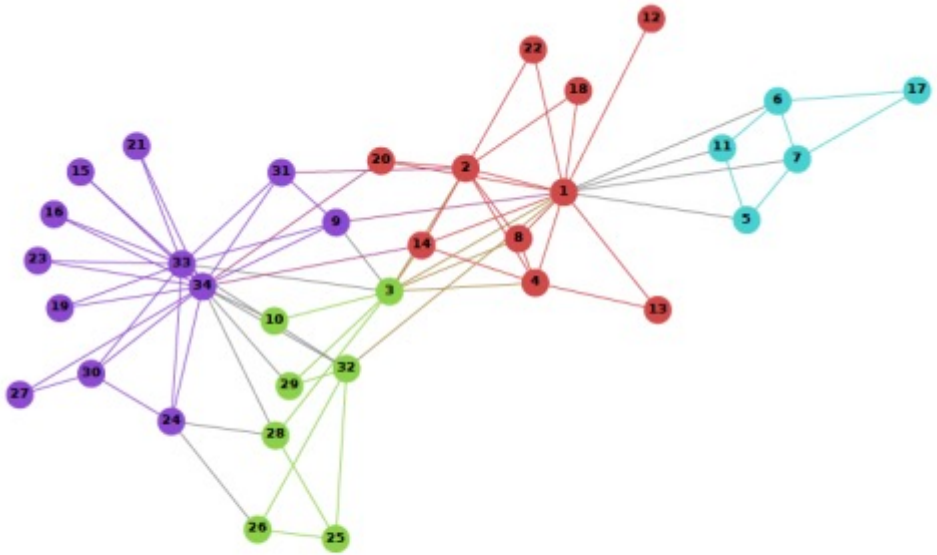
Project 2 Social Network Mining

TA: Qi HU(qhuaf@connect.ust.hk)

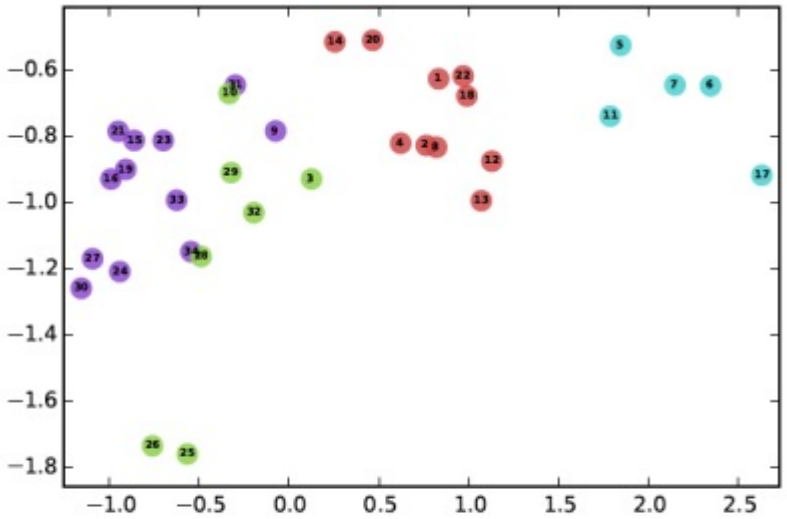
Social Network

https://www.datacenterknowledge.com/archives/2012/02/02/facebooks-1-billion-data-center-network

