Graph Convolutional Networks Achieve Information-Theoretic Limits under SBMs with Non-Informative Node Features

Zhongtian Ma, Qiaosheng Zhang, Lin Zhou, Xuelong Li, and Zhen Wang *

Abstract

We study semi-supervised node classification by graph convolutional networks (GCNs) on the symmetric stochastic block model (SBM) with n nodes and K communities, where edges connect nodes within the same community with probability p and across different communities with probability q. In the semi-supervised setting, the community labels of an η fraction of nodes are revealed *a priori*. We focus on a *non-informative feature* regime in which each node feature follows an i.i.d. Gaussian feature vector in $\mathbb{R}^d$ that is independent of the underlying community label. Despite the complete absence of feature signal, we show that a single-layer linear GCN trained by least squares admits strong theoretical guarantees when the feature dimension d exceeds a certain threshold. In the logarithmic-degree regime $p, q = \Theta(\log n/n)$, we prove that the GCN achieves exact recovery of node labels if $(\sqrt{p}-\sqrt{q})^2 > K \log n/n$ with only a vanishing proportion η of labeled nodes. We further prove a matching converse result under this semi-supervised setting, showing that the proposed GCN attains the information-theoretic limit for exact recovery. In the broader regime $1/n \ll p, q \ll 1$, we further show that, under explicit scaling conditions on the feature dimension d and *still* a vanishing η, the expected misclassification rate achieved by the GCN is $\exp\big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\big)$. This coincide with the optimal minimax misclassification rate of the unsupervised SBM (i.e., the community detection problem).

1 Introduction

Graph convolutional networks (GCNs) have emerged as a powerful framework for semi-supervised node classification on graphs, with strong empirical performance in diverse domains [1, 2, 3]. A central theoretical question is to *understand when and why GCNs succeed*, and *how their performance depends on the interplay between graph structure, node features, and the availability of labeled data*.

A growing body of work analyzes GCNs under probabilistic graph models, most notably the stochastic block model (SBM) and its extensions [4]. In a classical symmetric SBM with n nodes and K communities, each node belongs to one of the K communities with equal probability. Edges are generated independently: two nodes in the same community are connected with probability p, while two nodes in different communities are connected with probability q. The parameters n, p, q therefore control both the sparsity of the graph and the strength of the community signal. Since the operating mechanism of GCNs is to aggregate neighborhood node features through a *message-passing operator*, most existing theoretical analyses relies on the contextual SBM (i.e., SBM with node features) in which node features are generated from a Gaussian mixture model aligned with community labels [5]. Such node features are *informative* in the sense that their distributions differ across communities and therefore carry explicit information about the underlying labels.

In contrast, much less is understood in the extreme regime where node features are completely *non-informative*, in the sense that their distributions are identical across communities. This setting isolates the contribution of graph topology alone, thereby enabling a clean investigation of how GCNs exploit graph information and what fundamental performance limits such exploitation can achieve. As such, *it provides a stringent test of GCN expressivity and offers theoretical insights into the design of architectures that more effectively leverage graph structure.*

*Z. Ma is with the School of Cyberspace Security, Northwestern Polytechnical University, Xi'an, Shaanxi, China (mazhongtian@mail.nwpu.edu.cn). Q. Zhang is with the Shanghai AI Laboratory and Shanghai Innovation Institute, Shanghai, China (zhangqiaosheng@pjlab.org.cn). Z. Lin is with the School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China (zhoul9@sustech.edu.cn). X. Li is with the Institute of Artificial Intelligence of China Telecom (TeleAI), China Telecom Corp Ltd, China (xuelong_li@ieee.org). Z. Wang is with the School of Cyberspace Security, Northwestern Polytechnical University, Xi'an, Shaanxi, China (w-zhen@nwpu.edu.cn).

In this paper, we provide an information-theoretic characterization of semi-supervised node classification by GCNs under the symmetric SBM, where node features are i.i.d. Gaussian and independent of the labels. We focus on a linearized single-layer GCN and analyze its closed-form prediction rule induced by a regularized least-squares loss. We study two fundamental performance metrics—*exact recovery*, which requires correctly classifying all nodes with high probability, and *misclassification rate*, which quantifies the fraction of incorrectly classified nodes—that are widely used in unsupervised node classification (i.e., community detection) and whose information-theoretic limits are by now well understood [6, 7, 8]. However, it remains unclear whether, and under what conditions, modern GCN-based semi-supervised methods can attain these information-theoretic limits in the non-informative feature setting.

A key insight from our analysis is the critical role played by the feature dimension d and the labeled proportion η. We show that sufficiently high-dimensional random features can amplify the structural signal encoded in the graph, allowing the noise induced by the features to be effectively averaged out. Our main contributions are summarized as follows:

- We establish a *sharp threshold* for exact recovery in semi-supervised node classification using single-layer linear GCNs in the logarithmic-degree regime (i.e., $p, q = \Theta(\log n/n)$). When the feature dimension d is sufficiently large (i.e., $d = \omega(1/\log n)$), we show that exact recovery is achievable with a vanishing labeled proportion η if and only if $(\sqrt{p} - \sqrt{q})^2 > K \log n/n$ (see Theorems 1 and 2). This demonstrates that the proposed GCN attains the information-theoretic limit for exact recovery.

- We characterize the misclassification rate of GCNs in the broader regime $1/n \ll p, q \ll 1$. We derive a non-asymptotic upper bound on the expected misclassification rate and show that, under appropriate scaling of the feature dimension d and a vanishing labeled proportion η, the resulting misclassification rate satisfies $\exp\big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2\big)$, which coincides with the *minimax-optimal rate* for the unsupervised node classification (i.e., community detection) established in [7]. This result demonstrates that, using only a vanishing fraction of labeled nodes, a simple single-layer GCN can achieve the same performance as optimal unsupervised estimators.[1]

1.1 Detailed Comparison with Prior Work

1.1.1 Relation to CSBM-Based Analyses of GCNs

Most existing theoretical analyses of GCNs are conducted under the CSBM, where node features are informative and aligned with community labels [5, 9]. Under this assumption, prior work studies how message passing and architectural choices improve classification performance, including effects of nonlinear activation function, oversmoothing, and attention mechanisms [10, 11, 12, 13, 14, 15], as well as exact recovery thresholds in the semi-supervised CSBM [16].

In contrast, while following the CSBM assumption that node features are Gaussian, we focus on an extreme regime in which the features are completely non-informative (i.e., i.i.d. Gaussian and independent of the node labels), yielding a special case of the CSBM. This regime is of independent interest, as it provides a natural baseline for assessing how effectively GCNs exploit *purely graph-structural information* when node features carry no label signal.

Although our setting is formally a restriction of the CSBM, existing CSBM-based analyses *do not apply*. Prior works typically assume that node features contain sufficient initial information and study how graph structure can further amplify feature separability. This assumption breaks down in the non-informative regime considered here, rendering our setting technically more challenging despite being a special case.

1.1.2 Relation to Community Detection and Information-Theoretic Limits

Semi-supervised node classification on the SBM is closely related to the classical problem of community detection. When the labeled proportion $\eta = 0$, it reduces exactly to unsupervised community detection; more generally, it can be interpreted as community detection with side information in the form of partially observed node labels [17, 18].

[1] We note that it remains an interesting open question whether the minimax misclassification rate remains unchanged in the semi-supervised setting when only a vanishing fraction of node labels is revealed.

Community detection under the SBM is a canonical problem with well-understood information-theoretic limits, including sharp thresholds for exact recovery and minimax misclassification rates [6, 19, 4, 7, 8]. In the logarithmic-degree regime $p, q = \Theta(\log n/n)$, exact recovery is possible if and only if $(\sqrt{p} - \sqrt{q})^2 > K \log n/n$ [4]; in the sparse regime, the optimal misclassification rate scales as $\exp\big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2\big)$ [7]. A variety of polynomial-time algorithms are known to achieve these limits under suitable conditions, including spectral methods [20, 21, 22], semidefinite relaxations [23, 24], and belief propagation [25, 26].

1.2 Notations

We use $[n] := \{1, 2, \ldots, n\}$ to denote the index set. The indicator function is denoted by $\mathbb{1}\{\cdot\}$, which equals 1 if the condition holds and 0 otherwise. Let δ_{ij} denote the Kronecker delta. We use $\mathbf{1}$ to denote the all-ones vector. Standard asymptotic notations $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\omega(\cdot)$ are used in their conventional sense. The notation $\gg$ has the same meaning as $\omega(\cdot)$, while $\ll$ corresponds to $o(\cdot)$. Throughout the paper, all asymptotic notations are taken with respect to $n \to \infty$, unless otherwise specified.

2 Semi-Supervised Node Classification on the Symmetric SBM

2.1 K-Symmetric Stochastic Block Model

We consider an undirected random graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = [n]$ and adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$. Each node $i \in \mathcal{V}$ is associated with a latent community label $\sigma_i \in [K]$, forming the assignment vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$.

Community labels are generated independently and uniformly at random:

$$\mathbb{P}(\sigma_i = k) = 1/K, \quad k \in [K],\ i \in [n].$$

As a consequence, each community $\mathcal{V}_k := \{i : \sigma_i = k\}$ satisfies $|\mathcal{V}_k| = (1 + o(1))\, n/K$ with high probability.

Conditional on $\boldsymbol{\sigma}$, edges are generated independently as

$$\mathbb{P}(\mathbf{A}_{ij} = 1 \mid \boldsymbol{\sigma}) = \mathbf{B}_{\sigma_i\sigma_j}, \quad \mathbf{A}_{ij} = \mathbf{A}_{ji}, \quad \mathbf{A}_{ii} = 0,$$

where the connectivity matrix $\mathbf{B} \in [0,1]^{K\times K}$ satisfies

$$\mathbf{B}_{kk} = p, \quad \mathbf{B}_{kl} = q, \quad k \neq l.$$

We denote this model by $\mathrm{SBM}(n, K, p, q)$.

2.2 Semi-Supervised Node Classification

We consider a semi-supervised learning setting, where only a subset of nodes

$$\mathcal{L} := \{i \in \mathcal{V} : \sigma_i \text{ is observed}\}$$

have known labels, while the remaining nodes $\mathcal{U} := \mathcal{V} \setminus \mathcal{L}$ are unlabeled. Under the K-symmetric SBM, we assume that each class contributes the same proportion of labeled nodes. Specifically, for a given $\eta \in (0, 1)$ and for each $k \in [K]$, we define

$$\mathcal{L}_k := \mathcal{L} \cap \mathcal{V}_k,$$

where $\mathcal{L}_k$ consists of a uniformly random subset of $\mathcal{V}_k$ of fixed size $|\mathcal{L}_k| = \eta\, |\mathcal{V}_k|$.

An estimator is any measurable mapping

$$\hat{\sigma} = \hat{\sigma}\big(\mathbf{A}, \mathcal{L}, \{\sigma_i : i \in \mathcal{L}\}\big) \in [K]^n,$$

which outputs a predicted label $\hat{\sigma}_i$ for each node $i \in \mathcal{V}$.

We define the misclassification rate as

$$\mathrm{err}(\hat{\sigma}, \sigma) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\sigma}_i \neq \sigma_i\}. \tag{1}$$

This metric measures the fraction of nodes whose predicted labels differ from their true community assignments. The label permutation ambiguity in unsupervised setting is absent in our semi-supervised setting, as each community contains at least one node with a revealed true label.

