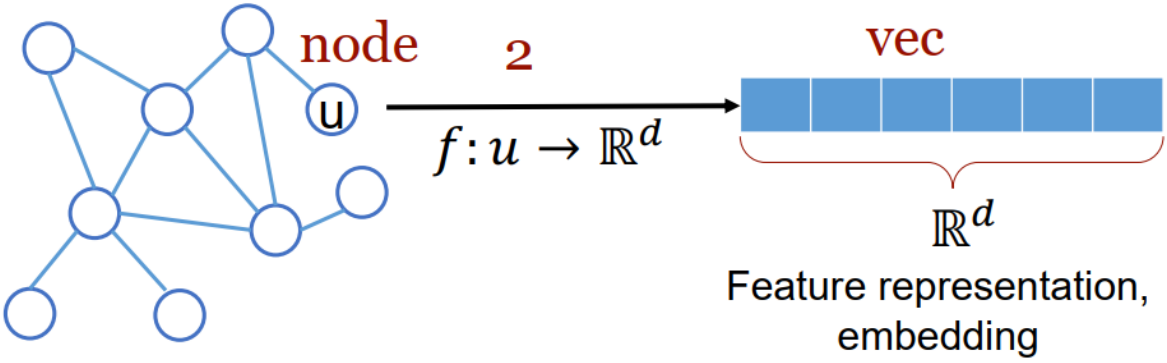
COMP 4332 / RMBI 4310
Big Data Mining and Management
Advanced Data Mining for Risk Management and Business Intelligence (2025 Spring)

Tutorial 5: Network Representation and Random Walk Based Network Embeddings

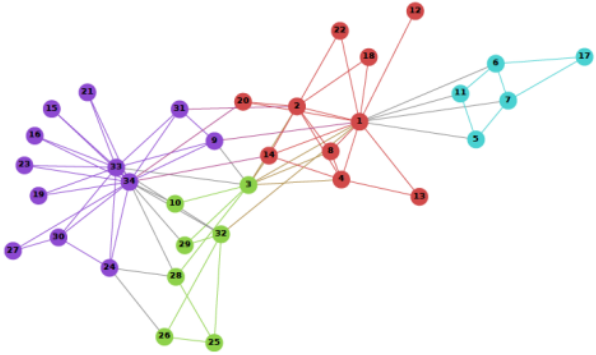
TA: Qing ZONG (qzong@connect.ust.hk)

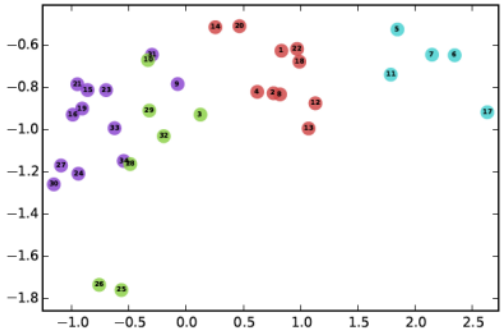
Network Representations



Goal: efficient task-independent feature learning for machine learning in networks.

Network Representations





Input **Output**

Intuition: find embedding of nodes to d-dimensions so that “similar” nodes in the graph have embeddings that are close together.

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

Network Representations

- Input Examples (user.csv)

user_id	friends
ZNZ7dxIsbHCbcbqTKQlUpg	['iDlkZO2iILS8Jwfdy7DP9A']
MmeVq6bWhogNpVPuEemrBA	['NwaJgVXNKZsQMVsgs6SOnA', '40m0541Z_KexYkdvvqHJIQ']
7aiIlfpcQZG7ea10R5i4RQ	['Q4Qfu-3vYtL1LRm2X1b0Gg', ‘PeLGa5vUR8_mcsn-fn42Jg']