Network Representations



Input

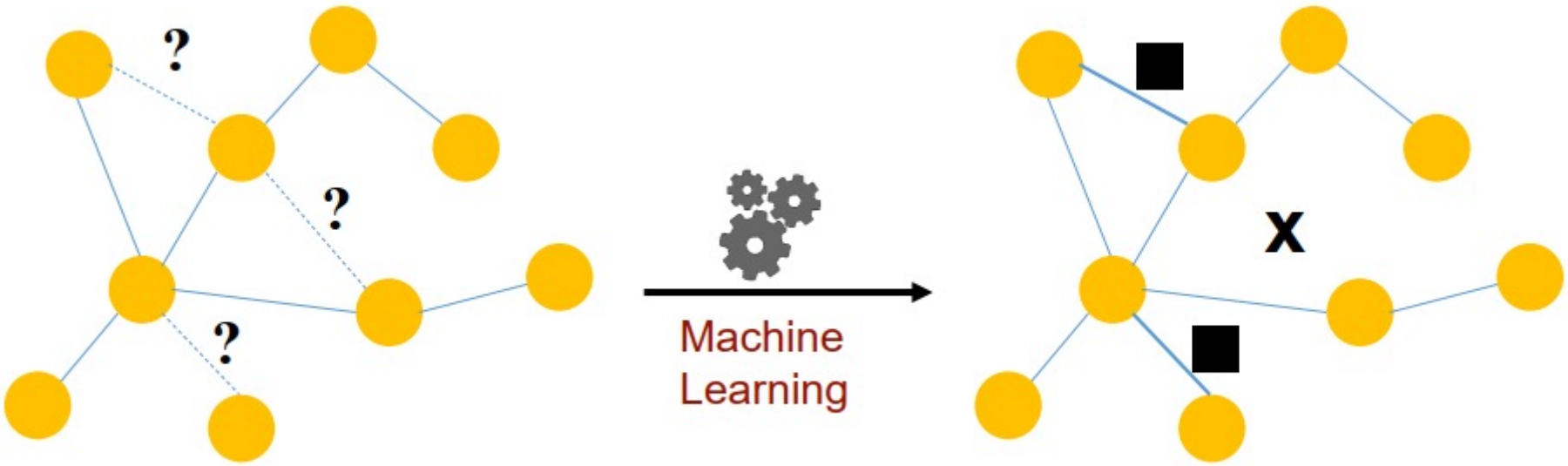


Output

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In *KDD*, pp. 701-710. 2014.

Link Prediction

- Predict the relation between nodes with their similarity and calculate the AUC-ROC score.



Pipeline

- Data loader
- Random walk generator (first-order, second-order, …)
- Embedding algorithm (DeepWalk, node2vec, …)
- Scorer

Dataset

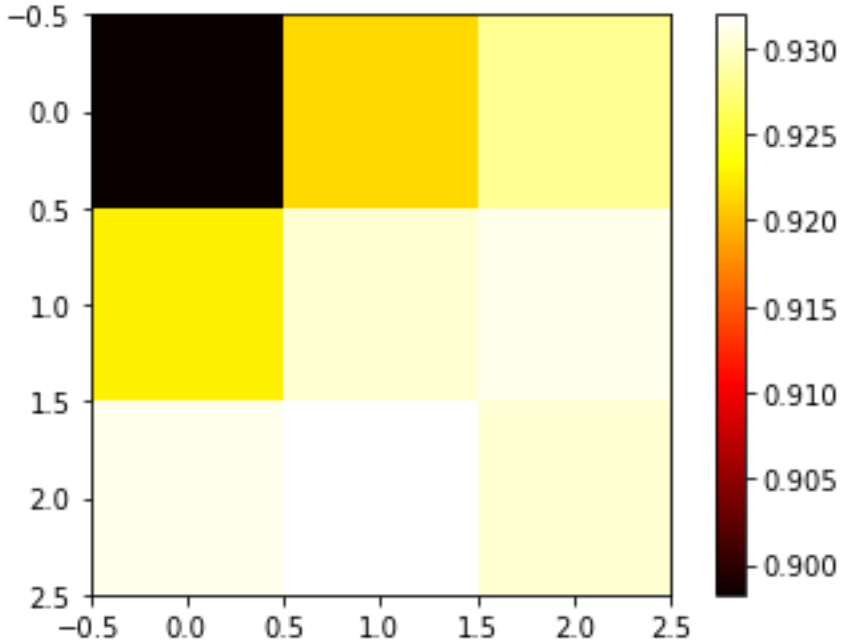
- training data: 8328 nodes, 100000 edges
- validation data: 5440 nodes, 19268 edges
- test data: 8509 nodes, 40000 edges (19267 positive edges)
- score: [0, 1]
- given features: `user_id, friends`