3 Graph Convolutional Networks as Estimators

3.1 Node Feature Generation

Since GCNs operate on attributed graphs where each node contains a feature vector. Importantly, we consider a non-informative node features setting, where node features are independent of the community structure.

Let

$$\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$$

denote the node feature matrix, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector of node i. Node features are generated as i.i.d. Gaussian random vectors:

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad i \in [n].$$

3.2 Linearized Graph Convolutional Networks

Let $\mathbf{D}$ be a diagonal matrix with $\mathbf{D}_{ii} = \frac{1}{n}\sum_{i,j=1}^{n} \mathbf{A}_{ij}$, [2] and define the message-passing operator

$$\mathbf{P} := \mathbf{D}^{-1}\mathbf{A}.$$

Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{n \times d_\ell}$ denote the node representations at layer ℓ, with each row corresponding to the embedding of a node. An L-layer GCN computes these representations recursively via

$$\mathbf{H}^{(0)} = \mathbf{X}, \quad \mathbf{H}^{(\ell)} = \phi\Big(\mathbf{P}\,\mathbf{H}^{(\ell-1)}\mathbf{W}^{(\ell-1)}\Big), \quad \ell = 1, \ldots, L,$$

where $\{\mathbf{W}^{(\ell-1)}\}_{\ell \in [L]}$ are trainable weight matrices and $\phi(\cdot)$ is an activation function.

For theoretical tractability, we consider the linearized GCN obtained by removing nonlinearities:

$$\mathbf{H}^{(L)} = \mathbf{P}^L \mathbf{X}\mathbf{W},$$

where $\mathbf{W} := \prod_{\ell=1}^{L} \mathbf{W}^{(\ell-1)} \in \mathbb{R}^{d \times K}$. We denote the resulting node embeddings by

$$\mathbf{Z} := \mathbf{P}^L \mathbf{X} \in \mathbb{R}^{n \times d}.$$

Although the formulation allows for general depth L, we find that a single-layer GCN is already sufficient to achieve the information-theoretic limits for community label recovery in the regimes of interest. Increasing the depth does not further improve the achievable performance in our setting. Therefore, we focus exclusively on the single-layer case ($L = 1$) in the following, for which

$$\mathbf{Z} = \mathbf{P}\mathbf{X}.$$

3.3 Least-Squares Training and Score-Based Prediction

Let $\mathbf{Y} \in \{0,1\}^{n \times K}$ be the *one-hot label matrix* with

$$\mathbf{Y}_{ik} = \mathbb{1}\{\sigma_i = k\},$$

and let $\mathbf{Z}_{\mathcal{L}}$ and $\mathbf{Y}_{\mathcal{L}}$ denote the submatrices of $\mathbf{Z}$ and $\mathbf{Y}$ restricted to the labeled node set $\mathcal{L}$.

We train the linearized GCN by minimizing the regularized least-squares loss[3]

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times K}} \ \|\mathbf{Z}_{\mathcal{L}}\mathbf{W} - \mathbf{Y}_{\mathcal{L}}\|_F^2 + \lambda\|\mathbf{W}\|_F^2, \quad \lambda > 0. \tag{2}$$

[2] Here we use the global average degree to ensure concentration for each node in sparse graphs, whereas it is known that simultaneous concentration of all node degrees cannot be guaranteed in the sparse regime; this choice is justified under the symmetric SBM setting and is inspired by [16].

[3] The regularized least-squares loss, equivalent to ridge regression, is a standard choice in practice for supervised and semi-supervised classification. While cross-entropy loss is more widely used in empirical GCN implementations, it does not admit a closed-form solution. To enable a clean and tractable theoretical analysis, we therefore adopt the regularized least-squares formulation in this work.

```
Algorithm 1: Single-layer linearized GCN estimator
Input: Adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$; feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$; labeled set $\mathcal{L} \subseteq \mathcal{V}$ and one-hot
       labels $\mathbf{Y}_{\mathcal{L}} \in \{0,1\}^{|\mathcal{L}|\times K}$.
Output: Predicted labels $\hat{\sigma} \in [K]^n$.
(Message-passing operator)
Compute $\mathbf{D} \in \mathbb{R}^{n\times n}$ with $\mathbf{D}_{ii} = \frac{1}{n}\sum_{i,j=1}^{n} \mathbf{A}_{ij}$ for all $i \in [n]$
Set $\mathbf{P} \leftarrow \mathbf{D}^{-1}\mathbf{A}$
(Single-layer linearized embedding)
Compute $\mathbf{Z} \leftarrow \mathbf{P}\mathbf{X}$
(Gram matrix and simplified score)
Compute $\mathbf{G} \leftarrow \mathbf{Z}\mathbf{Z}^\top$
Compute $\hat{\mathbf{S}} \leftarrow \mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}}$
(Prediction)
for $i = 1$ to $n$ do
    $\hat{\sigma}_i \leftarrow \arg\max_{k\in[K]} \hat{\mathbf{S}}_{ik}$
return $\hat{\sigma}$
```

This objective admits a closed-form solution given by

$$\mathbf{W}^\star = (\mathbf{Z}_{\mathcal{L}}^\top \mathbf{Z}_{\mathcal{L}} + \lambda \mathbf{I})^{-1} \mathbf{Z}_{\mathcal{L}}^\top \mathbf{Y}_{\mathcal{L}}. \tag{3}$$

Define the Gram matrix $\mathbf{G} := \mathbf{Z}\mathbf{Z}^\top$. The corresponding score matrix produced by the trained GCN is

$$\mathbf{S}^\star = \mathbf{Z}\mathbf{W}^\star = \mathbf{G}_{:\mathcal{L}}\big(\mathbf{G}_{\mathcal{L}\mathcal{L}} + \lambda \mathbf{I}\big)^{-1}\mathbf{Y}_{\mathcal{L}},$$

where $\mathbf{S}^\star_{ik}$ represents the score assigned to node i for class k.

In the symmetric SBM setting, the inverse factor $\big(\mathbf{G}_{\mathcal{L}\mathcal{L}} + \lambda \mathbf{I}\big)^{-1}$ primarily rescales scores within the subspace spanned by labeled nodes and does not affect the relative ordering of class scores under the regimes of interest. To streamline the analysis, we therefore focus on the simplified (unnormalized) score matrix

$$\hat{\mathbf{S}} := \mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}}. \tag{4}$$

The predicted label of node i is then given by

$$\hat{\sigma}_i = \arg\max_{k\in[K]} \hat{\mathbf{S}}_{ik}, \quad i \in \mathcal{V}.$$

The computational procedure of the resulting GCN estimator is summarized in Algorithm 1.

4 Main Results

In this section, we analyze the performance of a single-layer linearized GCN estimator. We first study exact recovery in the logarithmic-degree regime and show that the GCN attains the sharp information-theoretic threshold, matching the information-theoretic lower bounds for semi-supervised node classification. We then turn to a broader regime and derive bounds on the misclassification rate, demonstrating that the GCN achieves the information-theoretic misclassification rate of the unsupervised setting using only a vanishing fraction of labeled nodes.

4.1 Sharp Exact Recovery Threshold in the Logarithmic-Degree Regime

Theorem 1 (Exact recovery by a single-layer GCN). *Consider a symmetric* SBM(n, K, p, q) *with* $K = O(1)$ *almost equal-sized communities, where*

$$p = \frac{a\log n}{n}, \quad q = \frac{b\log n}{n}, \quad |p-q| = \Theta(p+q).$$

Assume that

$$\frac{a}{K} + \frac{K-1}{K} b > 3 + \alpha, \quad \eta = \omega\left(\frac{1}{\log n}\right), \quad d = \omega(n).$$

Then, a single-layer linearized GCN estimator achieves exact recovery with probability at least $1 - o(1)$ *if and only if*

$$(\sqrt{a} - \sqrt{b})^2 > K, \quad \textit{or equivalently} \quad (\sqrt{p} - \sqrt{q})^2 > K\,\frac{\log n}{n}.$$

Proof. See Appendix for the detailed proof. □

Theorem 2 (Information-theoretic lower bound for semi-supervised exact recovery)**.** *Consider a symmetric* $\mathrm{SBM}(n, K, p, q)$ *with* K *equal-sized communities, where*

$$p = \frac{a \log n}{n}, \quad q = \frac{b \log n}{n}, \quad a > b > 0.$$

Suppose we are in the semi-supervised setting described in Section 2, where a fraction $\eta \in (\frac{\log^2 n}{n}, 1)$ *of node labels is revealed uniformly at random. Define* $\beta := \frac{\log \frac{1}{1-\eta}}{\log n}$*. If*

$$(\sqrt{a} - \sqrt{b})^2 + K\beta < K,$$

then exact recovery is information-theoretically impossible. In particular, for any estimator $\hat{\sigma}$ *that has access to the graph and the revealed labels,* $\liminf_{n\to\infty} \mathbb{P}(\hat{\sigma} = \sigma) = 0$.

Proof. The proof of Theorem 2 is obtained by extending Theorem 1 of [17] from two symmetric communities to K symmetric communities, following the technique developed in [27]. Moreover, [17] assumes that each node is labeled independently with probability η; when $\eta \in \left(\frac{\log^2 n}{n}, 1\right)$, the labeled fraction concentrates around η with high probability, which is consistent with our setting. □

Theorems 1 and 2 together provide a refined understanding of exact recovery under the stochastic block model in the semi-supervised setting.

Firstly, Theorem 1 shows that a single-layer GCN achieves exact recovery whenever $(\sqrt{a} - \sqrt{b})^2 > K$, even when only a vanishing fraction of node labels is revealed, for instance by choosing $\eta = 1/\log^2 n$.

Secondly, Theorem 2 establishes an information-theoretic lower bound for semi-supervised exact recovery. By choosing an appropriate η, such as $\frac{1}{\log n} \ll \eta \ll 1 - \frac{1}{n}$, the correction term β becomes asymptotically negligible. In this regime, the information-theoretic lower bound reduces to $(\sqrt{a} - \sqrt{b})^2 > K$. According to Theorem 1, a single-layer GCN achieves exact recovery whenever this condition holds. Therefore, the proposed GCN attains the information-theoretic limit for exact recovery.

Finally, the requirement $d = \omega(n)$ highlights the role of high-dimensional random features in enabling the GCN to extract and preserve structural information through the aggregation $\mathbf{AX}$. In this regime, the randomness introduced by the features does not degrade performance, and the single-layer GCN effectively reduces the problem to community detection under the SBM at the information-theoretic scale.

4.2 Misclassification Rate in the Broader Regime

Theorem 3 (Misclassification rate of a single-layer GCN)**.** *Consider a symmetric* $\mathrm{SBM}(n, K, p, q)$ *with* $K = O(1)$ *almost equal-sized communities, where*

$$\frac{1}{n} \ll p, q \ll 1, \quad |p - q| = \Theta(p + q).$$

If the feature dimension d *and labeled proportion* η *satisfies*

$$d = \omega\left(\frac{2 \log n}{\log\!\big(1 + R^*(\mathbf{A})\big)}\right), \quad \text{and} \quad \eta = \omega\left(\frac{1}{np}\right),$$

where

$$R^*(\mathbf{A}) = \min_{i\in[n],\, k\neq r} \frac{\gamma^2}{\alpha\beta - \gamma^2}$$

and

$$\begin{aligned}\alpha(i) &:= \mathrm{Var}(Z \mid \mathbf{A}) = (\mathbf{A}\mathbf{A}^\top)_{ii},\\ \beta(k,r) &:= \mathrm{Var}(V \mid \mathbf{A}) = (\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r})^\top \mathbf{A}\mathbf{A}^\top (\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r}),\\ \gamma(i,k,r) &:= \mathrm{Cov}(Z,V \mid \mathbf{A}) = \boldsymbol{e}_i^\top \mathbf{A}\mathbf{A}^\top (\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r}).\end{aligned}$$

Then the expected misclassification rate of a single-layer linearized GCN estimator $\hat\sigma$ obeys

$$\mathbb{E}[\mathrm{err}(\hat\sigma,\sigma)] \;\le\; \exp\Big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\Big)\,.$$

Proof. Refer to Appendix for the detailed proof. □

Theorem 3 establishes an upper bound on the expected misclassification rate, characterizing how it decays as a function of the model parameters. Several remarks are in order:

(1). Theorem 3 shows that when the feature dimension d is sufficiently large, one may take $\eta = \frac{1}{\sqrt{np}}$, which vanishes asymptotically since $np = \omega(1)$. Under this choice, the resulting misclassification rate upper bound matches the information-theoretic limit of unsupervised setting derived in [7], which provides solid theoretical guarantees for GCNs.

