GRAPH CONVOLUTIONAL NETWORKS ACHIEVE INFORMATION-THEORETIC LIMITS UNDER STOCHASTIC BLOCK MODELS WITH NON-INFORMATIVE NODE FEATURES

Anonymous authors
Paper under double-blind review

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 p and q denote the within- and between-community edge probabilities. An η fraction of community labels is revealed, while the node features are i.i.d. standard Gaussian vectors in $\mathbb{R}^d$ independent of the labels. Despite the absence of feature signal, we show that a single-layer one-step gradient GCN estimator based on regularized least squares can exploit graph structure when d is sufficiently large. In the logarithmic-degree regime $p, q = \Theta(\log n/n)$ with $p > q$, the estimator achieves exact recovery when $(\sqrt{p} - \sqrt{q})^2 > K \log n/n$, using only a vanishing proportion $\eta = o(1)$ of labeled nodes. A matching information-theoretic lower bound shows that this threshold is sharp. In the broader regime $1/n \ll p, q \ll 1$, under a sufficient condition on d and again with only a vanishing proportion of labeled nodes, its expected misclassification rate is at most $\exp\big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2(1 + o(1))\big)$. We further establish an information-theoretic lower bound on the expected misclassification rate when the labeled proportion vanishes. In the common regime, this lower bound matches the preceding upper bound and shows that our GCN estimator attains the optimal expected misclassification rate. We also extend the analysis to multi-layer GCNs and reveal a tradeoff between stronger feature concentration and reduced separation between communities under repeated propagation, which helps explain oversmoothing. We conduct numerical experiments on both single-layer and multi-layer GCNs that corroborate our theoretical findings.

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 Kipf & Welling (2017); Hamilton (2020); Cheng et al. (2025). 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 Abbe (2018). 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 rely 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 Deshpande et al. (2018). Such node features are *informative* in the sense that their distributions differ across communities and therefore carry explicit information about the underlying labels.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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*.

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. Our results focus on a single-layer one-step gradient GCN estimator obtained from the first negative-gradient direction of the regularized least-squares loss at the zero initialization. We also extend the analysis to multi-layer GCNs. We study two fundamental performance metrics, namely *exact recovery*, which requires correctly classifying all nodes with high probability, and *misclassification rate*, which quantifies the fraction of incorrectly classified nodes. These two metrics are widely used in unsupervised node classification (i.e., community detection). Their information-theoretic limits are by now well understood Abbe et al. (2015); Zhang & Zhou (2016); Gao et al. (2017). However, it remains unclear **(i)** how these information-theoretic limits change in the semi-supervised setting and **(ii)** whether modern GCN-based methods can attain them and, if so, under what conditions.

Our main contributions are summarized below.

• We establish a *sharp threshold* for exact recovery in semi-supervised node classification using our GCN estimator in the logarithmic-degree regime (i.e., $p, q = \Theta(\log n/n)$). When the feature dimension d is sufficiently large (i.e., $d = \omega(n)$), 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 4.1 and 4.2). This demonstrates that our GCN estimator attains the information-theoretic limit for exact recovery.

• We derive an upper bound on the expected misclassification rate of our GCN estimator in the broader regime $1/n \ll p, q \ll 1$. Under appropriate scaling of the feature dimension d and a vanishing labeled proportion η, we show that its expected misclassification rate is at most $\exp\big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2(1 + o(1))\big)$ (see Theorem 4.3).

• Theorem 4.4 gives the first semi-supervised lower bound showing that, with a vanishing proportion of revealed labels, the optimal expected misclassification rate has the same leading exponential order as in unsupervised node classification under the SBM. This conclusion holds for all estimators when $\eta = o(1)$. Under the condition on d in Theorem 4.3 and in the common regime $1/(np) \ll \eta \ll 1$, Theorems 4.3 and 4.4 give matching upper and lower bounds. Together, these results establish the optimal expected misclassification rate for semi-supervised node classification and prove that our GCN estimator attains it.

• We extend our analysis to multi-layer GCNs. Theorem 4.5 gives an explicit upper bound on the expected misclassification rate and a sufficient condition for exact recovery, which clearly describes two competing effects of increasing depth. Repeated propagation strengthens feature concentration by averaging feature noise, while it also reduces the separation between communities. This tradeoff provides a theoretical explanation of oversmoothing.

• Single-layer experiments support our exact-recovery and misclassification analyses and show that our estimator and trained linearized GCNs exhibit similar performance and dependence on d under the SBM. Multi-layer experiments support the depth tradeoff and oversmoothing behavior predicted by our analysis.

1.1 Detailed Comparison with Prior Work

Relation to CSBM-Based Analyses of GCNs. Most existing theoretical analyses of GCNs are conducted under the contextual SBM (CSBM), where node features are informative and aligned with community labels Deshpande et al. (2018); Wang et al. (2024). 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 Baranwal et al. (2021); Wu et al. (2023); Wei et al. (2022); Shi et al. (2024); Fountoulakis et al. (2023); Ma et al. (2025), as well as exact recovery thresholds in the semi-supervised CSBM Wang & Wang (2024).

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 special case 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. A more detailed comparison with existing works that analyze GCNs under the SBM/CSBM framework is provided in Appendix A (Table 1).

Relation to Community Detection and Information-Theoretic Limits. Semi-supervised node classification on the SBM is closely related to the classical (unsupervised) community detection. When the labeled proportion $\eta = 0$, it reduces exactly to community detection; more generally, it can be interpreted as community detection with side information in the form of partially observed node labels Saad & Nosratinia (2018); Gaudio & Joshi (2025).

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 Abbe et al. (2015); Mossel et al. (2015); Abbe (2018); Zhang & Zhou (2016); Gao et al. (2017). In the logarithmic-degree regime $p, q = \Theta(\log n/n)$, exact recovery is possible if and only if $(\sqrt{p} - \sqrt{q})^2 > K \log n/n$ Abbe (2018);

in the sparse regime, under the corresponding signal conditions, the optimal expected misclassification rate scales as $\exp\big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2(1 + o(1))\big)$ Zhang & Zhou (2016).

A variety of polynomial-time algorithms are known to achieve these limits under suitable conditions, including spectral methods Lei & Rinaldo (2015); Zhang (2024); Zhang & Tan (2022), semidefinite relaxations Hajek et al. (2016); AMINI & LEVINA (2018), and belief propagation Decelle et al. (2011); Mossel et al. (2014).