- Output
  - d-dimensions vector

Applications

- **Link Prediction**
- Node Classification
- Recommendation
- Visualization

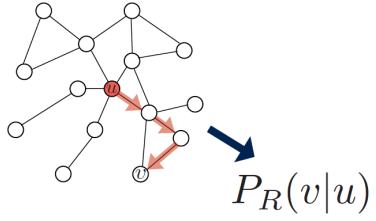
Work in the History

- ISOMap [Tenenbaum et al., Science'00]
- LLE [Roweis and Saul, Science'00]
- Laplacian EigenMap [Belkin et al., NIPS'01]
- (t)-SNE [Maaten and Hinton, JMLR'08]
- Deepwalk [Perozzi et al., KDD'14]
- LINE [Tang et al., WWW'15]
- Node2vec [Grover and Leskovec, KDD'16]

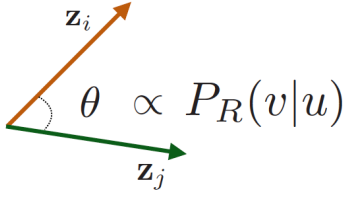
General Idea

1. Estimate the probability of visiting node *v* on a random walk starting from node *u* using some random walk strategy *R*.
2. Optimize embeddings to encode these random walk statistics.

$$\boldsymbol{Z}_u \boldsymbol{Z}_v \approx \mathrm{P\{u, v\ co\_occur\ on\ a\ random\ walk\}}$$



Convert graph into list of nodes



Train embedding with Word2Vec

Why using Random Walks

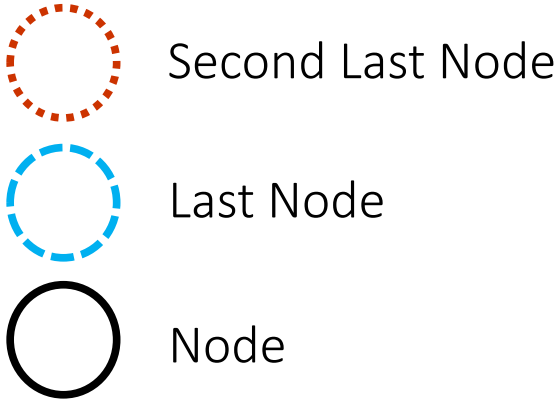
- **Expressivity**: Flexible stochastic definition of node similarity that incorporates both local and higher-order neighborhood information.

- **Efficiency**: Do not need to consider all node pairs when training; only need to consider pairs that co-occur on random walks.

First-order Random Walk

A random walker moves to next node based on the last node.

$$\pi_{c_i|c_{i-1},c_{i-2}} = \pi_{c_i|c_{i-1}} = \begin{cases} 1, & \text{if } (c_{i-1}, c_i) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases}$$

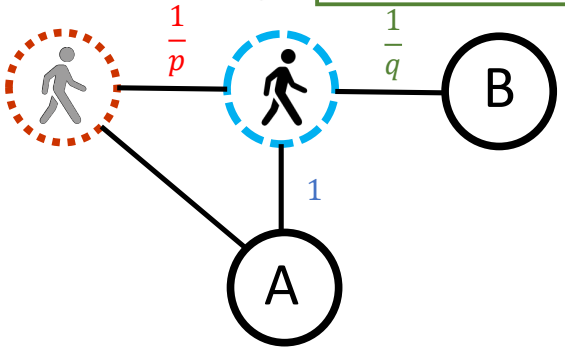


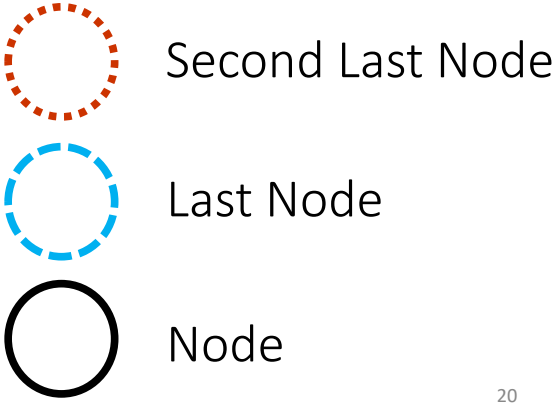
Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In *KDD*, pp. 701-710. 2014.