(2). Unlike the exact recovery result in the denser regime, in the broader setting we are unable to provide a simple and intuitive condition on d expressed solely in terms of n. This is because the concentration of $\mathbf{A}\mathbf{A}^\top$ can no longer be guaranteed. Consequently, the required condition on d must be determined based on the realized adjacency matrix $\mathbf{A}$, as quantified by $R^*(\mathbf{A})$, which characterizes the corresponding algorithmic requirement.

(3). Although an explicit threshold for d is not available, a direct calculation shows that $\mathbb{E}[\gamma(i,k,\sigma(i))] = -\frac{\eta n^2}{K}(p-q)^2$, which indicates that the required feature dimension d is negatively correlated with $|p-q|$, n, and η.

5 Conclusion

We provided a theoretical characterization of single-layer linear GCNs in the non-informative feature regime under the symmetric SBM. Despite completely non-informative node features, a GCN can leverage graph structure together with a vanishing fraction of labeled nodes to achieve exact recovery at the SBM information-theoretic threshold and attain the minimax-optimal misclassification exponent in the broader regime.

Several directions remain open. First, it is of interest to understand whether deeper GCN architectures can further improve performance in sparser regimes by exploiting multi-hop structure. Second, extending the analysis to richer message-passing operators and to SBM extensions (e.g., degree heterogeneity or overlapping communities) may yield a broader understanding of when GNN-style semi-supervised methods are statistically optimal.

A Proofs of Theorem 1 and Theorem 3

Recall that for any node i, the predicted label is given by

$$\hat\sigma_i = \arg\max_{k\in[K]} \hat{\mathbf{S}}_{ik} = \arg\max_{k\in[K]} (\mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}})_{ik}.$$

We now analyze the term $(\mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}})_{ik}$ in detail. For clarity, we begin with the case of a single-layer GCN. In this setting,

$$\hat{\mathbf{S}}_{ik} = (\mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}})_{ik} = \sum_{\ell\in\mathcal{L}_k} \mathbf{G}_{i\ell} = \sum_{\ell\in\mathcal{L}_k} (\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top)_{i\ell}. \tag{5}$$

A.1 Local error event analysis of score matrix $\hat{\mathbf{S}}$

We begin with a local analysis of the $\hat{\mathbf{S}}$ by examining its (i,k)-th entry $\hat{\mathbf{S}}_{ik}$. For any $k \neq \sigma(i)$, let $\Delta_i^{(k)} := \hat{\mathbf{S}}_{ik} - \hat{\mathbf{S}}_{i\sigma(i)}$. Then node i is misclassified if and only if $\max_{k\neq\sigma(i)} \Delta_i^{(k)} = \max_{k\neq\sigma(i)} \hat{\mathbf{S}}_{ik} - \hat{\mathbf{S}}_{i\sigma(i)} > 0$. The randomness of $\Delta_i^{(k)}$ arises from two independent sources: the SBM-generated matrix $\mathbf{P}$ and the Gaussian feature matrix $\mathbf{X}$. Owing to their independence, we decompose the analysis into two steps: first, we establish the concentration of $\Delta_i^{(k)} \mid \mathbf{P}$; second, we analyze the concentration properties of $\mathbf{P}\mathbf{P}^\top$.

First, by Lemma 1 and a union bound, we have that with high probability, uniformly over all nodes, the following statements hold:

$$\begin{aligned} &\text{(i)} \quad |\mathcal{V}_k| = (1+o(1))\,\frac{n}{K}, \qquad \text{for all } k \in [K],\\ &\text{(ii)} \quad \mathbf{D}_{ii} = (1+o(1))\,\overline{\deg} := (1+o(1))\,n\left(\frac{p}{K} + \frac{(K-1)q}{K}\right). \end{aligned}$$

Consequently,

$$\mathbf{P}\mathbf{P}^\top \asymp (1+o(1))\,\frac{1}{\overline{\deg}^2}\,\mathbf{A}\mathbf{A}^\top.$$

All subsequent analysis is conducted under these high-probability events.

Consider the randomness of X. Since the rows of $\mathbf{X}$ are i.i.d. as $\mathcal{N}(0,\mathbf{I}_d)$, each entry of $\mathbf{X}$ is i.i.d. $\mathcal{N}(0,1)$. Consequently, the columns of $\mathbf{X}$ are also i.i.d. and each follows $\mathcal{N}(0,\mathbf{I}_n)$. Therefore,

$$\mathbf{X}\mathbf{X}^\top = \sum_{r=1}^{d} \boldsymbol{x}^{(r)}(\boldsymbol{x}^{(r)})^\top,$$

where $\{\boldsymbol{x}^{(r)}\}_{r=1}^d$ are i.i.d. $\mathcal{N}(0,\mathbf{I}_n)$.

Then we obtain

$$\hat{\mathbf{S}}_{ik} = \mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top = \frac{1}{\overline{\deg}^2}\sum_{\ell\in\mathcal{L}_k}\sum_{r=1}^{d}(\mathbf{A}\boldsymbol{x}^{(r)})_i\,(\mathbf{A}\boldsymbol{x}^{(r)})_\ell.$$

This further implies

$$\Delta_i^{(k)} := \hat{\mathbf{S}}_{ik} - \hat{\mathbf{S}}_{i\sigma(i)} = \frac{1}{\overline{\deg}^2}\sum_{r=1}^{d}(\mathbf{A}\boldsymbol{x}^{(r)})_i\left(\sum_{\ell\in\mathcal{L}_k}(\mathbf{A}\boldsymbol{x}^{(r)})_\ell - \sum_{\ell\in\mathcal{L}_{\sigma(i)}}(\mathbf{A}\boldsymbol{x}^{(r)})_\ell\right). \tag{6}$$

Define

$$Z_r := (\mathbf{A}\boldsymbol{x}^{(r)})_i, \qquad V_r := \sum_{\ell\in\mathcal{L}_k}(\mathbf{A}\boldsymbol{x}^{(r)})_\ell - \sum_{\ell\in\mathcal{L}_{\sigma(i)}}(\mathbf{A}\boldsymbol{x}^{(r)})_\ell.$$

Conditioned on $\mathbf{A}$, both Z and V are mean-zero Gaussian random variables. A straightforward calculation shows that

$$\begin{aligned} \alpha(i) &:= \mathrm{Var}(Z \mid \mathbf{A}) = (\mathbf{A}\mathbf{A}^\top)_{ii},\\ \beta(k,\sigma(i)) &:= \mathrm{Var}(V \mid \mathbf{A}) = (\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_{\sigma(i)}})^\top\mathbf{A}\mathbf{A}^\top(\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_{\sigma(i)}}),\\ \gamma(i,k,\sigma(i)) &:= \mathrm{Cov}(Z,V \mid \mathbf{A}) = \boldsymbol{e}_i^\top\mathbf{A}\mathbf{A}^\top(\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_{\sigma(i)}}), \end{aligned} \tag{7}$$

where $\mathbf{1}_{\mathcal{L}_k}$ denotes the indicator vector whose entries are equal to 1 for indices in $\mathcal{L}_k$ and 0 otherwise, and $\boldsymbol{e}_i$ denotes the i-th standard basis vector, with a 1 in the i-th position and zeros elsewhere.

For an error event $\Delta_i^{(k)} > 0$, we have

$$\mathbb{P}(\Delta_i^{(k)} > 0) = \mathbb{P}(e^{t\,\Delta_i^{(k)}} > 1) \le \mathbb{E}[e^{t\,\Delta_i^{(k)}}] := M_t(\Delta_i^{(k)}),$$

where $M_t(\Delta_i^{(k)})$ is the Moment Generating Function (MGF) of $\Delta_i^{(k)}$.

First, consider the conditional MGF of ZV, since Z, V are two zero-mean gaussian variables, according to Lemma 2, we have

$$M_t(ZV \mid \mathbf{A}) = \frac{1}{\sqrt{1 - 2t\gamma - t^2(\alpha\beta - \gamma^2)}}.$$

Since $\{\boldsymbol{x}^{(r)}\}_{r=1}^d$ are i.i.d., it follows that $\{Z_r\}_{r=1}^d$ and $\{V_r\}_{r=1}^d$ are also i.i.d.. Then, by (6), we obtain

$$M_t\Big(\Delta_i^{(k)} \mid \mathbf{A}\Big) = \mathbb{E}\Big[\exp\Big(t\Delta_i^{(k)}\Big) \mid \mathbf{A}\Big] = \prod_{r=1}^{d} \mathbb{E}\left[\exp\left(\frac{t}{\deg^2} Z_r V_r\right) \middle| \mathbf{A}\right] = \frac{1}{\big(1 - 2t\gamma - t^2(\alpha\beta - \gamma^2)\big)^{d/2}}.$$

The optimal choice of the parameter can be obtained as

$$t^* = \frac{-\gamma}{\alpha\beta - \gamma^2}.$$

Substituting this value yields

$$\mathbb{P}\Big(\Delta_i^{(k)} > 0 \mid \mathbf{A}\Big) \le \exp\Big(-\frac{d}{2}\log\big(1 - 2t^*\gamma - (t^*)^2(\alpha\beta - \gamma^2)\big)\Big) = \exp\Big(-\frac{d}{2}\log\Big(1 + \frac{\gamma}{\alpha\beta - \gamma^2}\Big)\Big).$$

Let

$$R_i(k, \sigma(i); \mathbf{A}) := \frac{\gamma^2}{\alpha\beta - \gamma^2}.$$

Hence, if

$$\frac{d}{2}\log\Big(1 + \frac{\gamma^2}{\alpha\beta - \gamma^2}\Big) \gg \log n \quad \Longrightarrow \quad d \gg \frac{2\log n}{\log\big(1 + R_i(k, \sigma(i); \mathbf{A})\big)},$$

we obtain

$$\mathbb{P}\Big(\Delta_i^{(k)} > 0 \mid \mathbf{A}\Big) \le e^{-\omega(\log n)} = o(n^{-1}) \xrightarrow{\text{union bound}} \mathbb{P}\Big(\forall i \in [n] : \Delta_i^{(k)} > 0 \mid \mathbf{A}\Big) = o(1)$$

This shows that the randomness introduced by $\mathbf{X}$ has been effectively eliminated, and the number of misclassified nodes is entirely determined by the randomness of $\mathbf{A}$.