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

K-Symmetric Stochastic Block Model. We consider an undirected random graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = [n]$ and adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$. The latent labels σ_i are drawn independently and uniformly from $[K]$, forming $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$. Thus, the communities $\mathcal{V}_k := \{i : \sigma_i = k\}$ have approximately equal sizes, allowing small random fluctuations rather than exact balance to better reflect practical settings.

Conditional on $\boldsymbol{\sigma}$, edges are generated independently as $\mathbb{P}(\mathbf{A}_{ij} = 1 \mid \boldsymbol{\sigma}) = \mathbf{B}_{\sigma_i \sigma_j}, \quad i < j$, with $\mathbf{A}_{ji} = \mathbf{A}_{ij}$ and $\mathbf{A}_{ii} = 0$. Here, $\mathbf{B} \in [0, 1]^{K \times K}$ satisfies $\mathbf{B}_{kk} = p$ and $\mathbf{B}_{kl} = q$ for $k \neq l$. We denote this model by $\mathrm{SBM}(n, K, p, q)$.

Semi-Supervised Node Classification. We consider semi-supervised learning with labeled set $\mathcal{L} \subseteq \mathcal{V}$ and unlabeled set $\mathcal{U} := \mathcal{V} \setminus \mathcal{L}$. Under the K-symmetric SBM, for a given $\eta \in (0, 1)$, each $\mathcal{L}_k := \mathcal{L} \cap \mathcal{V}_k$ is a uniformly random subset of $\mathcal{V}_k$ of size $\lfloor \eta |\mathcal{V}_k| \rfloor$. Conditional on $\boldsymbol{\sigma}$, these subsets are sampled independently across classes and independently of $\mathbf{A}$.

An estimator is any measurable mapping $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}\big(\mathbf{A}, \mathbf{X}, \mathcal{L}, \{\sigma_i : i \in \mathcal{L}\}\big) \in [K]^n$, with *misclassification rate* defined as

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

In the regimes considered for our GCN estimator, every community contains a revealed label with probability $1 - o(1)$, eliminating label permutation ambiguity.

We say that an estimator an estimator achieves *exact recovery* if $\mathbb{P}(\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}) \to 1$ as $n \to \infty$.

3 Graph Convolutional Networks as Estimators

Node Feature Generation. Node feature generation is required, as GCNs are designed to operate on graphs equipped with node features. Importantly, we consider non-informative node features generated independently of $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A})$.

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

Linearized Graph Convolutional Networks. Let $\mathbf{D}$ be a diagonal matrix with $\mathbf{D}_{ii} = \frac{1}{n}\sum_{u,v=1}^{n} \mathbf{A}_{uv}$, [1] and define the message-passing operator $\mathbf{P} := \mathbf{D}^{-1}\mathbf{A}$. If the common diagonal entry of $\mathbf{D}$ is zero, we set $\mathbf{P} = \mathbf{0}$.

An L-layer GCN alternates graph propagation, trainable linear transformations, and nonlinear activations. We study its linearized version obtained by removing the activations and treating the layer weights as a single effective readout matrix $\mathbf{W} \in \mathbb{R}^{d \times K}$. The corresponding class-score matrix is $\mathbf{P}^L\mathbf{X}\mathbf{W}$. To separate graph propagation from the linear readout, we define the propagated feature matrix $\mathbf{Z}^{(L)} := \mathbf{P}^L\mathbf{X} \in \mathbb{R}^{n \times d}$.

Thus, the class-score matrix is $\mathbf{Z}^{(L)}\mathbf{W}$. In the analysis below, $\mathbf{P}^L$ is fixed and the gradient is taken with respect to the effective readout matrix $\mathbf{W}$. We write $\mathbf{z}_i^{(L)}$ for the ith row of $\mathbf{Z}^{(L)}$. For $L = 1$, we write $\mathbf{Z} := \mathbf{Z}^{(1)} = \mathbf{P}\mathbf{X}$.

One-Step Gradient GCN Estimator. 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 $\mathbf{Z}_{\mathcal{L}}^{(L)}$ and $\mathbf{Y}_{\mathcal{L}}$ denote the rows of $\mathbf{Z}^{(L)}$ and $\mathbf{Y}$ indexed by $\mathcal{L}$. For $L \geq 1$ and $\lambda > 0$, consider the regularized least-squares objective for the effective linear readout $\mathbf{W}$:

$$\mathcal{J}_\lambda^{(L)}(\mathbf{W}) := \frac{1}{2}\left\|\mathbf{Z}_{\mathcal{L}}^{(L)}\mathbf{W} - \mathbf{Y}_{\mathcal{L}}\right\|_F^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2. \tag{2}$$

One gradient step from zero with step size $s > 0$ gives

$$\mathbf{W}_1 = -s\nabla\mathcal{J}_\lambda^{(L)}(\mathbf{0}) = s\big(\mathbf{Z}_{\mathcal{L}}^{(L)}\big)^\top\mathbf{Y}_{\mathcal{L}}.$$

The regularization gradient vanishes at zero, and s only rescales all scores by a positive constant, so we set $s = 1$.

Define the Gram matrix $\mathbf{G}^{(L)} := \mathbf{Z}^{(L)}(\mathbf{Z}^{(L)})^\top$. The resulting linearized GCN score matrix is

$$\widehat{\mathbf{S}}^{(L)} := \mathbf{Z}^{(L)}\mathbf{W}_1 = \mathbf{G}_{:\mathcal{L}}^{(L)}\mathbf{Y}_{\mathcal{L}}. \tag{3}$$

Since $\mathbf{Y}$ is one-hot,

$$\widehat{\mathbf{S}}_{ik}^{(L)} = \sum_{j \in \mathcal{L}_k}\left\langle \mathbf{z}_i^{(L)}, \mathbf{z}_j^{(L)} \right\rangle,$$

so each score sums the similarities to labeled embeddings from that class.

The *one-step gradient GCN estimator* $\widehat{\boldsymbol{\sigma}}^{(L)}$ predicts

$$\widehat{\sigma}_i^{(L)} \in \arg\max_{k \in [K]} \widehat{\mathbf{S}}_{ik}^{(L)}, \qquad i \in \mathcal{U}, \tag{4}$$

with ties broken by a fixed deterministic rule, and sets $\widehat{\sigma}_i^{(L)} = \sigma_i$ for $i \in \mathcal{L}$.

