Graph Two-sample Testing with Node Embeddings

Qiucheng Wu[1] **Yuze Lou[1]** **Shucheng Zhong[1]** **Jiaxin Wang[1]**

[1] *Computer Science and Engineering*

EECS 576, F19
Advanced Data Mining

Problem Definition

- How can we apply node embedding methods to improve graph two-sample testing, i.e., determining if two populations of graphs are from the same distribution/random model?
- Evaluate various combinations of node embedding methods with hypothesis testing methods.

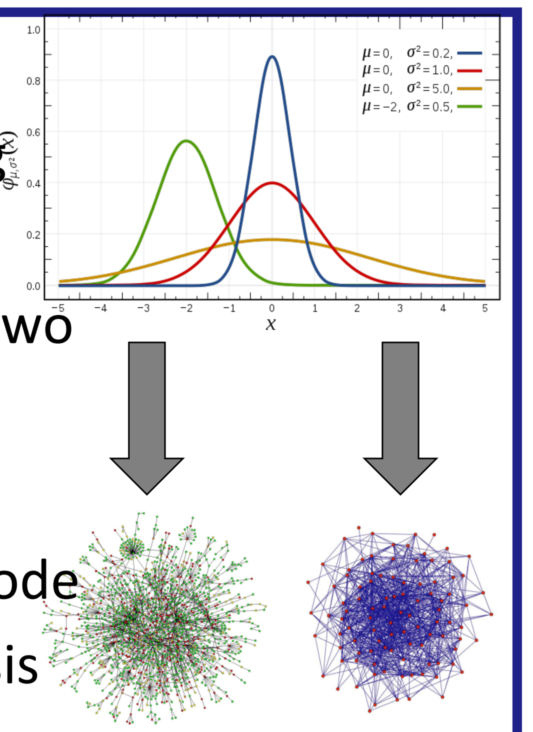


Image:https://en.wikipedia.org/wiki/Normal_distribution

Motivation

- Current work on graph two-sample testing focuses on theoretical approaches and tests on simple features.
- Node embedding is useful in many graph mining problems, so it may also be helpful to represent nodes by vectors in graph two-sample testing.

Datasets

- **ER**: Generated by Erdős–Rényi model.
  |N| = 500, |E| = 6318.
- **SBM**: Generated by stochastic block model.
  |N| = 500, |E| = 44,663.
- **Kronecker**: Generated by stochastic kronecker model.
  |N| = 512, |E| = 9838.
- **Arxiv GR-QC**: Collaboration network from e-print arXiv.
  |N| = 5242, |E| = 14496.
- **Arxiv Astro-ph:** Collaboration network from e-print arXiv.
  |N| = 18772, |E| = 198110.

Cornell University
arXiv.org

References

[1] Béla Bollobás, Svante Janson, and Oliver Riordan. 2007. The phase transition in inhomogeneous random graphs. Random Structures and Algorithms 31, 1 (2007), 3–122. https://doi.org/10.1002/rsa.20168
[2] Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. 2017. struc2vec. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17 (2017). 3097983.3098061

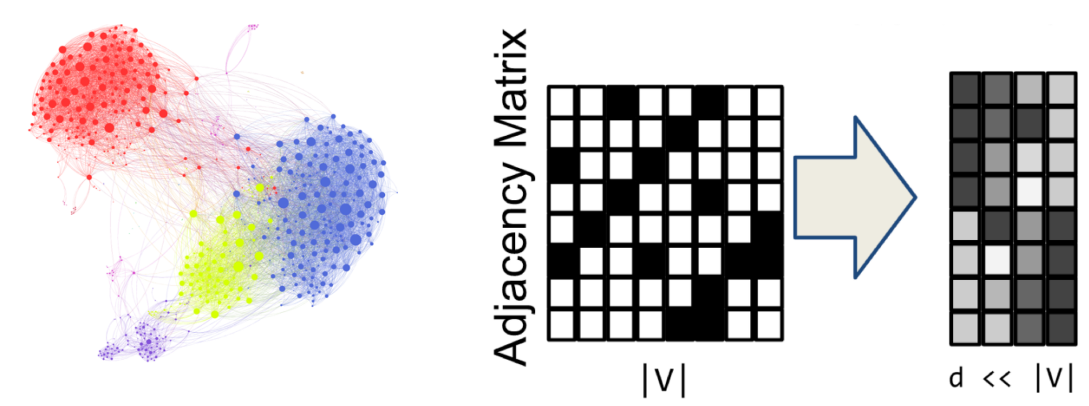
Use node embedding methods to improve graph two-sample testing