Since the true label $\sigma(i)$ is unknown in practice, we instead use

$$R^*(\mathbf{A}) := \min_{i \in [n],\, k \neq s} R_i(k, s; \mathbf{A})$$

as a worst-case lower bound over all nodes and cluster pairs. Consequently, the condition on d becomes

$$d \gg \frac{2\log n}{\log\big(1 + R^*(\mathbf{A})\big)}.$$

Next, we show the concentration result of $R_i(k, \sigma(i); \mathbf{A})$ in the denser regime. For the regime $p, q = \Theta(\log n/n)$, according to Lemma 4, if

$$\frac{a}{K} + \frac{K-1}{K} b > 3 + \alpha, \quad \text{and} \quad \eta = \omega(\frac{1}{\log n}),$$

simply applying chernoff bound and union bound yields that the following events hold for all nodes with high probability:

$$\begin{aligned} \alpha &= \Theta(\mathbb{E}_{\mathbf{A}}[\alpha]) = \Theta(np) = \Theta(\log n), \\ \beta &= \Theta(\mathbb{E}_{\mathbf{A}}[\beta]) = \Theta(\eta^2 n^3 p^2) = \Theta(\eta^2 n \log^2 n), \\ \gamma &= \Theta(\mathbb{E}_{\mathbf{A}}[\gamma]) = \Theta(\eta n^2 p^2) = \Theta(\eta \log^2 n). \end{aligned}$$

Thus, for all $i \in [n]$, we have

$$R_i(k, \sigma(i); \mathbf{A}) \overset{\text{w.h.p.}}{=} \Theta\left(\frac{\log n}{n}\right) = o(1).$$

Using the fact $\log(1+x) = x + o(x)$ for $x = o(1)$, it follows that

$$d = \omega\left(\frac{2\log n}{\log\bigl(1 + R_i(k, \sigma(i); \mathbf{A})\bigr)}\right) = \omega(n)$$

is sufficient to eliminate the effect of the randomness in $\mathbf{X}$.

In an even denser regime where $p, q = \omega(\log n/n)$, it suffices to assume $\eta = \omega\left(\frac{1}{np}\right)$ and $d = \omega\left(\frac{\log n}{p}\right)$.

Consider the randomness of P. At this point, the misclassification event is fully determined by the randomness of the adjacency matrix $\mathbf{A}$. Fix a node i with true label $\sigma(i) = r$, and consider any competing label $k \neq r$. Without considering the effect of $\mathbf{X}$, misclassification to class k can only occur if

$$\bar{\Delta}_i^{(k)} := \bar{\mathbf{S}}_{ik} - \bar{\mathbf{S}}_{ir} = \sum_{\ell \in \mathcal{L}_k} (\mathbf{A}\mathbf{A}^\top)_{i\ell} - \sum_{\ell \in \mathcal{L}_r} (\mathbf{A}\mathbf{A}^\top)_{i\ell} \geq 0,$$

and hence

$$\mathbb{P}(\widehat{\sigma}(i) = k) \ \leq \ \mathbb{P}\Bigl(\bar{\Delta}_i^{(k)} \geq 0\Bigr).$$

To analyze this probability, we expand the score difference into a sum of local contributions. For each $t \in [n]$, define

$$W_{j,k} := \sum_{\ell \in L_k} \mathbf{A}_{\ell j}, \qquad W_{j,r} := \sum_{\ell \in L_r} \mathbf{A}_{\ell j}, \qquad D_j := W_{j,k} - W_{j,r},$$

and set

$$X_j := \mathbf{A}_{ij}\, D_j.$$

With this notation,

$$\bar{\Delta}_i^{(k)} = \sum_{j=1}^{n} X_j.$$

We also bound the upper tail of this sum using a Chernoff argument. For any $t > 0$,

$$\mathbb{P}\Bigl(\bar{\Delta}_i^{(k)} \geq 0\Bigr) = \mathbb{P}\Bigl(e^{t\sum_t X_j} \geq 1\Bigr) \leq \mathbb{E}\exp\left(t\sum_{j=1}^{n} X_j\right) = M_t\bigl(\sum_{j=1}^{n} X_j\bigr).$$

Conditioning on the community label $u = \sigma(t)$, the edge indicator $\mathbf{A}_{ij}$ is Bernoulli with parameter

$$\rho_{r,u} = \begin{cases} p, & u = r, \\ q, & u \neq r, \end{cases}$$

and, up to negligible diagonal effects, the random variables $W_{j,k}$ and $W_{j,r}$ are independent binomials with

$$W_{j,k} \mid u \sim \mathrm{Bin}(m, \rho_{k,u}), \qquad W_{j,r} \mid u \sim \mathrm{Bin}(m, \rho_{r,u}),$$

where $m = |L_k| = \eta n/K$. Using the independence between $\mathbf{A}_{ij}$ and the edges contributing to $W_{j,k}$ and $W_{j,r}$, we obtain

$$\mathbb{E}\bigl[e^{tX_j} \mid u\bigr] = (1 - \rho_{r,u}) + \rho_{r,u}(1 - \rho_{k,u} + \rho_{k,u}e^t)^m(1 - \rho_{r,u} + \rho_{r,u}e^{-t})^m.$$

Evaluating this expression in the three cases $u = r$, $u = k$, and $u \notin \{r, k\}$ yields the functions

$$\begin{aligned} G_r(t) &= (1-p) + p(1 - q + qe^t)^m(1 - p + pe^{-t})^m, \\ G_k(t) &= (1-q) + q(1 - p + pe^t)^m(1 - q + qe^{-t})^m, \\ G_0(t) &= (1-q) + q(1 - q + qe^t)^m(1 - q + qe^{-t})^m. \end{aligned}$$

Summing the contributions over all $t \in [n]$ and grouping by community membership, we obtain the log-mgf bound

$$\log M_t(\bar{\Delta}_i^{(k)}) = \log \mathbb{E}\exp\Bigl(t\bar{\Delta}_i^{(k)}\Bigr) \leq \frac{n}{K}\log G_r(t) + \frac{n}{K}\log G_k(t) + \frac{(K-2)n}{K}\log G_0(t) + o\bigl(n(p+q)\bigr).$$

Consequently,

$$\mathbb{P}\Big(\bar{\Delta}_i^{(k)} \geq 0\Big) \leq \exp\big(\Psi_n(t) + o(n(p+q))\big),$$

where

$$\Psi_n(t) := \frac{n}{K}\log G_r(t) + \frac{n}{K}\log G_k(t) + \frac{(K-2)n}{K}\log G_0(t).$$

We now specialize to the regime $p, q = \omega(1/n)$, $p, q = o(1)$, and $(p+q)/(n(p-q)^2) = o(1)$. With $m = \eta n/K$, we choose

$$t = t(x) := \frac{x}{m(p-q)}, \qquad x > 0 \text{ constant.}$$

Under the assumption $\eta = \omega(\frac{1}{np})$, we have $m(p+q)t^2 = o(1)$, and a Taylor expansion yields

$$(1-q+qe^t)^m(1-p+pe^{-t})^m = \exp\big(-x + o(1)\big),$$

$$(1-p+pe^t)^m(1-q+qe^{-t})^m = \exp\big(x + o(1)\big), \qquad (1-q+qe^t)^m(1-q+qe^{-t})^m = 1 + o(1).$$

Substituting these expressions gives

$$\Psi_n(t(x)) = \frac{n}{K}\log\big((1-p) + pe^{-x}\big) + \frac{n}{K}\log\big((1-q) + qe^{x}\big) + o\big(n(p+q)\big).$$

Optimizing over $x > 0$ yields

$$\inf_{x>0}\Psi_n(t(x)) = \frac{2n}{K}\log\Big(\sqrt{(1-p)(1-q)} + \sqrt{pq}\Big) + o\big(n(p+q)\big).$$

Since $p, q = o(1)$,

$$\sqrt{(1-p)(1-q)} + \sqrt{pq} = 1 - \frac{(\sqrt{p}-\sqrt{q})^2}{2} + o(p+q),$$

and therefore

$$\mathbb{P}\Big(\bar{\Delta}_i^{(k)} \geq 0\Big) \leq \exp\Big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\Big).$$

A union bound over $k \neq r$ yields the single-node misclassification probability

$$\mathbb{P}(\widehat{\sigma}(i) \neq \sigma(i)) \leq (K-1)\exp\Big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\Big).$$

Averaging over all nodes, and note that K is a constant, we obtain

$$\mathbb{E}[\mathrm{Err}] \leq (K-1)\exp\Big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\Big) = \exp\Big(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1)),\Big)$$

which gives the result of the upper bound of misclassification rate under the regime $\frac{1}{n} \ll p, q \ll 1$.

For the exact recovery result, consider the logarithmic regime $p = a\log n/n$, $q = b\log n/n$ with $a > b > 0$,

$$\mathbb{P}\Big(\bar{\Delta}_i^{(k)} \geq 0\Big) \leq n^{-(\sqrt{a}-\sqrt{b})^2/K + o(1)}.$$

A union bound over all nodes and incorrect labels shows that if

$$(\sqrt{a}-\sqrt{b})^2 > K,$$

then $\widehat{\sigma}$ achieves exact recovery with high probability.

Combining the above analyses, which separately account for the randomness in $\mathbf{P}$ and $\mathbf{X}$, we obtain the following statements.

1. In the regime $\frac{1}{n} \ll p, q \ll 1$ with $|p-q| = \Theta(p+q)$, if

$$d \gg \frac{2\log n}{\log\bigl(1 + R^*(\mathbf{A})\bigr)}, \quad \text{and} \quad \eta = \omega\Bigl(\frac{1}{np}\Bigr)$$

then the misclassification rate of a single-layer GCN satisfies

$$\mathbb{E}[\mathrm{Err}] \le \exp\Bigl(-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\Bigr),$$

which matches the information-theoretic limit (min–max rate) in [7].

2. In the regime $p, q = \Theta(\log n/n)$ with $|p-q| = \Theta(p+q)$, if

$$\frac{a}{K} + \frac{K-1}{K}b > 3+\alpha, \qquad \eta = \omega\left(\frac{1}{\log n}\right), \qquad d = \omega(n),$$

then a single-layer GCN achieves exact recovery whenever

$$(\sqrt{a}-\sqrt{b})^2 > K,$$

which again coincides with the information-theoretic limit in [4].

This completes the proof.

B Lemmas and proofs

Lemma 1. *In the symmetric K-community stochastic block model with* uniform *community assignment, assume $K = O(1)$ and $p, q = \omega(1/n)$. Let*

$$\mathbf{D}_{ii} = \frac{1}{n}\sum_{u,v=1}^{n} \mathbf{A}_{uv} = \frac{2|\mathcal{E}|}{n}$$

denote the normalized total degree, and let $|\mathcal{V}_k|$ denote the number of nodes in community k. Then there exists an absolute constant $c > 0$ such that, with probability at least

$$1 - 2K\exp\Bigl(-\frac{n}{12K}\Bigr) - 2\exp\Bigl(-c\,\frac{n^2}{(\log n)^2}\cdot\frac{p+(K-1)q}{K}\Bigr),$$

the following statements hold simultaneously:

1. *$|\mathcal{V}_k| = \frac{n}{K}(1+o(1))$, for all $k = 1, \ldots, K$;*
2. *$\mathbf{D}_{ii} = \Bigl(1 + O\Bigl(\frac{1}{\log n}\Bigr)\Bigr)\Bigl[\frac{n}{K}p + n\Bigl(1-\frac{1}{K}\Bigr)q\Bigr]$.*

Proof. Since each node independently chooses a community with probability $1/K$, we have $|\mathcal{V}_k| \sim \mathrm{Bin}(n, 1/K)$. For $\delta = 1/2$, the Chernoff bound yields

$$\mathbb{P}\Bigl(\Bigl||\mathcal{V}_k| - \frac{n}{K}\Bigr| \ge \frac{n}{2K}\Bigr) \le 2\exp\Bigl(-\frac{\delta^2}{3}\cdot\frac{n}{K}\Bigr) = 2\exp\Bigl(-\frac{n}{12K}\Bigr).$$

A union bound over $k = 1, \ldots, K$ gives

$$\mathbb{P}\Bigl(\exists k : \Bigl||\mathcal{V}_k| - \frac{n}{K}\Bigr| \ge \frac{n}{2K}\Bigr) \le 2K\exp\Bigl(-\frac{n}{12K}\Bigr). \tag{8}$$

In particular, on the complement of this event we have $|\mathcal{V}_k| = \frac{n}{K}(1+o(1))$ for all k, proving 1.