For the single-layer case, we suppress the superscript and write $\mathbf{G} := \mathbf{G}^{(1)}, \widehat{\mathbf{S}} := \widehat{\mathbf{S}}^{(1)}, \widehat{\boldsymbol{\sigma}} := \widehat{\boldsymbol{\sigma}}^{(1)}$.

[1] 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 Wang & Wang (2024).

Remark 3.1. Related methods, such as graph ridge regression of Wang & Wang (2024), use the exact minimizer of a regularized least-squares objective. We instead use one gradient step from zero to isolate the graph signal in our non-informative feature setting. This estimator already attains the information-theoretic limits studied here, making exact minimization unnecessary for our purpose.

4 Main Results

We first analyze the sharp performance of our GCN estimator. In the logarithmic-degree regime, we show that it attains the information-theoretic threshold for exact recovery. In a broader regime, we establish the optimal expected misclassification rate by combining the upper bound for this estimator with a matching information-theoretic lower bound for the same semi-supervised setting. The lower bound also shows that, with an $o(1)$ proportion of revealed labels, this rate has the same leading exponential order as in unsupervised node classification under the SBM. We then extend the analysis to multi-layer GCNs.

4.1 Sharp Exact Recovery Threshold in the Logarithmic-Degree Regime

We first characterize the conditions under which the estimator achieves exact recovery (Theorem 4.1), and then establish a matching information-theoretic lower bound (Theorem 4.2).

Theorem 4.1 (Exact recovery by the single-layer one-step gradient GCN estimator)**.** *Consider a symmetric* $\mathrm{SBM}(n, K, p, q)$ *with* $K = O(1)$ *almost equal-sized communities, where* $a > b > 0$ *are fixed constants and* $p = \frac{a \log n}{n}$, $q = \frac{b \log n}{n}$, $|p - q| = \Theta(p + q)$. *Assume that* $\eta = \omega\Big(\frac{1}{\log n}\Big)$, $d = \omega(n)$.

Then the one-step gradient GCN estimator defined in Section 3 with $L = 1$ *achieves exact recovery with probability at least* $1 - o(1)$ *if*

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

Proof. See Appendix C.1 for the detailed proof. □

Theorem 4.2 (Information-theoretic lower bound for semi-supervised exact recovery)**.** *Consider the semi-supervised symmetric SBM described in Section 2, where* $K \geq 2$ *is fixed, the community labels are i.i.d. uniform on* $[K]$*, and* $p = \frac{a \log n}{n}$, $q = \frac{b \log n}{n}$*, where* $a > b > 0$ *are fixed constants. Allow* $\eta \in (0, 1)$ *to be a deterministic sequence. Conditional on* $\boldsymbol{\sigma}$*, the sets* $\mathcal{L}_k$ *are sampled independently across* k*, with each set uniform among the subsets of* $\mathcal{V}_k$ *with size* $|\mathcal{L}_k| = \lfloor \eta |\mathcal{V}_k| \rfloor$*. Define* $\beta_n := \frac{\log\big(1/(1-\eta)\big)}{\log n}$*. If* $\limsup_{n\to\infty} \Big(\frac{(\sqrt{a}-\sqrt{b})^2}{K} + \beta_n \Big) < 1$*, then*

$$\sup_{\widehat{\boldsymbol{\sigma}}} \mathbb{P}(\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}) \longrightarrow 0,$$

where the supremum is over all measurable, possibly randomized, estimators based on $(\mathbf{A}, \mathbf{X}, \mathcal{L}, \{\sigma_i : i \in \mathcal{L}\})$.

Proof. See Appendix C.2 for the detailed proof. □

Theorems 4.1 and 4.2 together provide a refined understanding of exact recovery under the stochastic block model in the semi-supervised setting. The following discussion highlights several key points.

(1). Theorem 4.1 shows that our GCN estimator 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 $\eta = 1/\sqrt{\log n}$.

(2). Theorem 4.2 establishes an information-theoretic lower bound for semi-supervised exact recovery. When $\frac{1}{\log n} \ll \eta \ll 1$, we have $\beta_n = o(1)$. In this regime, the information-theoretic lower bound reduces to $(\sqrt{a} - \sqrt{b})^2 < K$. Theorem 4.1 establishes exact recovery under the reverse strict inequality. Therefore, our GCN estimator attains the information-theoretic limit for exact recovery.

(3). The requirement $d = \omega(n)$ highlights the role of high-dimensional random features in enabling our GCN estimator 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 estimator effectively reduces the problem to community detection under the SBM at the information-theoretic scale.

4.2 Optimal Misclassification Rate in the Broader Regime

We next establish matching upper and lower bounds on the expected misclassification rate in a broader regime. Together, they identify the optimal expected misclassification rate. The GCN upper bound is given in Theorem 4.3, and the matching information-theoretic lower bound is given in Theorem 4.4. Before formally stating the theorem, we first define a quantity $R^*(\mathbf{A})$, which will appear in the condition on the feature dimension d, as follows,

$$R^*(\mathbf{A}) := \min_{\substack{i\in[n],\\ k\in[K]\setminus\{r\},\\ \gamma(i,k,r)<0}} \frac{\gamma^2(i,k,r)}{\alpha(i)\beta(k,r) - \gamma^2(i,k,r)},$$

where $r := \sigma_i$ is the true label of node i. Thus, $R^*(\mathbf{A})$ is the uniform minimum over graph-good node–class pairs.

The quantities $\alpha(i)$, $\beta(k,r)$, and $\gamma(i,k,r)$ are defined as

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

For convenience, define $\Lambda_n := \max\{\log n,\ \frac{n}{K}(\sqrt{p} - \sqrt{q})^2\}$.

Theorem 4.3 (Upper bound on the expected misclassification rate of the single-layer estimator)**.** *Consider a symmetric* $\mathrm{SBM}(n, K, p, q)$ *with* $K = O(1)$ *almost equal-sized communities, where* $\frac{1}{n} \ll p, q \ll 1, \quad |p - q| = \Theta(p + q)$. *Assume that* $\eta = \omega\Big(\frac{1}{np}\Big)$. *Suppose further that, for some deterministic sequence* $\zeta_n \to \infty$, $d \geq \frac{2\zeta_n \Lambda_n}{\log\big(1+R^*(\mathbf{A})\big)}$ *holds with probability at least* $1 - \exp[-\omega(\Lambda_n)]$ *over* $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A})$.