Our Approach

Step 1: Generate vector representation for nodes

Convert nodes to vectors with 2 and 128 dimensions. Node embedding methods include:

- node2vec
- struc2vec
- xNetMF
- GraphWave

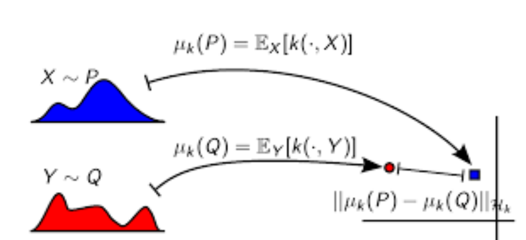


Step 2: Apply hypothesis testing methods to vectors

Use vector embeddings as input of test methods, such as

- Maximum Mean Discrepancy

$$MMD[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \lim (\mathbf{E}_x[f(y)] - \mathbf{E}_y[f(y)]$$

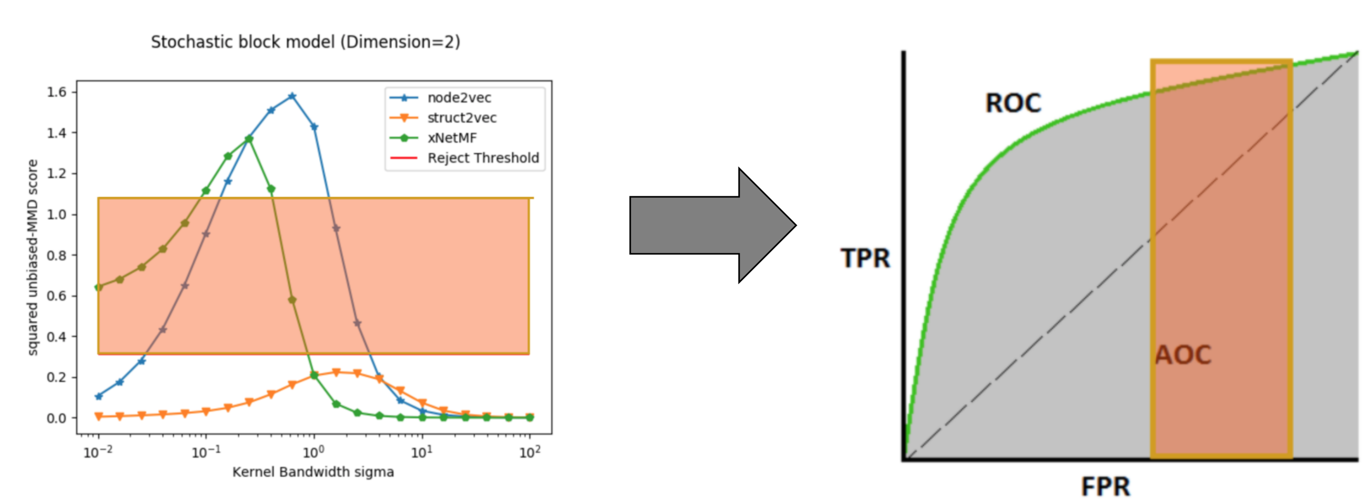


- Adjacency Spectral Embedding

$$T_{ASE} = \min \{ \|X_G - X_H W\|_F : W \in \mathbb{R}^{r \times r}, WW^T = I \}$$