Let

$$S := \sum_{1 \le i < j \le n} \mathbf{A}_{ij}, \qquad \text{so that} \qquad \mathbf{D}_{ii} = \frac{2S}{n}.$$

Condition on the community assignment σ. Given σ, the random variables $\{\mathbf{A}_{ij}\}_{i<j}$ are independent Bernoulli with parameters in $\{p, q\}$, hence

$$\mathbb{E}[S \mid \sigma] = p \sum_{k=1}^{K} \binom{|\mathcal{V}_k|}{2} + q\left(\binom{n}{2} - \sum_{k=1}^{K} \binom{|\mathcal{V}_k|}{2}\right), \qquad \mathrm{Var}(S \mid \sigma) = \sum_{i<j} \mathrm{Var}(\mathbf{A}_{ij} \mid \sigma) \le \mathbb{E}[S \mid \sigma].$$

Set

$$\varepsilon_n := \frac{1}{\log n}, \qquad t := \varepsilon_n \, \mathbb{E}[S \mid \sigma].$$

Bernstein's inequality (conditional on σ) yields

$$\mathbb{P}(\,|S - \mathbb{E}[S \mid \sigma]| \ge t \,|\, \sigma) \le 2 \exp\left(-\frac{t^2/2}{\mathrm{Var}(S \mid \sigma) + t/3}\right).$$

Using $\mathrm{Var}(S \mid \sigma) \le \mathbb{E}[S \mid \sigma]$ and $t = \varepsilon_n \mathbb{E}[S \mid \sigma]$, the denominator is at most $\mathbb{E}[S \mid \sigma] + \varepsilon_n \mathbb{E}[S \mid \sigma]/3 \le (4/3)\mathbb{E}[S \mid \sigma]$, and therefore, for some absolute constant $c > 0$,

$$\mathbb{P}\left(|S - \mathbb{E}[S \mid \sigma]| \ge \frac{1}{\log n} \mathbb{E}[S \mid \sigma] \,\middle|\, \sigma\right) \le 2 \exp\left(-c \, \frac{1}{(\log n)^2} \, \mathbb{E}[S \mid \sigma]\right). \tag{9}$$

On the event in (8), we have

$$\sum_{k=1}^{K} \binom{|\mathcal{V}_k|}{2} = \frac{n^2}{2K}(1 + o(1)), \qquad \binom{n}{2} - \sum_{k=1}^{K} \binom{|\mathcal{V}_k|}{2} = \frac{n^2}{2}\Big(1 - \frac{1}{K}\Big)(1 + o(1)),$$

and hence

$$\mathbb{E}[S \mid \sigma] = \frac{n^2}{2K}\big(p + (K-1)q\big)\,(1 + o(1)). \tag{10}$$

Substituting (10) into (9) and adjusting $c > 0$ gives, on the event in (8),

$$\mathbb{P}\left(|S - \mathbb{E}[S \mid \sigma]| \ge \frac{1}{\log n} \, \mathbb{E}[S \mid \sigma]\right) \le 2 \exp\Big(-c \, \frac{n^2}{(\log n)^2} \cdot \frac{p + (K-1)q}{K}\Big).$$

Finally, since $\mathbf{D}_{ii} = 2S/n$, the same relative deviation bound transfers to $\mathbf{D}_{ii}$: with the above probability,

$$\big|\mathbf{D}_{ii} - \mathbb{E}[\mathbf{D}_{ii} \mid \sigma]\big| \le \frac{1}{\log n} \, \mathbb{E}[\mathbf{D}_{ii} \mid \sigma].$$

Moreover, by (10),

$$\mathbb{E}[\mathbf{D}_{ii} \mid \sigma] = \frac{2}{n}\mathbb{E}[S \mid \sigma] = \left[\frac{n}{K}p + n\Big(1 - \frac{1}{K}\Big)q\right](1 + o(1)).$$

Combining the last two displays yields

$$\mathbf{D}_{ii} = \left(1 + O\Big(\frac{1}{\log n}\Big)\right)\left[\frac{n}{K}p + n\Big(1 - \frac{1}{K}\Big)q\right],$$

which proves 2. Taking a union bound with (8) concludes the proof. □

Lemma 2. *Let (Z, V) be jointly Gaussian with mean zero:*

$$\mathbb{E}[Z] = \mathbb{E}[V] = 0, \qquad \mathrm{Var}(Z) = a > 0, \quad \mathrm{Var}(V) = b > 0, \quad \mathrm{Cov}(Z, V) = c,$$

so that the covariance matrix is

$$\Sigma = \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \qquad ab - c^2 > 0.$$

Define $Y := ZV$. Then the moment generating function of Y is

$$M_Y(t) := \mathbb{E}[e^{tY}] = \frac{1}{\sqrt{1 - 2ct - (ab - c^2)t^2}},$$

for all t such that

$$1 - 2ct - (ab - c^2)t^2 > 0,$$

equivalently

$$t \in \left(\frac{-c - \sqrt{ab}}{ab - c^2}, \ \frac{-c + \sqrt{ab}}{ab - c^2} \right).$$

Proof. Let $X := (Z, V)^\top \sim \mathcal{N}(0, \Sigma)$. Note that

$$Y = ZV = X^\top B X, \qquad B := \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}.$$

A standard identity for a centered Gaussian quadratic form states that for any symmetric matrix B,

$$\mathbb{E}\big[e^{tX^\top B X}\big] = \det(I - 2t\Sigma B)^{-1/2},$$

whenever $I - 2t\Sigma B$ is positive definite.

We now compute the determinant explicitly. First,

$$\Sigma B = \begin{pmatrix} a & c \\ c & b \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} \frac{c}{2} & \frac{a}{2} \\ \frac{b}{2} & \frac{c}{2} \end{pmatrix}.$$

Hence

$$I - 2t\Sigma B = \begin{pmatrix} 1 - ct & -at \\ -bt & 1 - ct \end{pmatrix},$$

so

$$\det(I - 2t\Sigma B) = (1 - ct)^2 - abt^2 = 1 - 2ct - (ab - c^2)t^2.$$

Therefore,

$$M_Y(t) = \mathbb{E}[e^{tY}] = \det(I - 2t\Sigma B)^{-1/2} = \frac{1}{\sqrt{1 - 2ct - (ab - c^2)t^2}}.$$

The validity domain is exactly the set of t for which $\det(I - 2t\Sigma B) > 0$ and $I - 2t\Sigma B \succ 0$; in the 2×2 case this reduces to

$$1 - 2ct - (ab - c^2)t^2 > 0,$$

which is equivalent to

$$t \in \left(\frac{-c - \sqrt{ab}}{ab - c^2}, \ \frac{-c + \sqrt{ab}}{ab - c^2} \right).$$

This completes the proof. □

Lemma 3. *For the adjacency matrix $\mathbf{A}$ of a K-symmetric SBM(n, K, p, q), the expectations and variances of $(\mathbf{A}\mathbf{A}^\top)_{ij}$ satisfy, uniformly over all i, j, the following asymptotic formulas, with probability at least $1 - 2Ke^{-n/(4K)}$:*

1. *If $i = j$ and $i \in \mathcal{V}_r$, then*

$$\mathbb{E}[(\mathbf{A}\mathbf{A}^\top)_{ii}] = (1 + o(1))\frac{n}{K}\, p + (1 + o(1))\frac{K - 1}{K} n\, q,$$

$$\mathrm{Var}[(\mathbf{A}\mathbf{A}^\top)_{ii}] = (1 + o(1))\frac{n}{K}\, p(1 - p) + (1 + o(1))\frac{K - 1}{K} n\, q(1 - q).$$

2. If $i \neq j$ and $i, j \in \mathcal{V}_r$, then

$$\mathbb{E}[(\mathbf{A}\mathbf{A}^\top)_{ij}] = (1 + o(1))\frac{n}{K}\,p^2 + (1 + o(1))\frac{K-1}{K}n\,q^2,$$

$$\mathrm{Var}[(\mathbf{A}\mathbf{A}^\top)_{ij}] = (1 + o(1))\frac{n}{K}(p^2 - p^4) + (1 + o(1))\frac{K-1}{K}n(q^2 - q^4).$$

3. If $i \in \mathcal{V}_r$ and $j \in \mathcal{V}_s$ with $r \neq s$, then

$$\mathbb{E}[(\mathbf{A}\mathbf{A}^\top)_{ij}] = (1 + o(1))\frac{2n}{K}\,pq + (1 + o(1))\frac{K-2}{K}n\,q^2,$$

$$\mathrm{Var}[(\mathbf{A}\mathbf{A}^\top)_{ij}] = (1 + o(1))\frac{2n}{K}(pq - p^2q^2) + (1 + o(1))\frac{K-2}{K}n(q^2 - q^4).$$

Proof. Since the graph is undirected and loopless, the adjacency matrix $\mathbf{A}$ is symmetric with $\mathbf{A}_{uu} = 0$ for all u. For any $i, j \in \{1, \dots, n\}$ we can write

$$(\mathbf{A}\mathbf{A}^\top)_{ij} = \sum_{\ell=1}^{n} \mathbf{A}_{i\ell}\mathbf{A}_{j\ell} = \sum_{\ell \neq i,j} \mathbf{A}_{i\ell}\mathbf{A}_{j\ell}, \tag{11}$$

because if $\ell = i$ or $\ell = j$, then one of the factors in the product is $\mathbf{A}_{ii}$ or $\mathbf{A}_{jj}$, which is zero.

According to Lemma 1, with probability at least $1 - 2Ke^{-n/(4K)}$, we have

$$|\mathcal{V}_k| = \frac{n}{K}(1 + o(1)), \text{ for all } k = 1, \dots, K \tag{12}$$

In the following, We analyze the three cases separately.

Case (1): $i = j$. In this case we have

$$(\mathbf{A}\mathbf{A}^\top)_{ii} = \sum_{\ell \neq i} \mathbf{A}_{i\ell},$$

which is exactly the degree of vertex i. Suppose $i \in \mathcal{V}_r$ for some $r \in \{1, \dots, K\}$.