Then the expected misclassification rate of the estimator $\hat{\boldsymbol{\sigma}} := \widehat{\boldsymbol{\sigma}}^{(1)}$ *defined in Section 3 satisfies*

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

Proof. See Appendix C.1 for the detailed proof. □

Theorem 4.4 (Information-theoretic lower bound for the expected misclassification rate)**.** *Consider the semi-supervised symmetric SBM described in Section 2, with non-informative features* $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *independent of* $(\mathbf{A}, \boldsymbol{\sigma})$. *Assume that* $K \geq 2, K = O(1)$, $\frac{1}{n} \ll p, q \ll 1, |p-q| = \Theta(p+q)$, *and that* $\eta = o(1)$. *The infimum and expectation below are over all measurable estimators and all randomness in* $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A}, \mathbf{X})$. *Then*

$$\inf_{\hat{\boldsymbol{\sigma}}} \mathbb{E}_{\boldsymbol{\sigma},\mathcal{L},\mathbf{A},\mathbf{X}}\left[\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma})\right] \geq \exp\Big\{-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2(1 + o(1))\Big\}\,.$$

This lower bound has the same leading exponential order as the classical minimax rate for unsupervised community detection Zhang & Zhou (2016).

Proof. See Appendix C.3 for the detailed proof. □

We provide three comments on these results.

(1). The lower bound in Theorem 4.4 has the same leading exponential order as the classical minimax rate for unsupervised node classification under the SBM Zhang & Zhou (2016). Thus, revealing a vanishing proportion of labels does not change the optimal misclassification rate at this order. To the

best of our knowledge, this is the first such semi-supervised lower bound. Under the condition on d in Theorem 4.3 and in the common regime $1/(np) \ll \eta \ll 1$, combining the two theorems shows that the optimal expected misclassification rate is $\exp\{-\frac{n}{K}(\sqrt{p}-\sqrt{q})^2(1+o(1))\}$. The proposed GCN estimator attains this rate and is therefore information-theoretically optimal.

(2). The additional condition $\eta \gg 1/(np)$ comes from the upper bound. It is compatible with a vanishing labeled proportion. For example, $\eta = 1/\sqrt{np}$ satisfies both conditions since $np = \omega(1)$.

(3). Because $R^*(\mathbf{A})$ is the minimum over all nodes and wrong classes, deriving a deterministic bound on d requires uniform control of all graph margins and variance terms, which is difficult in this broader regime. The present condition requires $d\log(1+R^*(\mathbf{A}))$ to dominate Λ_n. A larger $R^*(\mathbf{A})$ represents a stronger graph margin relative to the feature variance and allows a smaller feature dimension. The condition also shows that d only needs to be large enough to make the feature noise negligible at the scale of the graph error. It is used only in the analysis and is not required by the estimator.

4.3 Extension to Multi-Layer GCNs

We now extend the propagation analysis to $L \geq 1$ layers, using the depth-L score $\widehat{\mathbf{S}}^{(L)}$ and prediction rule from Section 3.

Define the average-degree scale and the normalized community contrast by

$$\delta_n := \frac{n}{K}\big(p+(K-1)q\big), \qquad \gamma_n := \frac{|p-q|}{p+(K-1)q}.$$

We also define

$$B_{n,L} := \eta n\delta_n^L + \eta^2 n\delta_n^{2L},\ \widetilde{\mathcal{V}}_{n,L} := \min_{1\leq a\leq 2L}\left\{\frac{B_{n,L}}{\delta_n^{2L-a}}, \left(\frac{B_{n,L}}{\delta_n^{2L-a}}\right)^{1/a}\right\},\ \mathcal{V}_{n,L} := \min\left\{\delta_n^{1/2}, \widetilde{\mathcal{V}}_{n,L}\right\},$$

and $\mathcal{W}_{n,L} := \eta\delta_n\gamma_n^{4L}$, $\quad \mathcal{F}_{n,L} := \frac{d\eta\delta_n^{2L}\gamma_n^{4L}}{n(1+\eta\delta_n^L)}$.

Theorem 4.5 (Performance bound for the multi-layer one-step gradient GCN estimator)**.** *Consider the symmetric K-community SBM described above with $2 \leq K = O(1)$. Let $0 < \eta < 1$ and let d be a positive integer, both possibly depending on n. Fix an integer $L \geq 1$ that does not depend on n. Assume that $\frac{1}{n} \ll p, q \ll n^{-1+1/(2L)}$, and that there exists a constant $c_{\mathrm{pq}} \in (0,1)$ such that $c_{\mathrm{pq}} \leq \frac{p}{q} \leq c_{\mathrm{pq}}^{-1}$. Suppose further that $\frac{n|p-q|^{2L}}{\left(p+(K-1)q\right)^{2L-1}} \to \infty$, $\quad \frac{\eta n|p-q|^{2L}}{\left(p+(K-1)q\right)^{2L}} \to \infty$. Then there exist constants $C, c > 0$, depending only on K, L, and c_{pq}, such that*

$$\mathbb{E}\left[\mathrm{err}\left(\widehat{\boldsymbol{\sigma}}^{(L)}, \boldsymbol{\sigma}\right)\right] \leq C\exp\{-c\mathcal{W}_{n,L}\} + C\exp\{-c\mathcal{V}_{n,L}\} + C\exp\{-c\mathcal{F}_{n,L}\}.$$

In particular, if $\mathcal{W}_{n,L} \gg \log n$, $\qquad \mathcal{V}_{n,L} \gg \log n$, $\qquad \mathcal{F}_{n,L} \gg \log n$, then

$$\mathbb{P}\left(\widehat{\boldsymbol{\sigma}}^{(L)} = \boldsymbol{\sigma}\right) \to 1.$$

Proof. See Appendix C.4 for the detailed proof. □

Consequently, the theorem gives the following non-sharp lower bound on the negative logarithm of the expected misclassification rate

$$-\log\mathbb{E}\left[\mathrm{err}\left(\widehat{\boldsymbol{\sigma}}^{(L)}, \boldsymbol{\sigma}\right)\right] \geq c\min\left\{\mathcal{W}_{n,L}, \mathcal{V}_{n,L}, \mathcal{F}_{n,L}\right\} - \log(3C).$$

Theorem 4.5 describes how propagation depth affects the three exponent scales in the multi-layer bound. We provide two comments below.

