COMP 4332 / RMBI 4310
Big Data Mining (Spring 2024)

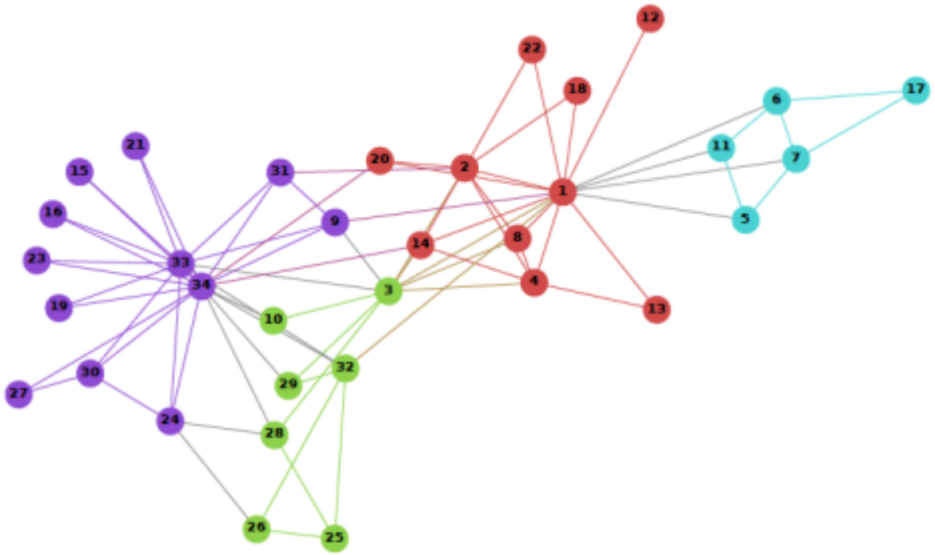
Project 2: Social Network Mining

TA: Sehyun Choi (schoiaj@connect.ust.hk)

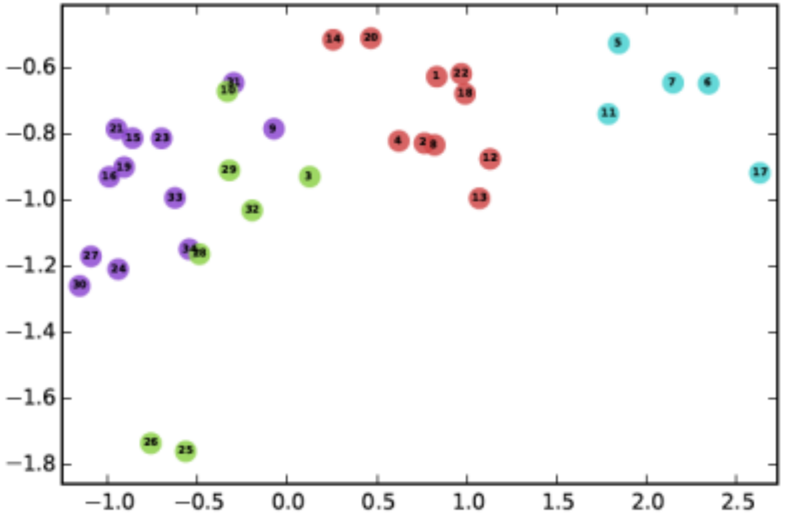
Social Network

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

Network Representations



Input

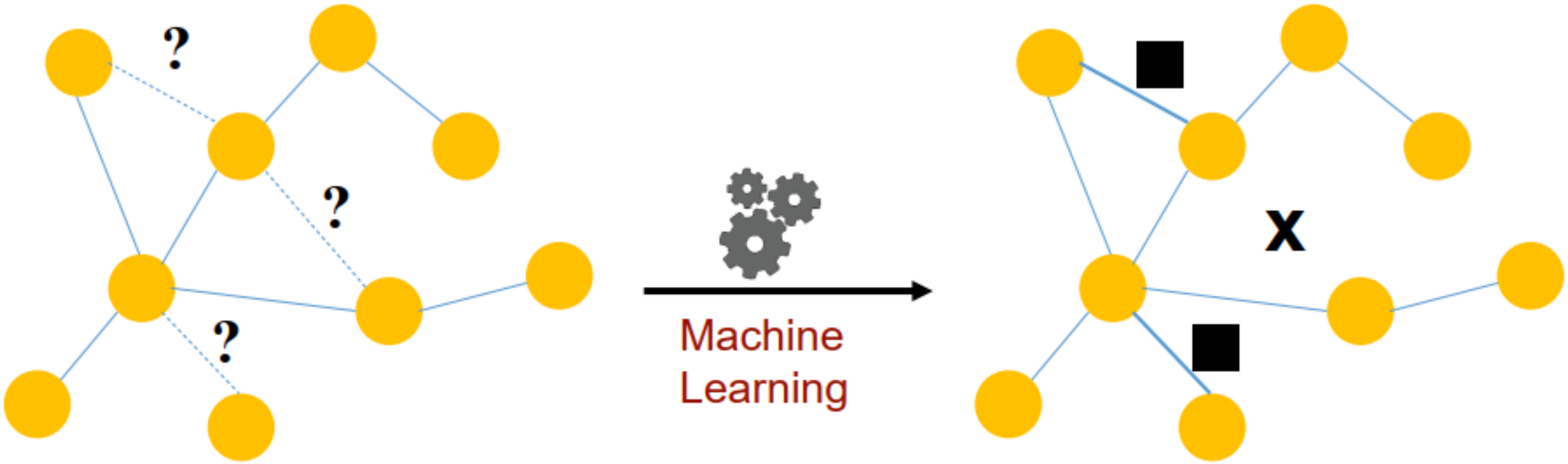


Output

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

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

Dataset

- Training data: 72795 nodes, 100000 edges
- Validation data: 17093 nodes, 20000 edges
- Test data: 32166 nodes, 40000 edges
- Score: [0, 1]
- Given features: `user_id, friends`

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: April 30, 2024
- Submission: Each **team leader** is required to submit the groupName.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 AUC scores. You will be graded based on your testing set performance and your report.

Grading Rule

Grade	Model (80%)	Report (20%)	Baseline (on test set)
60%		submission	
80%	an easy baseline that most students can outperform	detailed explanation	40000 edges (20000 negative): 65.00%
90%	a competitive baseline that about half students can surpass	detailed explanation and analysis	40000 edges (20000 negative): 67.00%
100%	a very competitive baseline	excellent visualization and analysis	40000 edges (20000 negative): 70.00%

Thank You