Expectation. We first compute the expectation of a single summand $\mathbf{A}_{i\ell}$. By the SBM construction,

$$\mathbb{E}[\mathbf{A}_{i\ell}] = \begin{cases} p, & \text{if } \ell \in \mathcal{V}_r,\ \ell \neq i, \\ q, & \text{if } \ell \notin \mathcal{V}_r. \end{cases}$$

There are exactly $|\mathcal{V}_r| - 1$ vertices ℓ in the same community as i (excluding i), and $n - |\mathcal{V}_r|$ vertices outside $\mathcal{V}_r$. Therefore,

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] = \sum_{\ell \neq i} \mathbb{E}[\mathbf{A}_{i\ell}] = (|\mathcal{V}_r| - 1)\,p + (n - |\mathcal{V}_r|)\,q.$$

By (12), the following asymptotic expression holds with probability at least $1 - 2Ke^{-n/(4K)}$:

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] = (1 + o(1))\frac{n}{K}\,p + (1 + o(1))\frac{K-1}{K}n\,q.$$

Variance. We next compute the variance. Since $(\mathbf{A}\mathbf{A}^\top)_{ii} = \sum_{\ell \neq i} \mathbf{A}_{i\ell}$, we have

$$\mathrm{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] = \mathrm{Var}\Big(\sum_{\ell \neq i} \mathbf{A}_{i\ell}\Big) = \sum_{\ell \neq i} \mathrm{Var}(\mathbf{A}_{i\ell}) \;+\; 2 \sum_{\ell < \ell',\, \ell, \ell' \neq i} \mathrm{Cov}(\mathbf{A}_{i\ell}, \mathbf{A}_{i\ell'}).$$

By the SBM, all edges are mutually independent. In particular, $\mathbf{A}_{i\ell}$ and $\mathbf{A}_{i\ell'}$ are independent whenever $\ell \neq \ell'$, so all covariance terms are zero. Hence

$$\operatorname{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] = \sum_{\ell \neq i} \operatorname{Var}(\mathbf{A}_{i\ell}).$$

For each Bernoulli variable $\mathbf{A}_{i\ell}$ we have

$$\operatorname{Var}(\mathbf{A}_{i\ell}) = \begin{cases} p(1-p), & \text{if } \ell \in \mathcal{V}_r,\ \ell \neq i, \\ q(1-q), & \text{if } \ell \notin \mathcal{V}_r. \end{cases}$$

Therefore,

$$\begin{aligned} \operatorname{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] &= (|\mathcal{V}_r| - 1)\, p(1-p) + (n - |\mathcal{V}_r|)\, q(1-q) \\ &\overset{\text{w.h.p.}}{=} (1+o(1))\frac{n}{K}\, p(1-p) + (1+o(1))\frac{K-1}{K} n\, q(1-q), \end{aligned}$$

which is the desired variance formula in Case (1).

Case (2): $i \neq j$ and $i, j \in \mathcal{V}_r$. In this case we consider the off-diagonal entry

$$(\mathbf{A}\mathbf{A}^\top)_{ij} = \sum_{\ell \neq i,j} \mathbf{A}_{i\ell}\mathbf{A}_{j\ell}.$$

For convenience, define for each $\ell \neq i, j$ the random variable

$$X_\ell := \mathbf{A}_{i\ell}\mathbf{A}_{j\ell}.$$

Then we can write

$$(\mathbf{A}\mathbf{A}^\top)_{ij} = \sum_{\ell \neq i,j} X_\ell. \tag{13}$$

Note that $X_\ell \in \{0,1\}$, and $X_\ell = 1$ if and only if both edges (i,ℓ) and (j,ℓ) are present.

Expectation. We first compute $\mathbb{E}[X_\ell]$ for different locations of ℓ.

(a) $\ell \in \mathcal{V}_r \setminus \{i,j\}$. Both vertices i and ℓ lie in the same community $\mathcal{V}_r$, and so do j and ℓ. Hence

$$\mathbb{P}(\mathbf{A}_{i\ell} = 1) = p, \qquad \mathbb{P}(\mathbf{A}_{j\ell} = 1) = p.$$

The two edges (i,ℓ) and (j,ℓ) are distinct and, by the model assumption, independent. Therefore,

$$\mathbb{P}\big(X_\ell = 1\big) = \mathbb{P}(\mathbf{A}_{i\ell} = 1, \mathbf{A}_{j\ell} = 1) = \mathbb{P}(\mathbf{A}_{i\ell} = 1)\,\mathbb{P}(\mathbf{A}_{j\ell} = 1) = p^2.$$

Thus, $X_\ell \sim \operatorname{Bern}(p^2)$, which implies $\mathbb{E}[X_\ell] = \mathbb{P}(X_\ell = 1) = p^2$.

(b) $\ell \notin \mathcal{V}_r$. In this case, both edges (i,ℓ) and (j,ℓ) are cross-community edges, so

$$\mathbb{P}(\mathbf{A}_{i\ell} = 1) = q, \qquad \mathbb{P}(\mathbf{A}_{j\ell} = 1) = q.$$

Again, these edges are distinct and independent, hence

$$\mathbb{P}(X_\ell = 1) = \mathbb{P}(\mathbf{A}_{i\ell} = 1)\,\mathbb{P}(\mathbf{A}_{j\ell} = 1) = q^2,$$

and hence $\mathbb{E}[X_\ell] = q^2$.

Collecting these cases, we have

$$\mathbb{E}[X_\ell] = \begin{cases} p^2, & \ell \in \mathcal{V}_r \setminus \{i,j\}, \\ q^2, & \ell \notin \mathcal{V}_r. \end{cases}$$

The number of indices ℓ in $\mathcal{V}_r \setminus \{i,j\}$ is $|\mathcal{V}_r| - 2$, and the number of indices $\ell \notin \mathcal{V}_r$ is $n - |\mathcal{V}_r|$. By applying the linearity of expectation to (13) and using the high-probability event in (12), we obtain

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] = \sum_{\ell \neq i,j} \mathbb{E}[X_\ell] = (|\mathcal{V}_r| - 2)p^2 + (n - |\mathcal{V}_r|)q^2$$
$$\overset{\text{w.h.p.}}{=} (1 + o(1))\frac{n}{K}\, p^2 + (1 + o(1))\frac{K-1}{K} n\, q^2.$$

Variance. We now compute $\mathrm{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big]$. From (13), we have

$$\mathrm{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] = \mathrm{Var}\Big(\sum_{\ell \neq i,j} X_\ell\Big) = \sum_{\ell \neq i,j} \mathrm{Var}(X_\ell) \; + \; 2 \sum_{\ell < \ell',\, \ell,\ell' \neq i,j} \mathrm{Cov}(X_\ell, X_{\ell'}).$$

We first compute $\mathrm{Var}(X_\ell)$ for a fixed ℓ. Since X_ℓ is a Bernoulli random variable, we have

$$\mathrm{Var}(X_\ell) = \mathbb{E}[X_\ell] - \mathbb{E}[X_\ell]^2.$$

From the expectation computation above,

$$\mathrm{Var}(X_\ell) = \begin{cases} p^2 - p^4, & \ell \in \mathcal{V}_r \setminus \{i,j\}, \\ q^2 - q^4, & \ell \notin \mathcal{V}_r. \end{cases}$$

Next we show that the covariance terms vanish. Each X_ℓ is a function only of the two edges $\mathbf{A}_{i\ell}$ and $\mathbf{A}_{j\ell}$. For $\ell \neq \ell'$, the sets of edges $\{\mathbf{A}_{i\ell}, \mathbf{A}_{j\ell}\}$ and $\{\mathbf{A}_{i\ell'}, \mathbf{A}_{j\ell'}\}$ are disjoint, and by the SBM assumption all edges are independent. Hence X_ℓ and $X_{\ell'}$ are independent, and

$$\mathrm{Cov}(X_\ell, X_{\ell'}) = 0 \quad \text{for all } \ell \neq \ell'.$$

Thus

$$\mathrm{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] = \sum_{\ell \neq i,j} \mathrm{Var}(X_\ell) = (|\mathcal{V}_r| - 2)(p^2 - p^4) + (n - |\mathcal{V}_r|)(q^2 - q^4).$$

Substituting $|\mathcal{V}_r| = (1 + o(1))\frac{n}{K}$ and $n - |\mathcal{V}_r| = (1 + o(1))\frac{K-1}{K} n$ yields

$$\mathrm{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] = (1 + o(1))\frac{n}{K}(p^2 - p^4) + (1 + o(1))\frac{K-1}{K} n(q^2 - q^4),$$

as claimed in Case (2).

Case (3): $i \in \mathcal{V}_r$, $j \in \mathcal{V}_s$, $r \neq s$. We again set

$$X_\ell := \mathbf{A}_{i\ell}\mathbf{A}_{j\ell}, \qquad \ell \neq i,j,$$

so that $(\mathbf{A}\mathbf{A}^\top)_{ij} = \sum_{\ell \neq i,j} X_\ell$.

Expectation. We distinguish three types of indices ℓ according to which community they belong to.

(a) $\ell \in \mathcal{V}_r \setminus \{i\}$. Then the edge (i,ℓ) is within-community while the edge (j,ℓ) is cross-community. Thus

$$\mathbb{P}(\mathbf{A}_{i\ell} = 1) = p, \qquad \mathbb{P}(\mathbf{A}_{j\ell} = 1) = q.$$

The two edges are independent, so

$$\mathbb{P}(X_\ell = 1) = \mathbb{P}(\mathbf{A}_{i\ell} = 1, \mathbf{A}_{j\ell} = 1) = \mathbb{P}(\mathbf{A}_{i\ell} = 1)\,\mathbb{P}(\mathbf{A}_{j\ell} = 1) = pq,$$

and therefore $\mathbb{E}[X_\ell] = pq$.

(b) $\ell \in \mathcal{V}_s \setminus \{j\}$. This case is symmetric to (a): the edge (j,ℓ) is within-community and (i,ℓ) is cross-community, so again

$$\mathbb{E}[X_\ell] = pq.$$

(c) $\ell \notin \mathcal{V}_r \cup \mathcal{V}_s$. Now both (i,ℓ) and (j,ℓ) are cross-community edges, hence

$$\mathbb{P}(\mathbf{A}_{i\ell}=1)=q, \qquad \mathbb{P}(\mathbf{A}_{j\ell}=1)=q,$$

and by independence,

$$\mathbb{P}(X_\ell=1)=\mathbb{P}(\mathbf{A}_{i\ell}=1)\,\mathbb{P}(\mathbf{A}_{j\ell}=1)=q^2.$$

Thus $\mathbb{E}[X_\ell]=q^2$ for such ℓ.

Counting the number of indices in each category, we obtain:

- $|\mathcal{V}_r|-1$ indices in $\mathcal{V}_r \setminus \{i\}$,
- $|\mathcal{V}_s|-1$ indices in $\mathcal{V}_s \setminus \{j\}$,
- $n-|\mathcal{V}_r|-|\mathcal{V}_s|$ indices in neither $\mathcal{V}_r$ nor $\mathcal{V}_s$.