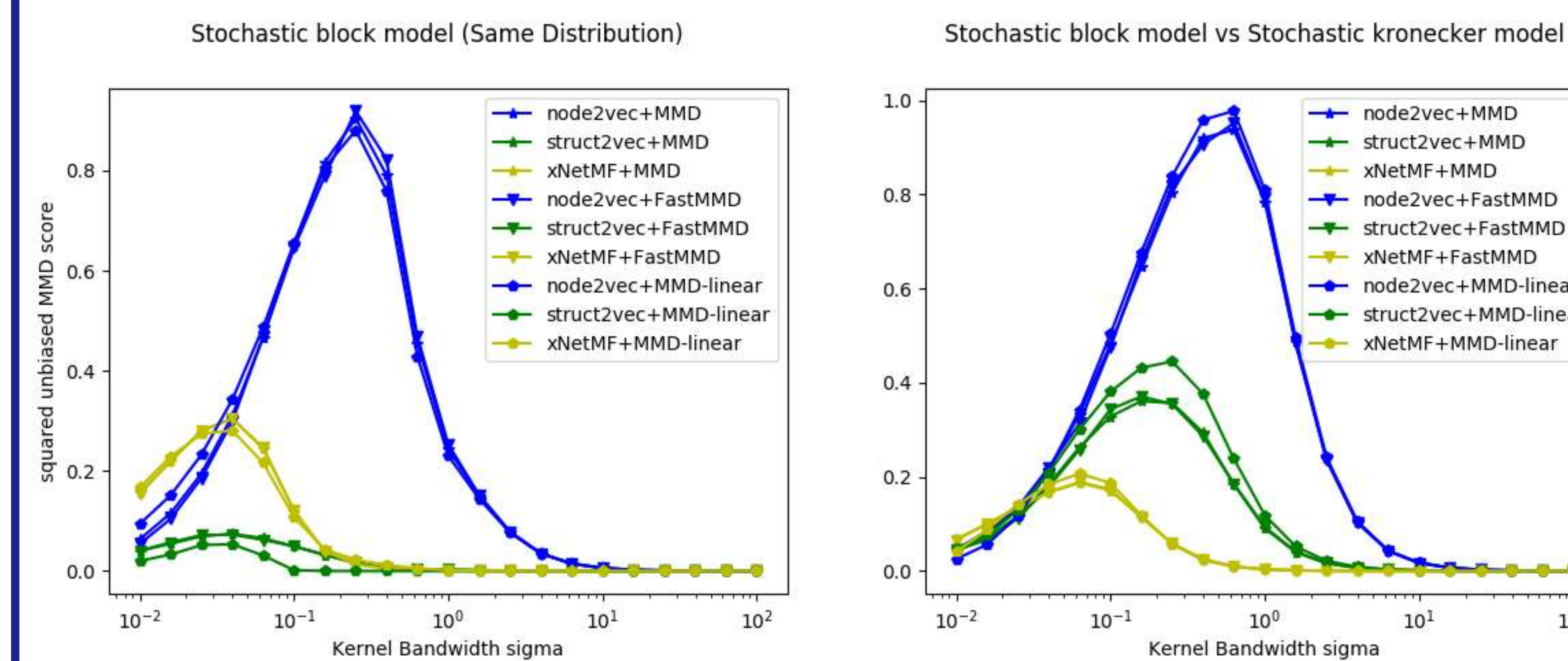
Step 3: Find the threshold of each testing method

Use AUC-ROC curve to try different thresholds and display the performance of our method.