Second-order Random Walk

A random walker moves to next node based on the last two nodes.

$$\pi_{c_i|c_{i-1},c_{i-2}} = \begin{cases} \frac{1}{p}, & \text{if } d_{c_{i-2}c_i} = 0 \\ 1, & \text{if } d_{c_{i-2}c_i} = 1 \\ \frac{1}{q}, & \text{if } d_{c_{i-2}c_i} = 2 \end{cases}$$





Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." In *KDD*, pp. 855-864. 2016.

Pipeline

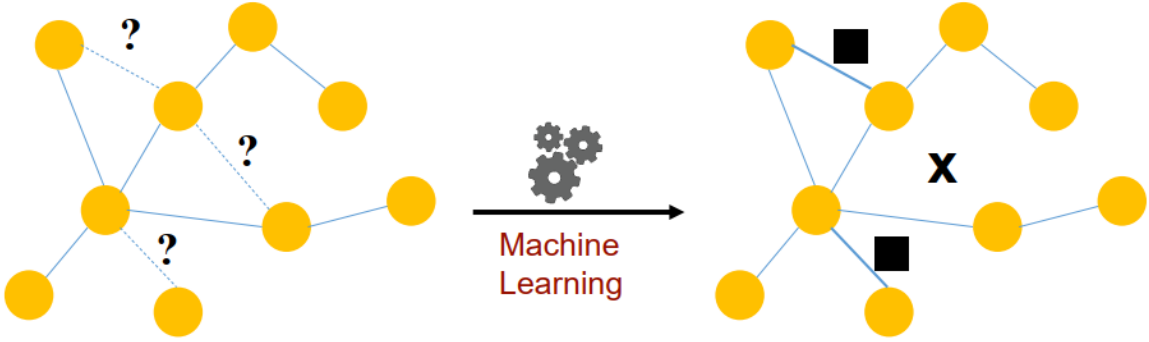
- Data loader
- Random walk generator
- Embedding algorithm
- Scorer

Implementation of the Whole Pipeline

See the jupyter-notebook.

Link Prediction

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

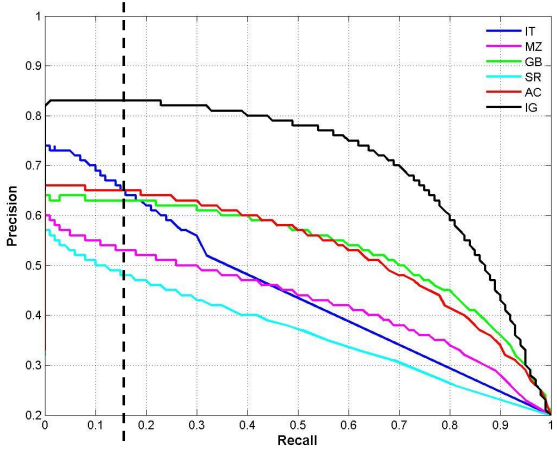


AUC-ROC Score

- ROC and AUC are terms commonly used in the context of machine learning and statistics, particularly for evaluating the performance of **classification** models.
- **ROC** stands for **Receiver Operating Characteristic**. It's a graphical plot that illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. In simpler terms, it shows how well a model can distinguish between two classes
  - **True Positive Rate (TPR)** = TP / (TP + FN), TP is true positives and FN is false negatives.
  - **False Positive Rate (FPR)** = FP / (FP + TN), FP is false positives and TN is true negatives.
- The ROC curve is created by plotting TPR (y-axis) against FPR (x-axis) at different thresholds.

AUC-ROC Score

- Area Under Curve (AUC) – ROC (Receiver Operating Characteristic Curve)



Under each recall level, we prefer a higher precision

AUC-ROC Score

- **AUC** stands for **Area Under the Curve**. Specifically, it refers to the area under the ROC curve. AUC provides a single scalar value summarizing the overall performance of a model across all possible thresholds.
  - AUC ranges from 0 to 1, A higher AUC indicates a better-performing model:
    - **1.0**: Perfect model (perfectly separates classes).
    - **0.5**: Model with no discriminative power (random guessing).
    - **< 0.5**: Model performing worse than random guessing (rare, indicates a flipped prediction).

Implementation of the Link Prediction Task

See the Jupyter Notebook.

Thank You