Using linearity of expectation,

$$\begin{aligned}\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] &= \sum_{\ell\neq i,j}\mathbb{E}[X_\ell]\\ &= (|\mathcal{V}_r|-1)\,pq+(|\mathcal{V}_s|-1)\,pq+\big(n-|\mathcal{V}_r|-|\mathcal{V}_s|\big)q^2\\ &= (|\mathcal{V}_r|+|\mathcal{V}_s|-2)\,pq+\big(n-|\mathcal{V}_r|-|\mathcal{V}_s|\big)q^2.\end{aligned}$$

According to (12), we obtain

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] \overset{\text{w.h.p.}}{=} (1+o(1))\frac{2n}{K}\,pq+(1+o(1))\frac{K-2}{K}n\,q^2.$$

Variance. As before,

$$\operatorname{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big]=\operatorname{Var}\Big(\sum_{\ell\neq i,j}X_\ell\Big)=\sum_{\ell\neq i,j}\operatorname{Var}(X_\ell)\;+\;2\sum_{\ell<\ell',\,\ell,\ell'\neq i,j}\operatorname{Cov}(X_\ell,X_{\ell'}).$$

For a fixed ℓ, we again use that $X_\ell\in\{0,1\}$, so

$$\operatorname{Var}(X_\ell)=\mathbb{E}[X_\ell]-\mathbb{E}[X_\ell]^2.$$

From the expectation values above, we get

$$\operatorname{Var}(X_\ell)=\begin{cases}pq-p^2q^2, & \ell\in\mathcal{V}_r\setminus\{i\} \text{ or } \ell\in\mathcal{V}_s\setminus\{j\},\\ q^2-q^4, & \ell\notin\mathcal{V}_r\cup\mathcal{V}_s.\end{cases}$$

To handle the covariance terms, observe that each X_ℓ depends only on the edges $\mathbf{A}_{i\ell}$ and $\mathbf{A}_{j\ell}$. For $\ell\neq\ell'$, the pairs of edges $\{\mathbf{A}_{i\ell},\mathbf{A}_{j\ell}\}$ and $\{\mathbf{A}_{i\ell'},\mathbf{A}_{j\ell'}\}$ are disjoint, and by the independence of all edges in the SBM, the corresponding variables X_ℓ and $X_{\ell'}$ are independent. Thus all covariance terms vanish, and we obtain

$$\operatorname{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big]=\sum_{\ell\neq i,j}\operatorname{Var}(X_\ell)=(|\mathcal{V}_r|-1)(pq-p^2q^2)+(|\mathcal{V}_s|-1)(pq-p^2q^2)+\big(n-|\mathcal{V}_r|-|\mathcal{V}_s|\big)(q^2-q^4).$$

Rewriting the first two terms and using the high-probability concentration of community size yields

$$\begin{aligned}\operatorname{Var}\big[(\mathbf{A}\mathbf{A}^\top)_{ij}\big] &= (|\mathcal{V}_r|+|\mathcal{V}_s|-2)(pq-p^2q^2)+\big(n-|\mathcal{V}_r|-|\mathcal{V}_s|\big)(q^2-q^4)\\ &= (1+o(1))\frac{2n}{K}(pq-p^2q^2)+(1+o(1))\frac{K-2}{K}n\,(q^2-q^4),\end{aligned}$$

which is exactly the variance claimed in Case (3).

Combining the results from Cases (1)–(3) completes the proof. □

Lemma 4. *Let $\mathbf{A} \in \{0,1\}^{n\times n}$ denote the adjacency matrix of a random graph generated by a symmetric* SBM(n, K, p, q)*. For each vertex $i \in \mathcal{V}$, let $\sigma(i) \in \{1, \ldots, K\}$ denote its community index, so that $i \in \mathcal{V}_{\sigma(i)}$. For each $k \in \{1, \ldots, K\}$, let $\mathcal{L}_k \subseteq \mathcal{V}_k$ be a (possibly random) subset and define*

$$\bar{\mathbf{S}}_{i,k} := \sum_{\ell \in \mathcal{L}_k} (\mathbf{A}\mathbf{A}^\top)_{i\ell}, \qquad i \in \mathcal{V},\ k \in \{1, \ldots, K\}.$$

Fix $i \in \mathcal{V}$ and $k \in \{1, \ldots, K\}$ and write $r = \sigma(i)$. Conditional on the community partition $(\mathcal{V}_1, \ldots, \mathcal{V}_K)$ and on the event

$$|\mathcal{V}_k| = (1 + o(1))\frac{n}{K} \text{ for all } k, \qquad |\mathcal{L}_k| = (1 + o(1))\,\eta\,\frac{n}{K} \text{ for all } k,$$

the following hold.

1. Expectation. *If $r = k$, then*

$$\mathbb{E}[\bar{\mathbf{S}}_{i,k} \mid \sigma(i) = k] = \big(1 + o(1)\big)\,\eta\,\frac{n^2}{K^2}\big(p^2 + (K-1)q^2\big).$$

   If $r \neq k$, then

$$\mathbb{E}[\bar{\mathbf{S}}_{i,k} \mid \sigma(i) \neq k] = \big(1 + o(1)\big)\,\eta\,\frac{n^2}{K^2}\big(2pq + (K-2)q^2\big).$$

2. Variance. *There exist constants $v_{r,k}$ and $c_{r,k}$ such that*

$$\mathrm{Var}(\bar{\mathbf{S}}_{i,k} \mid \sigma(i) = r) = |\mathcal{L}_k|\, v_{r,k} + |\mathcal{L}_k|\,(|\mathcal{L}_k| - 1)\, c_{r,k} = \big(1 + o(1)\big)\,\eta\,\frac{n}{K}\Big(v_{r,k} + \big(\eta\frac{n}{K} - 1\big)c_{r,k}\Big).$$

   Moreover:

   - *If $r = k$, then*

$$v_{k,k} = \big(1+o(1)\big)\frac{n}{K}\Big[(p^2-p^4)+(K-1)(q^2-q^4)\Big],\ c_{k,k} = \big(1+o(1)\big)\frac{n}{K}\Big[p^3(1-p)+(K-1)q^3(1-q)\Big].$$

   - *If $r \neq k$, then*

$$v_{r,k} = \big(1+o(1)\big)\frac{n}{K}\Big[2(pq-p^2q^2)+(K-2)(q^2-q^4)\Big],\ c_{r,k} = \big(1+o(1)\big)\frac{n}{K}\Big[pq^2(1-p)+qp^2(1-q)+(K-2)q^3(1-q)\Big].$$

3. Concentration of $\bar{\mathbf{S}}_{i,k}$.
   Suppose $a = p\,n/\log n$ and $b = q\,n/\log n$. For any sufficiently small constant $\alpha > 0$, if

$$\frac{a}{K} + \frac{K-1}{K}b > 3 + \alpha, \quad \text{and} \quad \eta = \omega\Big(\frac{1}{\log n}\Big)$$

   then, with probability at least $1 - n^{-1-t/2-t^2/4}$, where $t = \sqrt{\frac{3+\alpha}{3}} - 1$, it holds that

$$\bar{\mathbf{S}}_{i,k} = \Theta\big(\mathbb{E}[\bar{\mathbf{S}}_{i,k}]\big).$$

   Hence, using union bound, $\bar{\mathbf{S}}_{i,k} = \Theta\big(\mathbb{E}[\bar{\mathbf{S}}_{i,k}]\big)$ hold for all nodes with high probability.

   Moreover, for the denser regime $p, q = \omega(\log n/n)$, $\eta = \omega(\frac{1}{np})$ is enough to guarantee the same concentration result holds for all nodes.

Proof. Throughout, all expectations and variances are taken conditional on the community partition $(\mathcal{V}_1, \ldots, \mathcal{V}_K)$ and on the size events $|\mathcal{V}_k| = (1 + o(1))\frac{n}{K}$ and $|\mathcal{L}_k| = (1 + o(1))\eta\frac{n}{K}$ for all k. Fix $i \in \mathcal{V}$ and $k \in \{1, \ldots, K\}$, and write $r = \sigma(i)$. For any vertex $j \in \mathcal{V}_k$ with $j \neq i$, Lemma 3 yields

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ij} \mid \sigma(i) = r, \sigma(j) = k\big] = \begin{cases} (1 + o(1))\frac{n}{K}\big(p^2 + (K-1)q^2\big), & r = k, \\ (1 + o(1))\frac{n}{K}\big(2pq + (K-2)q^2\big), & r \neq k. \end{cases}$$

Since $\bar{\mathbf{S}}_{i,k} = \sum_{\ell \in \mathcal{L}_k} (\mathbf{A}\mathbf{A}^\top)_{i\ell}$ and, by symmetry within $\mathcal{V}_k$, the above conditional mean does not depend on the particular choice of $\ell \in \mathcal{V}_k$ (up to $o(1)$ boundary effects), we obtain

$$\mathbb{E}[\bar{\mathbf{S}}_{i,k} \mid \sigma(i) = r] = |\mathcal{L}_k| \, \mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ij} \mid \sigma(i) = r, \sigma(j) = k\big] \ (1 + o(1)),$$

and substituting $|\mathcal{L}_k| = (1 + o(1))\eta \frac{n}{K}$ gives the stated expectation formulas.

For the variance, write $Y_\ell := (\mathbf{A}\mathbf{A}^\top)_{i\ell}$ for $\ell \in \mathcal{L}_k$, so that $\bar{\mathbf{S}}_{i,k} = \sum_{\ell \in \mathcal{L}_k} Y_\ell$. Then

$$\operatorname{Var}(\bar{\mathbf{S}}_{i,k} \mid \sigma(i) = r) = \sum_{\ell \in \mathcal{L}_k} \operatorname{Var}(Y_\ell \mid \sigma(i) = r) + 2 \sum_{\substack{\ell, m \in \mathcal{L}_k \\ \ell < m}} \operatorname{Cov}(Y_\ell, Y_m \mid \sigma(i) = r).$$

By exchangeability of vertices within $\mathcal{V}_k$, $\operatorname{Var}(Y_\ell \mid \sigma(i) = r)$ is the same for all $\ell \in \mathcal{L}_k$, and $\operatorname{Cov}(Y_\ell, Y_m \mid \sigma(i) = r)$ is the same for all distinct unordered pairs $\{\ell, m\} \subseteq \mathcal{L}_k$. Hence there exist $v_{r,k}$ and $c_{r,k}$ such that

$$\operatorname{Var}(Y_\ell \mid \sigma(i) = r) = v_{r,k}, \qquad \operatorname{Cov}(Y_\ell, Y_m \mid \sigma(i) = r) = c_{r,k} \quad (\ell \neq m),$$

which implies

$$\operatorname{Var}(\bar{\mathbf{S}}_{i,k} \mid \sigma(i) = r) = |\mathcal{L}_k| \, v_{r,k} + |\mathcal{L}_k|(|\mathcal{L}_k| - 1) \, c_{r,k}.$$

The single-entry variance $v_{r,k}$ is exactly the conditional variance of $(\mathbf{A}\mathbf{A}^\top)_{ij}$ given $\sigma(i) = r, \sigma(j) = k$, and therefore coincides with the expressions stated above (as in Lemma 3). To identify $c_{r,k}$, take two distinct vertices $v, w \in \mathcal{V}_k$ and set $Y_v = (\mathbf{A}\mathbf{A}^\top)_{iv}$ and $Y_w = (\mathbf{A}\mathbf{A}^\top)_{iw}$. Writing $Y_v = \sum_{t \in \mathcal{V} \setminus \{i,v\}} \mathbf{A}_{it}\mathbf{A}_{vt}$ and $Y_w = \sum_{t \in \mathcal{V} \setminus \{i,w\}} \mathbf{A}_{it}\mathbf{A}_{wt}$, one checks that all cross-terms with different summation indices are independent and thus have zero covariance, so only the common-index contributions remain. For $t \in \mathcal{V}_u$, define $\rho_{rs} := p\delta_{rs} + q(1 - \delta_{rs})$, where δ_{rs} is the Kronecker delta. Using independence of the three edges $\mathbf{A}_{it}$, $\mathbf{A}_{vt}$, and $\mathbf{A}_{wt}$, we obtain

$$\operatorname{Cov}(\mathbf{A}_{it}\mathbf{A}_{vt}, \ \mathbf{A}_{it}\mathbf{A}_{wt} \mid t \in \mathcal{V}_u) = \rho_{ru}\rho_{ku}^2(1 - \rho_{ru}),$$

and summing over $u \in \{1, \dots, K\}$ and over $t \in \mathcal{V}_u$ yields

$$\operatorname{Cov}(Y_v, Y_w \mid \sigma(i) = r) = (1 + o(1)) \frac{n}{K} \sum_{u=1}^{K} \rho_{ru}\rho_{ku}^2(1 - \rho_{ru}).$$

Evaluating the last sum gives $c_{k,k} = (1 + o(1)) \frac{n}{K} \big[p^3(1 - p) + (K - 1)q^3(1 - q)\big]$ when $r = k$, and $c_{r,k} = (1 + o(1)) \frac{n}{K} \big[pq^2(1 - p) + qp^2(1 - q) + (K - 2)q^3(1 - q)\big]$ when $r \neq k$. Substituting $v_{r,k}$ and $c_{r,k}$ into the variance decomposition completes the proof of (1) and (2).