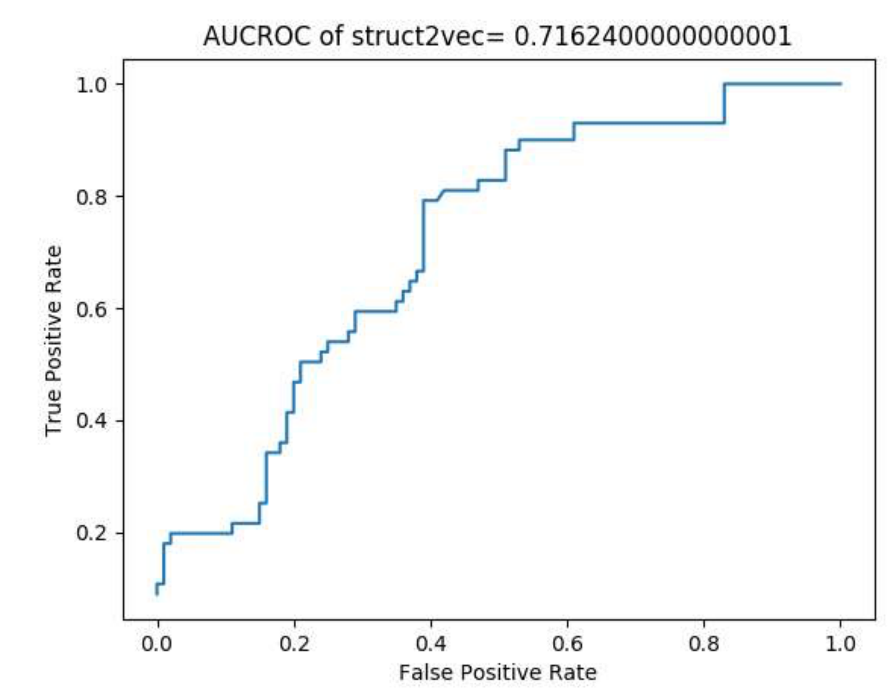
Experimental Results

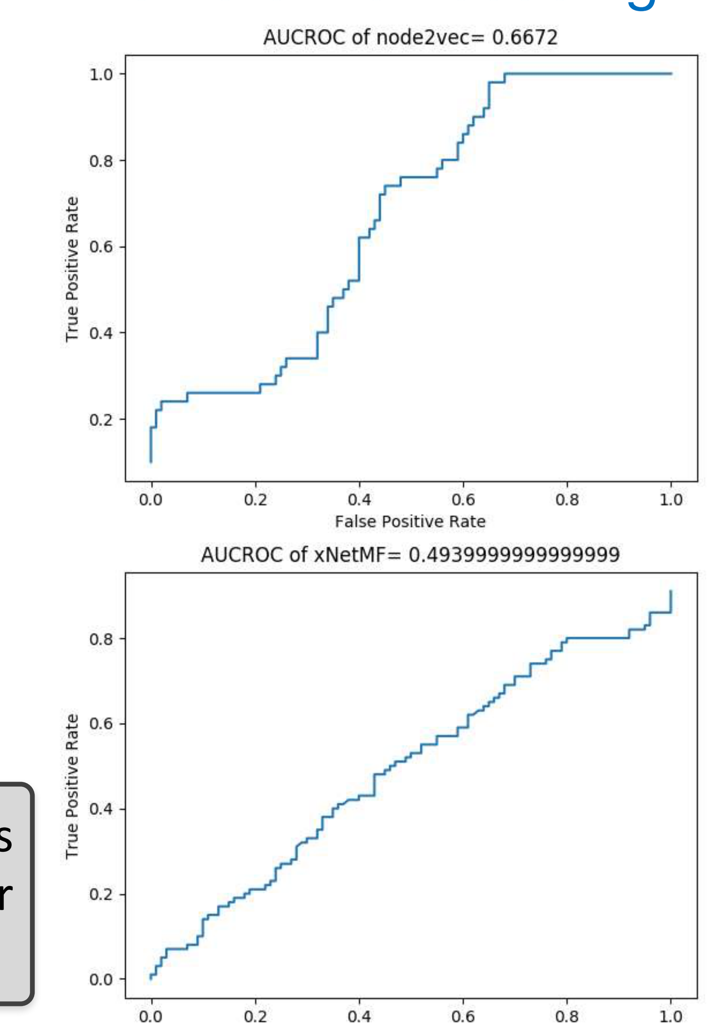
MMD score with different kernel bandwidth



Observation: We prefer a small MMD score for graphs from the same distribution (left figure) and a large score for graphs from different models (right figure).
struc2vec outperforms other two embedding methods.

AUC-ROC curve of unbiased MMD with different embedding methods





Observation: With unbiased MMD, in terms of AUROC, struc2vec also outperforms other two embeddings.

Conclusions

- struc2vec+MMD provides the best performance over other embedding methods in low dimension.
- Structural node embedding methods may not fit the two sample test since it is hard to interpret the distances between node embeddings
- Some heuristic methods may help the testing like using principal component analysis to reduce dimension in hypothesis test.