Graph Attention is Not Always Beneficial:

A Theoretical Analysis of Graph Attention Mechanisms via Contextual Stochastic Block Models (CSBMs)

Zhongtian Ma, Qiaosheng Zhang*, Bocheng Zhou, Yexin Zhang, Shuyue Hu, Zhen Wang*

Contact US! mazhongtian@mail.nwpu.edu.cn, w-zhen@nwpu.edu.cn, zhangqiaosheng@pjlab.org.cn

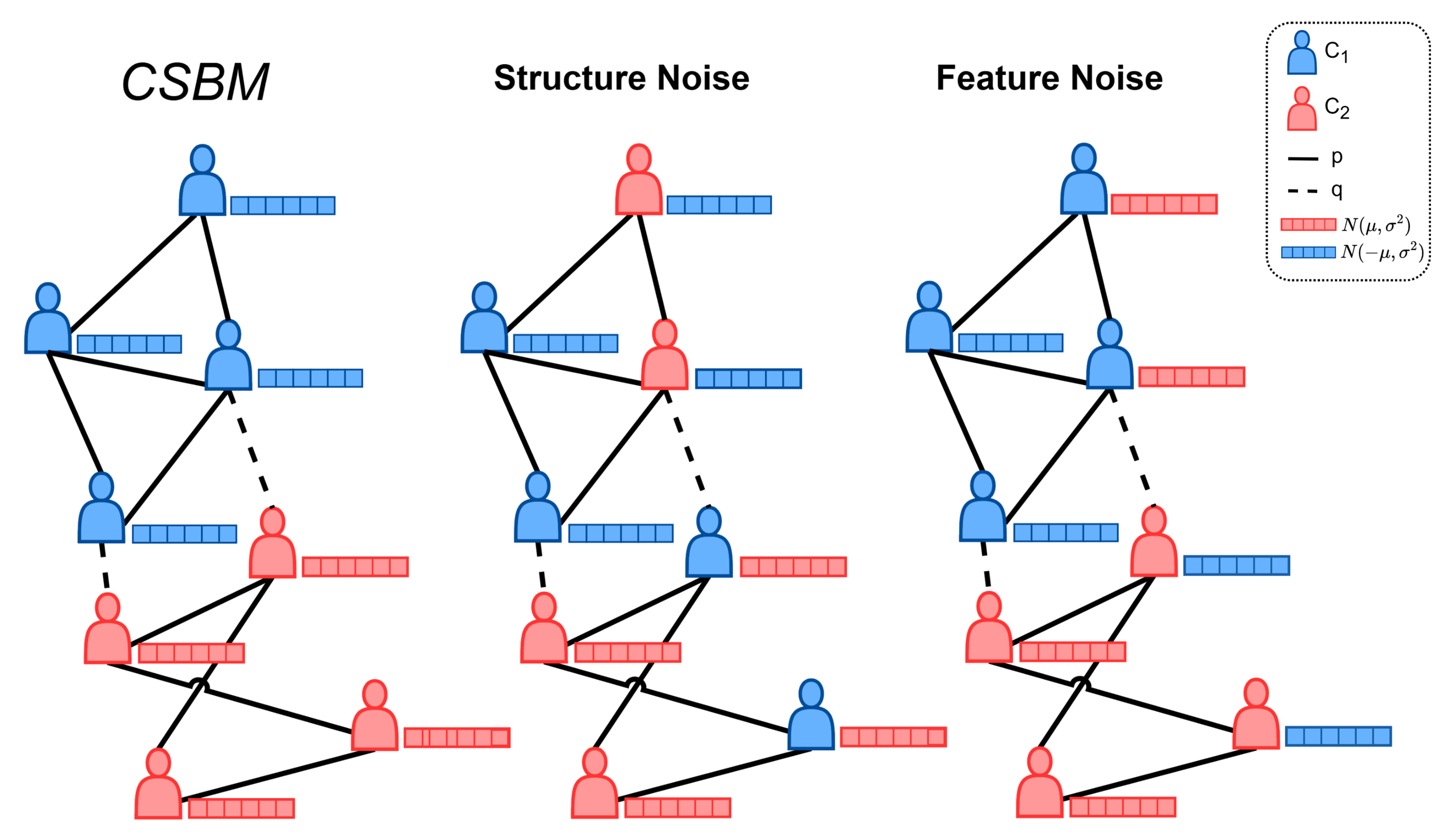
OpenReview WeChat

Motivation

- Despite the growing popularity of graph attention mechanisms (GAT), their **theoretical understanding** remains limited.
- Understand when and why **graph attention mechanism** works.

✓ Why CSBM?