(1). Increasing the depth L has two competing effects on the bound. The graph-margin exponent $\mathcal{W}_{n,L}$ contracts by a factor $\gamma_n^4 < 1$ per layer, since repeated propagation reduces the separation between communities. The feature-noise exponent $\mathcal{F}_{n,L}$ instead grows by roughly $\delta_n\gamma_n^4$ per layer when $\eta\delta_n^L \gg 1$. The two cross when $d\delta_n^{L-1} \asymp \eta n$: below this threshold additional depth mainly averages out feature noise, while above it the graph term becomes the bottleneck and further depth

erodes community separation — a concrete mechanism for oversmoothing (see Appendix C.4 for the derivation and the graph-variance constraint $\mathcal{V}_{n,L}$).

(2). Theorem 4.5 does not recover the sharp single-layer results of Theorems 4.1 and 4.3 at $L = 1$. The single-layer proof exploits the length-two structure of $\mathbf{A}\mathbf{A}^\top$ to obtain an exact conditional MGF; for $L > 1$ this edgewise factorization breaks down, and the general concentration inequality we use instead yields only the conservative constraint $\mathcal{V}_{n,L} \le \delta_n^{1/2}$ (at $L = 1$, $p, q = \Theta(\log n/n)$, this is $O(\sqrt{\log n})$, well short of the sharp threshold). Recovering the $L = 1$ results therefore requires the dedicated length-two argument used in Theorems 4.1 and 4.3 (full multi-layer argument in Appendix C.4).

Practical interpretation. Taken together, the above results offer three practical insights. First, even when node features contain no information about the community labels, we show that our GCN estimator can use graph topology for accurate node classification. In contrast, an MLP that ignores the graph cannot classify the unlabeled nodes better than random guessing. Second, our analysis shows that the feature dimension d is a key factor in the classification performance of the GCN estimator under this setting. Although increasing d does not add label information, it reduces the effect of feature randomness and allows the Gram matrix of the propagated features to better reflect graph structure. Third, the multi-layer result suggests a tradeoff in depth. Shallow propagation can improve feature averaging, whereas excessive propagation reduces the separation between communities and may lead to oversmoothing.

5 Experiments

We conduct six experiments, summarized in Figures 1 and 2. All experiments use $K = 4$, independent standard Gaussian features, and $n = 2000$ (Figure 2(b) additionally uses $n = 10000$). Throughout, GCN* denotes the one-step gradient GCN estimator of Section 3.

Exact recovery experiments. Panels (a)–(b) evaluate exact recovery over 100 trials each, with $\eta = 2/\log n$, $p = a\log n/n$, and $q = b\log n/n$ for $a, b \in \{2, 6, \ldots, 62\}$. With $d = n\log n$ (panel (a), satisfying $d = \omega(n)$), we observe a sharp phase transition near the information-theoretic threshold (red curve), verifying Theorem 4.1. With the smaller $d = n$ (panel (b)), exact recovery fails even at the threshold, indicating that $d = \omega(n)$ is essentially tight.

Misclassification rate experiments. Panels (c)–(d) report the misclassification rate with $\eta = 0.2$, $p = a\sqrt{\log n}/n$, and $q = 4\sqrt{\log n}/n$ (only a varies). Panel (c) shows the negative log misclassification rate approaching the exponent in Theorem 4.3 as d grows. Panel (d) trains a single-layer linearized GCN on the objective in (2) via Adam (learning rate 10^{-2}, 100 epochs) and compares it (blue) with GCN* (black) under $d = n\log n$ and $d = \sqrt{n}$: the trained GCN is slightly worse but shows the same dependence on d, supporting GCN* as a proxy for trained GCNs.

Multi-layer experiments. Panels (a)–(b) of Figure 2 examine how depth L affects GCN* on a symmetric SBM with $K = 4$. Panel (a) reports, over 30 trials, the exact-recovery fraction for $n = 2000$, $\eta = 2/\log n$, $d = n\log n$, $p = a\log n/n$, $q = b\log n/n$ ($a \in \{10, \ldots, 42\}$, $b \in \{2, \ldots, 14\}$), and $L \in \{1, 3, 4, 6\}$: as L increases, exact recovery becomes harder to attain near the information-theoretic threshold $(\sqrt{a} - \sqrt{b})^2 = K$ (red curve), even for pairs above it. Panel (b) reports, over 20 trials, the misclassification rate for $a = 14$, $b = 4$, $\eta = 0.2$, $p = a\sqrt{\log n}/n$, $q = b\sqrt{\log n}/n$, $n \in \{2000, 10000\}$, $d \in \{n\log n, \sqrt{n}\}$, and $L \in \{1, 2, 3, 4, 5, 6, 8, 10, 12, 16\}$: on these sparser graphs, a small depth (e.g., $L = 2$ or 3, depending on n, d) reduces the misclassification rate, but larger L increases it rapidly. This oversmoothing pattern is consistent with the tradeoff identified after Theorem 4.5, where neighborhood averaging initially reduces feature noise while repeated propagation reduces community separation.

6 Conclusion

We provided a theoretical characterization of linear GCNs in the non-informative feature regime under the symmetric SBM. Despite completely non-informative node features, our GCN estimator can leverage graph structure together with a vanishing fraction of labeled nodes to achieve exact recovery at the SBM information-theoretic threshold. In a broader regime, an information-theoretic lower bound shows that, with an $o(1)$ proportion of revealed labels, the optimal expected misclassification

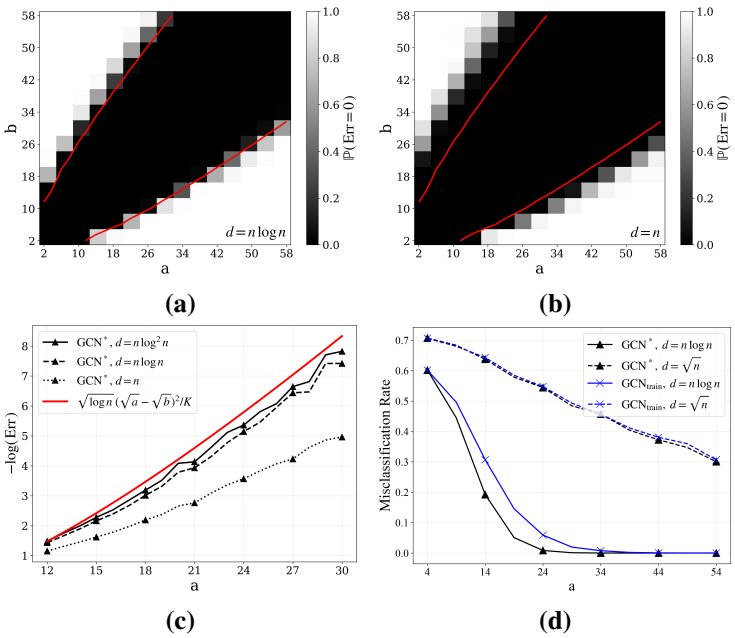


Figure 1: Single-layer experimental results ($n = 2000$, $K = 4$; GCN* uses $L = 1$). (a)–(b) Exact recovery (100 trials/point); $d = n \log n$ in (a), $d = n$ in (b). (c)–(d) Misclassification rate (10 trials/point); (d) also shows $\text{GCN}_{\text{train}}$, a trained single-layer linearized GCN.