Next, we prove the statement (3). Under the assumption $p, q = \Omega(\log n / n)$, we have

$$\mathbb{E}[\bar{\mathbf{S}}_{i,k}] = \Theta\big(\eta \log^2 n\big)$$

uniformly over all i and k. For later reference, we denote

$$\mu_{i,k} := \mathbb{E}[\bar{\mathbf{S}}_{i,k}] = \Theta(\eta \log^2 n).$$

To analyze concentration, observe that for each $v \in \mathcal{V}_k$,

$$(\mathbf{A}\mathbf{A}^\top)_{iv} = \sum_{u=1}^{n} \mathbf{A}_{iu}\mathbf{A}_{vu}.$$

Consequently,

$$\bar{\mathbf{S}}_{i,k} = \sum_{u \in N(i)} \sum_{v \in \mathcal{L}_k} \mathbf{A}_{vu},$$

where

$$N(i) := \{u \in \mathcal{V} : \mathbf{A}_{iu} = 1\}$$

denotes the (random) neighborhood of vertex i. Conditioning on $N(i)$, define the index set

$$\mathcal{I}(i, k) := \{(v, u) : v \in \mathcal{V}_k, \ u \in N(i)\}, \qquad Z_{v,u} := \mathbf{A}_{vu}.$$

Then

$$\bar{\mathbf{S}}_{i,k} = \sum_{(v,u)\in\mathcal{I}(i,k)} Z_{v,u},$$

where $\{Z_{v,u}\}_{(v,u)\in\mathcal{I}(i,k)}$ are mutually independent Bernoulli random variables.

Under the assumption $p, q \geq c_0 \log n/n$, we have

$$\mathbb{P}(Z_{v,u} = 1) = \rho_{\sigma(v)\sigma(u)} \geq c_0 \frac{\log n}{n}, \qquad (v,u) \in \mathcal{I}(i,k).$$

Therefore, conditioning on $N(i)$,

$$\mu_{i,k}(N(i)) := \mathbb{E}[\bar{\mathbf{S}}_{i,k} \mid N(i)] = \sum_{(v,u)\in\mathcal{I}(i,k)} \mathbb{P}(Z_{v,u} = 1) \geq \eta\, |\mathcal{V}_k|\, |N(i)|\, c_0\, \frac{\log n}{n} = \frac{c_0\, \eta}{K} |N(i)| \log n.$$

By Lemma 5,

$$\mathbb{P}\big(N(i) = \Theta(\mathbb{E}[N(i)]) = \Omega(\log n)\big) \geq 1 - n^{-1-t/2-t^2/4}, \tag{14}$$

which implies $\mu_{i,k}(N(i)) = \Omega(\eta \log^2 n)$ on this event. Conditioned on $N(i) = \Theta(\mathbb{E}[N(i)])$, the random variable $\bar{\mathbf{S}}_{i,k}$ is thus a sum of $\Omega(n \log n)$ independent Bernoulli variables with expectation of order $\eta \log^2 n$.

Applying the Chernoff bound, there exist constants $C, c > 0$ such that for any $0 < \varepsilon < 1$,

$$\mathbb{P}\big(\big|\bar{\mathbf{S}}_{i,k} - \mathbb{E}[\bar{\mathbf{S}}_{i,k}]\big| \geq \varepsilon\, \mathbb{E}[\bar{\mathbf{S}}_{i,k}] \mid N(i)\big) \leq C \exp\big(-c\, \varepsilon^2 \eta \log^2 n\big) + n^{-1-t/2-t^2/4}.$$

Note that $\eta = \omega\left(\frac{1}{\log n}\right)$. By choosing

$$\varepsilon = \frac{c_1}{\sqrt{\eta \log n}} = o(1),$$

for a sufficiently large constant c_1, we obtain

$$\mathbb{P}\big(\big|\bar{\mathbf{S}}_{i,k} - \mathbb{E}[\bar{\mathbf{S}}_{i,k}]\big| \geq o(1)\, \mathbb{E}[\bar{\mathbf{S}}_{i,k}] \mid N(i)\big) \leq n^{-1-t/2-t^2/4}.$$

Consequently, with probability at least $1 - n^{-1-t/2-t^2/4}$,

$$\bar{\mathbf{S}}_{i,k} = \big(1 + o(1)\big)\, \mathbb{E}\big[\bar{\mathbf{S}}_{i,k} \mid N(i)\big] \overset{(i)}{=} \Theta\big(\mathbb{E}[\bar{\mathbf{S}}_{i,k}]\big),$$

where (i) holds for (14).

For the denser regime where $p, q = \omega(\log n/n)$ and $\eta = \omega(1)$, choosing $\varepsilon = \frac{c_1}{\sqrt{\eta n p}}$ yields the same concentration result, without imposing any assumptions on a and b.

This completes the proof of (3). □

Lemma 5. *Suppose $a = p\,n/\log n$ and $b = q\,n/\log n$. For any sufficiently small constant $\alpha > 0$, if*

$$\frac{a}{K} + \frac{K-1}{K} b > 3 + \alpha,$$

then

$$(\mathbf{A}\mathbf{A}^\top)_{ii} = \Theta\big(\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big]\big) \tag{15}$$

holds with probability at least $1 - n^{-1-t/2-t^2/4}$, where $t = \sqrt{\frac{3+\alpha}{3}} - 1$.

Proof. Since $p, q = \Omega(\log n/n)$, it follows from Lemma 3 that

$$\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big] = (1 + o(1)) \left(\frac{a}{K} + \frac{K-1}{K} b\right) \log n.$$

Under the stated condition, this expectation is bounded below by $(3 + \alpha) \log n$ for all sufficiently large n.

Let

$$\varepsilon = \frac{1}{2}\left(1 + \sqrt{\frac{3}{3+\alpha}}\right),$$

which satisfies $0 < \varepsilon < 1$. Applying the Chernoff bound yields

$$\mathbb{P}\big(\big|(\mathbf{A}\mathbf{A}^\top)_{ii} - \mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big]\big| \geq \varepsilon\,\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big]\big) \leq 2\exp\left(-\frac{\varepsilon^2(3+\alpha)}{3}\log n\right).$$

A direct calculation shows that the exponent can be rewritten as

$$-\frac{\varepsilon^2(3+\alpha)}{3}\log n = -\left(1 + \frac{t}{2} + \frac{t^2}{4}\right)\log n,$$

where $t = \sqrt{\frac{3+\alpha}{3}} - 1$. Therefore,

$$\mathbb{P}\big(\big|(\mathbf{A}\mathbf{A}^\top)_{ii} - \mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big]\big| \geq \varepsilon\,\mathbb{E}\big[(\mathbf{A}\mathbf{A}^\top)_{ii}\big]\big) \leq 2n^{-1-t/2-t^2/4}.$$

This completes the proof. □

References

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[2] W. L. Hamilton, *Graph representation learning.* Morgan & Claypool Publishers, 2020.

[3] L. Cheng, P. Zhu, Y. Guo, K. Tang, C. Gao, and Z. Wang, "Hyperdet: Source detection in hypergraphs via interactive relationship construction and feature-rich attention fusion," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, J. Kwok, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2025, pp. 2758–2766, main Track.

[4] E. Abbe, "Community detection and stochastic block models: recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[5] Y. Deshpande, S. Sen, A. Montanari, and E. Mossel, "Contextual stochastic block models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on information theory*, vol. 62, no. 1, pp. 471–487, 2015.

[7] A. Y. Zhang and H. H. Zhou, "Minimax rates of community detection in stochastic block models," *The Annals of Statistics*, vol. 44, no. 5, pp. 2252–2280, 2016.

[8] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block models," *Journal of Machine Learning Research*, vol. 18, no. 60, pp. 1–45, 2017.

[9] R. J. Wang, A. Baranwal, and K. Fountoulakis, "Analysis of corrected graph convolutions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 128 015–128 052, 2024.

[10] A. Baranwal, K. Fountoulakis, and A. Jagannath, "Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization," in *International Conference on Machine Learning.* PMLR, 2021, pp. 684–693.

[11] X. Wu, Z. Chen, W. W. Wang, and A. Jadbabaie, "A non-asymptotic analysis of oversmoothing in graph neural networks," in *The Eleventh International Conference on Learning Representations.*

[12] R. Wei, H. YIN, J. Jia, A. R. Benson, and P. Li, "Understanding non-linearity in graph neural networks from the bayesian-inference perspective," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 34 024–34 038.

[13] C. Shi, L. Pan, H. Hu, and I. Dokmanić, "Homophily modulates double descent generalization in graph convolution networks," *Proceedings of the National Academy of Sciences*, vol. 121, no. 8, p. e2309504121, 2024.

[14] K. Fountoulakis, A. Levi, S. Yang, A. Baranwal, and A. Jagannath, "Graph attention retrospective," *Journal of Machine Learning Research*, vol. 24, no. 246, pp. 1–52, 2023.

[15] Z. Ma, Q. Zhang, B. Zhou, Y. Zhang, S. Hu, and Z. Wang, "Graph attention is not always beneficial: A theoretical analysis of graph attention mechanisms via contextual stochastic block models," in *Forty-second International Conference on Machine Learning.*

[16] H. Wang and Z. Wang, "Optimal exact recovery in semi-supervised learning: A study of spectral methods and graph convolutional networks," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 21–27 Jul 2024, pp. 51 614–51 649.

[17] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.

[18] J. Gaudio and N. Joshi, "Exact community recovery under side information: Optimality of spectral algorithms," in *The Thirteenth International Conference on Learning Representations*, 2025.

[19] E. Mossel, J. Neeman, and A. Sly, "Reconstruction and estimation in the planted partition model," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 431–461, 2015.

[20] J. Lei and A. Rinaldo, "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, pp. 215–237, 2015.

[21] A. Y. Zhang, "Fundamental limits of spectral clustering in stochastic block models," *IEEE Transactions on Information Theory*, 2024.

[22] Q. Zhang and V. Y. Tan, "Exact recovery in the general hypergraph stochastic block model," *IEEE Transactions on Information Theory*, vol. 69, no. 1, pp. 453–471, 2022.

[23] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

[24] A. A. AMINI and E. LEVINA, "On semidefinite relaxations for the block model," *The Annals of Statistics*, vol. 46, no. 1, pp. 149–179, 2018.

[25] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 84, no. 6, p. 066106, 2011.

[26] E. Mossel, J. Neeman, and A. Sly, "Belief propagation, robust reconstruction and optimal recovery of block models," in *Conference on Learning Theory.* PMLR, 2014, pp. 356–370.

[27] V. Jog and P.-L. Loh, "Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence," *arXiv preprint arXiv:1509.06418*, 2015.