- CSBM combines SBM and GMM to generate realistic graph structures and node features, ideal for both empirical and theoretical studies.
- In CSBM, nodes are split into several communities. **Intra-community** edges appear with probability p, **inter-community** edges with q; node features in each community are drawn from **a distinct Gaussian distribution**.



✓ Two types of noises

- We define two types of noise: **feature noise** and **structure noise**, as shown above.
- In CSBMs: $\mathcal{F}_{noise} = \frac{p+q}{p-q}$, $\mathcal{S}_{noise} = SNR^{-1} = \frac{\sigma}{\mu}$.
- We study node classification task with **perfect node classification**(i.e. **exact recovery**) as the metric, and show that feature and structure noise **are key to** the effectiveness of graph attention.

✓ A simplified graph attention mechanism:

- For a node i and its neighbor j, with X_i and X_j representing their respective features, a simplified graph attention mechanism used in this paper is defined as:

$$\Psi(X_i, X_j) \triangleq \begin{cases} t, & if \;\; X_i \cdot X_j \geq 0, \\ -t, & if \;\; X_i \cdot X_j < 0. \end{cases}$$

- $t > 0$ is referred to as the *attention intensity*.

Theoretical and Experimental Results

✓ The regimes that GAT works and fails.

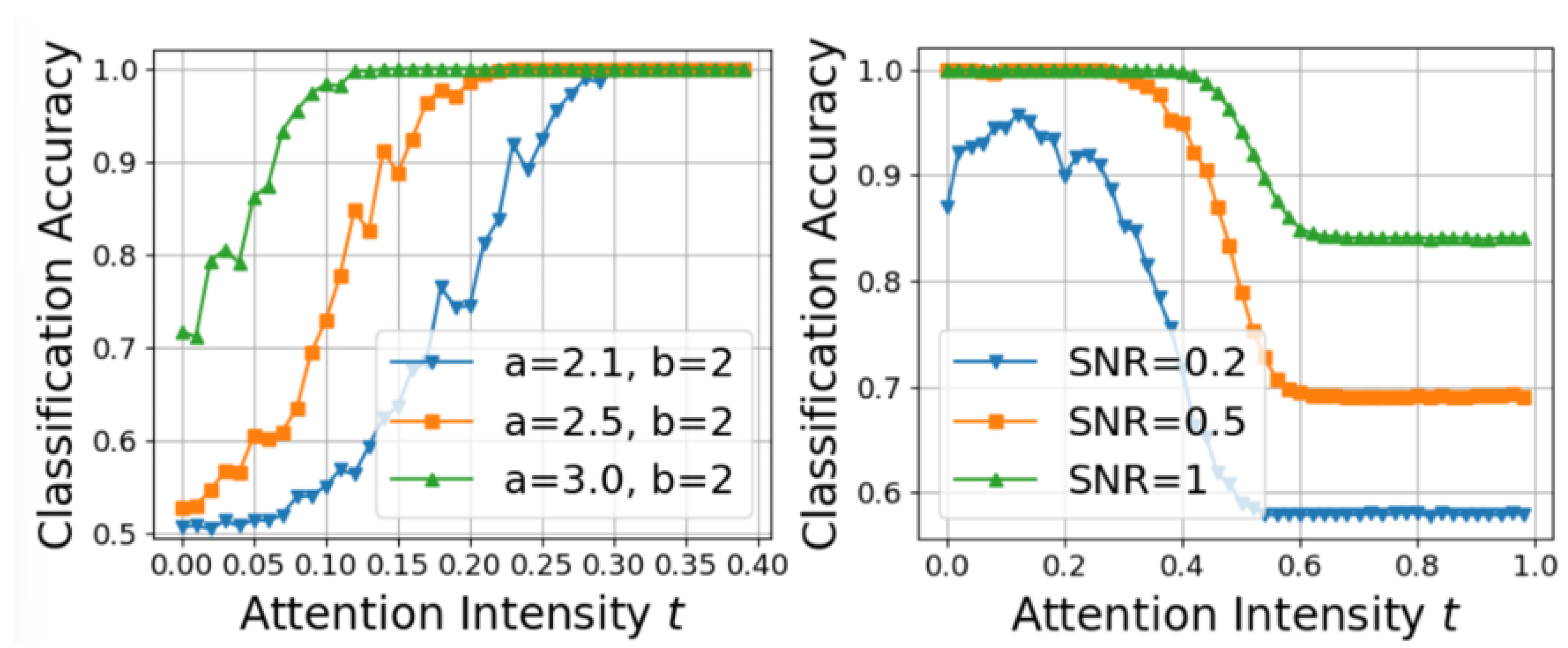
Theorem 2 and Corollary 1:

- Graph attention mechanism helps when
$$\mathcal{F}_{noise} = o\left(\frac{1}{\sqrt{\log n}}\right) \text{ and } \mathcal{S}_{noise} = \omega(1);$$
- Graph attention mechanism does not help when
$$\mathcal{F}_{noise} = \omega(1) \text{ and } \mathcal{S}_{noise} = O(1).$$

Insight:

- When structure noise dominates ($\mathcal{S}_{noise} \gg \mathcal{F}_{noise}$), graph attention mechanism is effective; when feature noise dominates ($\mathcal{F}_{noise} \gg \mathcal{S}_{noise}$) , GAT fails to work.

Validation Experiments on Synthetic Dataset



➤ $\mathcal{S}_{noise} \gg \mathcal{F}_{noise}$, **GAT works!** ➤ $\mathcal{F}_{noise} \gg \mathcal{S}_{noise}$, **GAT fails!**

✓ The impact on over-smoothing problem.

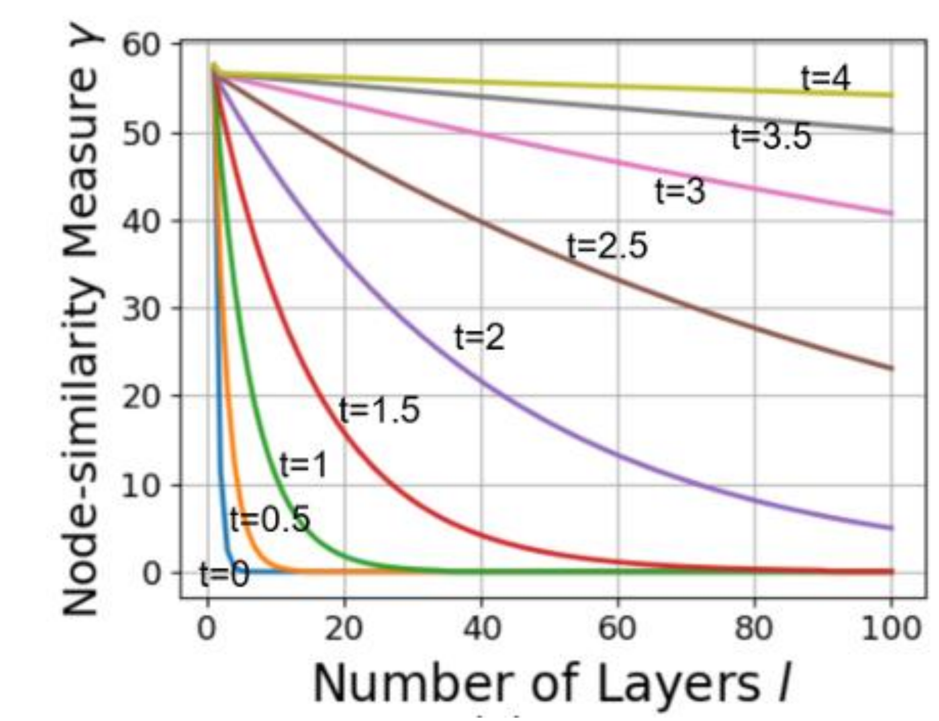
Theorem 3:

- Assume that $SNR = \omega(\sqrt{\log n})$. The graph convolutional networks suffer from over-smoothing. However, when $t = \omega(\sqrt{\log n})$, networks with graph attention mechanism can prevent this over smoothing problem.

Insight:

- In regimes where GAT works, with sufficiently strong attention intensity, GAT can solve the over-smoothing problem.

Validation Experiments on Synthetic Dataset



➤ γ measures node feature variance; smaller values imply greater similarity.

➤ $t = 0$ refers to GCN.

➤ As t increases, γ stops decaying exponentially with depth l, indicating the alleviation of over-smoothing problem.

✓ A new upper bound of exact recovery.

Theorem 4:

- When $SNR = \omega\left(\frac{\sqrt{\log n}}{\sqrt[3]{n}}\right)$, there exists a multi-layer GAT capable of achieving perfect node classification.

Insight:

- We provide the **first** upper bound for achieving exact recovery with **multi-layer** GAT networks on CSBM.
- Our result improves the bound from $SNR = \omega(\sqrt{\log n})$ (in [1]) to $\omega\left(\frac{\sqrt{\log n}}{\sqrt[3]{n}}\right)$, highlighting the benefit of using multiple layers in GAT.

[1] Fountoulakis K, et al. Graph attention retrospective. JMLR 2023.