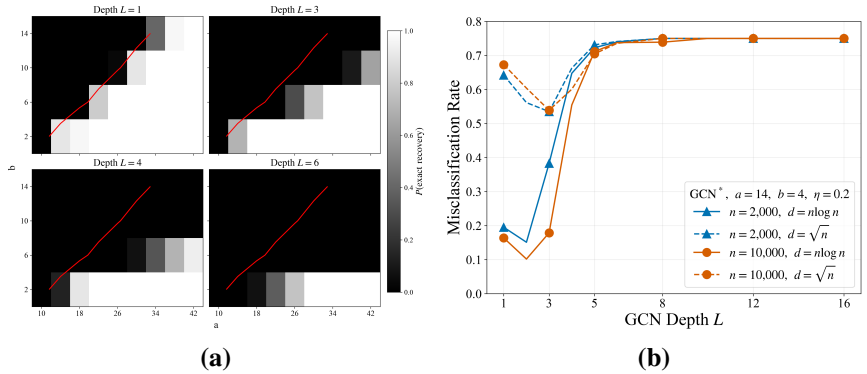


Figure 2: Multi-layer experimental results for GCN*. (a) Exact-recovery fraction (30 trials/point) for $n = 2000$, $d = n \log n$, $\eta = 2/\log n$, $L \in \{1, 3, 4, 6\}$; red curve marks the threshold $(\sqrt{a} - \sqrt{b})^2 = K$. (b) Misclassification rate (20 trials/point) versus L for $a = 14$, $b = 4$, $\eta = 0.2$, $n \in \{2000, 10000\}$, $d \in \{n \log n, \sqrt{n}\}$.

rate has the same leading exponential order as in unsupervised node classification under the SBM. In the common regime of the upper and lower bounds, the estimator attains this optimal rate. We also extended the analysis to multi-layer GCNs. The multi-layer result shows how feature averaging competes with the reduction in separation between communities under repeated propagation.

Several directions remain open. First, it is of interest to characterize the optimal choice of depth across broader graph regimes. Second, extending the analysis to richer message-passing operators and to SBM extensions (e.g., degree heterogeneity, unbalanced community sizes or overlapping communities) may yield a broader understanding of when GNN-style semi-supervised methods are statistically optimal.

REFERENCES

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

ARASH A AMINI and ELIZAVETA LEVINA. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 684–693. PMLR, 2021.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Effects of graph convolutions in multi-layer networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=P-73JPgRs0R.

Le Cheng, Peican Zhu, Yangming Guo, Keke Tang, Chao Gao, and Zhen Wang. Hyperdet: Source detection in hypergraphs via interactive relationship construction and feature-rich attention fusion. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 2758–2766. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/307. Main Track.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(6):066106, 2011.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. *Advances in Neural Information Processing Systems*, 31, 2018.

O. Duranthon and L. Zdeborová. Statistical physics analysis of graph neural networks: Approaching optimality in the contextual stochastic block model. *Phys. Rev. X*, 15:041026, Nov 2025. doi: 10.1103/lfxj-hbsk.

Kimon Fountoulakis, Amit Levi, Shenghao Yang, Aseem Baranwal, and Aukosh Jagannath. Graph attention retrospective. *Journal of Machine Learning Research*, 24(246):1–52, 2023.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(60): 1–45, 2017.

Julia Gaudio and Nirmit Joshi. Exact community recovery under side information: Optimality of spectral algorithms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, pp. 215–237, 2015.

Wei Lu. Learning guarantees for graph convolutional networks on the stochastic block model. In *International Conference on Learning Representations*, 2022.

Zhongtian Ma, Qiaosheng Zhang, Bocheng Zhou, Yexin Zhang, Shuyue Hu, and Zhen 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*, 2025.

Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pp. 356–370. PMLR, 2014.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

Hussein Saad and Aria Nosratinia. Community detection with side information: Exact recovery under the stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5): 944–958, 2018.

Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pp. 437–446, USA, 2012. Society for Industrial and Applied Mathematics.

Cheng Shi, Liming Pan, Hong Hu, and Ivan Dokmanić. Homophily modulates double descent generalization in graph convolution networks. *Proceedings of the National Academy of Sciences*, 121(8):e2309504121, 2024.

Haixiao Wang and Zhichao 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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51614–51649. PMLR, 21–27 Jul 2024.

Robert J Wang, Aseem Baranwal, and Kimon Fountoulakis. Analysis of corrected graph convolutions. *Advances in Neural Information Processing Systems*, 37:128015–128052, 2024.

Rongzhe Wei, Haoteng YIN, Junteng Jia, Austin R Benson, and Pan Li. Understanding non-linearity in graph neural networks from the bayesian-inference perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34024–34038. Curran Associates, Inc., 2022.

Xinyi Wu, Zhengdao Chen, William Wei Wang, and Ali Jadbabaie. A non-asymptotic analysis of oversmoothing in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016. ISSN 00905364.

Anderson Ye Zhang. Fundamental limits of spectral clustering in stochastic block models. *IEEE Transactions on Information Theory*, 2024.

Qiaosheng Zhang and Vincent YF Tan. Exact recovery in the general hypergraph stochastic block model. *IEEE Transactions on Information Theory*, 69(1):453–471, 2022.

APPENDIX

The appendix is organized as follows. Appendix A provides a detailed comparison with prior work on GCNs under the SBM/CSBM framework. Appendix B summarizes the notations used throughout the paper. Appendix C provides detailed proofs of Theorems 4.1, 4.2, 4.3, 4.4, and 4.5. Appendix D presents the auxiliary lemmas used in these proofs.