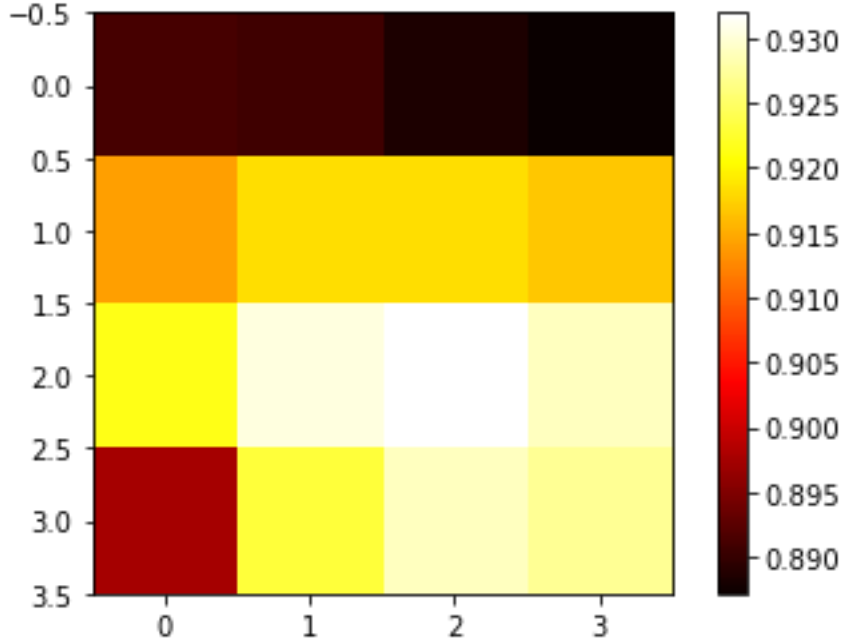
user_id	friends
--UOvCH5qEgdNQ8lzR8QYQ	[DBHCFW3mSmmOEpONHVu1rQ, QPJJohtGqkMkaN0Gt3TRI...
-05T0q5BxB9g0RCKiGYoyQ	[V7uS5US4oTf-S9u36HJQCQ]
-0HhZbPBlB1YZx3BhAfaEA	[uG35h72BAMutvXAWdRpqCQ, Sv48jgljDP-CRfXmU8uSg...
-1ZMRA0N01rqZL0TWk3fgA	[no2kFt4TEEzZDVaM8haSDA]
-267Yx8RmdP6io2-ql4UcQ	[E3pXvQwKsPBQGQ7RkLrN3g, deL6e_z9xqZTIODKqnvRX...
...	...

src	dst	score
Nu5188fyBvHZHvgEgZT2bQ	lKSmm5MzHF8cMhMolKaIOw	0
rCWrxuRC8_pfagpchtHp6A	xQLy_wpqrR3etSXt61OIlg	0
maK3UBQczh33NuDjBYeHrA	K0sapHOlhlGNjx3GBeSf5A	0
siXOnFrtV0a_YjOJr-X2Mg	7hAhYoMPjHnxKCz6MQ95Bg	0
_4iFWWuZ6_RrzyXZrMq3Mw	rTK_sTPgBjJXkdKi2G3X-w	0
...	...	...

Analysis

- Heatmap (https://stackoverflow.com/questions/33282368/plotting-a-2d-heatmap-with-matplotlib)





Evaluation

- AUC-ROC score on **test data**

Submission

- Predictions on **test data** (please make sure your pred.csv's format is same as test.csv: src/dst/score)
- Report (1~2 pages)
- Code (Frameworks and even programming languages are not restricted.)
- DDL: May 3, 2021
- Submission: Each **team leader** is required to submit the groupNo.zip file that contains pred.csv, the report, and your team's code on canvas.
- we will check your report with your code and the accuracy.

Grading Rule

Grade	Model (80%)	Report (20%)	Baseline
60%		submission	
80%	an easy baseline that most students can outperform	detailed explanation	40000 edges (20732 negative): 92.30%
90%	a competitive baseline that about half students can surpass	detailed explanation and analysis	40000 edges (20732 negative): 92.80%
100%	a very competitive baseline	excellent visualization and analysis	40000 edges (20732 negative): 93.10%

Thank You