A DETAILED COMPARISON WITH PRIOR WORK

Work	Graph sparsity	Support for non-informative features	No. of communities	Arbitrary fixed-depth	Metric
Shi et al. (2024)	$p, q = \Theta(1/n)$	✗	$K = 2$	✗	misclassification rate (asymptotic value)
Duranthon & Zdeborová (2025)	$p, q = \Theta(1)$ (dense graph)	✓	$K = 2$	✓	misclassification rate (asymptotic value)
Wang & Wang (2024)	$p, q = \omega(1/n)$	✗	$K = 2$	✗	exact recovery
Lu (2022)	$p, q = \Theta(\log^3 n/n)$	✓	$K = 2$	✗	almost exact recovery
Baranwal et al. (2021)	$p, q = \Omega(\log^2 n/n)$	✗	$K = 2$	✗	exact recovery
Fountoulakis et al. (2023)	$p, q = \Omega(\log^2 n/n)$	✗	$K = 2$	✗	exact recovery
Ma et al. (2025)	$p, q = \Omega(\log^2 n/n)$	✗	$K = 2$	✓	exact recovery
Baranwal et al. (2023)	$p + q = \Theta(\log^3 n/n)$	✗	$K = 2$	✗	exact recovery
Ours	$p, q = \Theta(\log n/n)$; $1/n \ll p, q \ll 1$	✓	$K = O(1)$	✓	sharp exact recovery threshold; optimal misclassification rate

Table 1: Comparison of assumptions and objectives in theoretical analyses of GCNs under SBM/CSBM-type models. The single-layer estimator attains the sharp exact-recovery threshold and the optimal misclassification rate.

B NOTATION TABLE

Table 2: Summary of notations used in the paper.

Symbol	Meaning
n	Number of nodes.
K	Number of communities/classes.
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Undirected graph.
$\mathcal{V}$	Node set, $\mathcal{V} = [n]$.
$\mathcal{E}$	Edge set.
$\mathbf{A} \in \{0,1\}^{n\times n}$	Adjacency matrix; symmetric; $\mathbf{A}_{ii} = 0$.
$\boldsymbol{\sigma}$	Label vector $(\sigma_1, \ldots, \sigma_n)$.
$\hat{\boldsymbol{\sigma}}$	Estimator/predictor producing labels in $[K]^n$.
$\mathcal{V}_k$	Nodes in class k: $\{i : \sigma_i = k\}$.
$\mathbf{B} \in [0,1]^{K\times K}$	SBM connectivity matrix.
p	Within-community edge probability.
q	Cross-community edge probability.
$\mathcal{L}$	Labeled node set.
$\mathcal{U}$	Unlabeled node set, $\mathcal{U} = \mathcal{V} \setminus \mathcal{L}$.
η	Labeled proportion (per class).
$\mathcal{L}_k$	Labeled nodes in class k: $\mathcal{L} \cap \mathcal{V}_k$.
$\mathbf{X} \in \mathbb{R}^{n\times d}$	Feature matrix.
$\boldsymbol{x}_i \in \mathbb{R}^d$	Feature vector of node i.
d	Feature dimension.

Continued on next page.

Symbol	Meaning
$\mathbf{I}_d$	$d \times d$ identity matrix.
$\mathbf{D}$	Diagonal normalization matrix; $\mathbf{D}_{ii} = \frac{1}{n}\sum_{u,v} \mathbf{A}_{uv}$.
$\mathbf{P}$	Propagation operator, $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$.
L	Number of propagation layers.
$\mathbf{Z}^{(L)}$	Depth-L linearized embedding, $\mathbf{Z}^{(L)} = \mathbf{P}^L\mathbf{X}$.
$\mathbf{Z}$	Single-layer embedding, $\mathbf{Z} = \mathbf{P}\mathbf{X}$.
$\mathbf{Y} \in \{0,1\}^{n\times K}$	One-hot label matrix, $\mathbf{Y}_{ik} = \mathbb{1}\{\sigma_i = k\}$.
$\mathbf{Z}_{\mathcal{L}}^{(L)}$	Rows of $\mathbf{Z}^{(L)}$ restricted to $\mathcal{L}$.
$\mathbf{Y}_{\mathcal{L}}$	Rows of $\mathbf{Y}$ restricted to $\mathcal{L}$.
$\mathbf{W}$	Effective linear readout matrix, $\mathbf{W} \in \mathbb{R}^{d\times K}$.
λ	Regularization parameter in the regularized least-squares objective.
$\mathbf{G}^{(L)}$	Depth-L Gram matrix, $\mathbf{G}^{(L)} = \mathbf{Z}^{(L)}(\mathbf{Z}^{(L)})^\top$.
$\widehat{\mathbf{S}}^{(L)}$	Depth-L score matrix, $\widehat{\mathbf{S}}^{(L)} = \mathbf{G}^{(L)}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}}$.
$\widehat{\boldsymbol{\sigma}}^{(L)}$	Predicted labels obtained from $\widehat{\mathbf{S}}^{(L)}$.
$\mathbf{G}$	Single-layer Gram matrix, $\mathbf{G} = \mathbf{G}^{(1)} = \mathbf{Z}\mathbf{Z}^\top$.
$\widehat{\mathbf{S}}$	Single-layer score matrix, $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}^{(1)}$.
$\widehat{\boldsymbol{\sigma}}$	Single-layer prediction vector, $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}^{(1)}$.
$\widehat{\mathbf{S}}_{ik}$	Single-layer score of node i for class k.
$\Delta_i^{(k)}$	Score gap $\widehat{\mathbf{S}}_{ik} - \widehat{\mathbf{S}}_{ir}$ where $r = \sigma_i$.
$\Delta_{i,k}^{(L)}$	Depth-L score gap $\widehat{\mathbf{S}}_{i,k}^{(L)} - \widehat{\mathbf{S}}_{i,r}^{(L)}$ where $r = \sigma_i$.
δ_n	Average-degree scale $\frac{n}{K}(p + (K-1)q)$.
γ_n	Normalized community contrast $\lvert p - q\rvert/(p + (K-1)q)$.
$\mathcal{W}_{n,L}$	Graph-margin exponent term in the multi-layer bound.
$\mathcal{V}_{n,L}$	Graph-induced variance exponent term in the multi-layer bound.
$\mathcal{F}_{n,L}$	Feature-noise exponent term in the multi-layer bound.

C Proofs of the Main Results

C.1 Proofs of Theorems 4.1 and 4.3

Under the single-layer convention from Section 3, recall that the score matrix of the one-step gradient GCN estimator is

$$\widehat{\mathbf{S}} = \widehat{\mathbf{S}}^{(1)} = \mathbf{G}_{:\mathcal{L}}\mathbf{Y}_{\mathcal{L}}.$$

Thus, for any unlabeled node i, the predicted label is given by

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

For each labeled node, we set $\widehat{\sigma}_i = \sigma_i$, so its error probability is zero. 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,

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

Local error event analysis of score matrix $\widehat{\mathbf{S}}$. We begin with a local analysis of $\widehat{\mathbf{S}}$ by examining its (i,k)-th entry $\widehat{\mathbf{S}}_{ik}$. For any $k \neq \sigma(i)$, let $\Delta_i^{(k)} := \widehat{\mathbf{S}}_{ik} - \widehat{\mathbf{S}}_{i\sigma(i)}$. If node i is misclassified, then $\Delta_i^{(k)} \geq 0$ for at least one $k \neq \sigma(i)$. Hence

$$\{\widehat{\sigma}_i \neq \sigma(i)\} \subseteq \bigcup_{k\neq\sigma(i)} \{\Delta_i^{(k)} \geq 0\}.$$

The two independent sources of randomness in $\Delta_i^{(k)}$ are 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 control $\Delta_i^{(k)}$ conditional on the graph. Second, we analyze the resulting graph-only score margin.

Lemma D.1 records the basic concentration of the community sizes and the global average degree. In particular, with high probability,

$$|\mathcal{V}_k| = \frac{n}{K}\left(1 + O\left(\frac{1}{\log n}\right)\right) \quad (k \in [K]),$$

and the common diagonal entry of $\mathbf{D}$ has its stated deterministic order. For the score reduction, however, no conditioning on this event is needed. Define the realized global average degree

$$D_0 := \mathbf{D}_{11} = \frac{1}{n}\sum_{u,v=1}^{n} \mathbf{A}_{uv}.$$

By definition, $\mathbf{D} = D_0\mathbf{I}$. Hence, on the event $D_0 > 0$,

$$\mathbf{P}\mathbf{P}^\top = D_0^{-2}\mathbf{A}\mathbf{A}^\top$$

holds exactly. Moreover, since $p, q = \omega(1/n)$ in both regimes considered below,

$$\mathbb{P}(D_0 = 0) \le \exp\left[-\binom{n}{2}\min\{p, q\}\right] = \exp[-\omega(n)].$$

We may define the estimator arbitrarily on this negligible event. Thus the realized factor D_0^{-2} is a common positive score scaling and does not affect any class comparison. The stronger community-size event required for the graph large-deviation exponent is introduced separately below.

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, on $D_0 > 0$, we obtain

$$\widehat{\mathbf{S}}_{ik} = \sum_{\ell \in \mathcal{L}_k} (\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top)_{i\ell} = \frac{1}{D_0^2}\sum_{\ell \in \mathcal{L}_k}\sum_{r=1}^{d} (\mathbf{A}\boldsymbol{x}^{(r)})_i\,(\mathbf{A}\boldsymbol{x}^{(r)})_\ell.$$

Recalling the definition of $\Delta_i^{(k)}$ from the local error event above, we obtain

$$\Delta_i^{(k)} = \widehat{\mathbf{S}}_{ik} - \widehat{\mathbf{S}}_{i\sigma(i)} = \frac{1}{D_0^2}\sum_{r=1}^{d} (\mathbf{A}\boldsymbol{x}^{(r)})_i \left(\sum_{\ell \in \mathcal{L}_k} (\mathbf{A}\boldsymbol{x}^{(r)})_\ell - \sum_{\ell \in \mathcal{L}_{\sigma(i)}} (\mathbf{A}\boldsymbol{x}^{(r)})_\ell\right). \tag{6}$$

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 $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A})$, both Z and V are mean-zero Gaussian random variables. Throughout this conditional feature analysis, conditioning on $\mathbf{A}$ abbreviates conditioning on $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A})$, so that only $\mathbf{X}$ remains random. 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.

Since $D_0^2 > 0$, the pairwise event $\Delta_i^{(k)} \geq 0$ is equivalent to

$$\widetilde{\Delta}_i^{(k)} := D_0^2 \Delta_i^{(k)} = \sum_{r=1}^{d} Z_r V_r \geq 0.$$

Recall that

$$\gamma(i,k,\sigma(i)) = \bar{\Delta}_i^{(k)} := \sum_{\ell \in \mathcal{L}_k} (\mathbf{A}\mathbf{A}^\top)_{i\ell} - \sum_{\ell \in \mathcal{L}_{\sigma(i)}} (\mathbf{A}\mathbf{A}^\top)_{i\ell}.$$

We first work on the graph-good event

$$\mathcal{G}_{i,k} := \{\gamma(i,k,\sigma(i)) < 0\}.$$

For $t > 0$ and $\mathbf{A} \in \mathcal{G}_{i,k}$, Markov's inequality gives

$$\mathbb{P}_{\mathbf{X}}\Big(\widetilde{\Delta}_i^{(k)} \geq 0 \mid \mathbf{A}\Big) \leq \mathbb{E}_{\mathbf{X}}\Big[\exp\Big(t\widetilde{\Delta}_i^{(k)}\Big) \mid \mathbf{A}\Big].$$

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

$$M_t\Big(\widetilde{\Delta}_i^{(k)} \mid \mathbf{A}\Big) = \mathbb{E}\Big[\exp\Big(t\widetilde{\Delta}_i^{(k)}\Big) \mid \mathbf{A}\Big] = \prod_{r=1}^{d} \mathbb{E}[\exp(tZ_rV_r) \mid \mathbf{A}] = \frac{1}{(1 - 2t\gamma - t^2(\alpha\beta - \gamma^2))^{d/2}}.$$

On $\mathcal{G}_{i,k}$, the optimal Chernoff parameter is positive and is given by

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

Substituting this value yields

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

Here we have written the non-degenerate case $\alpha\beta - \gamma^2 > 0$. If $\alpha\beta - \gamma^2 = 0$ and $\gamma < 0$, then necessarily $\alpha > 0$ and

$$\mathrm{Var}\Big(V_r - \frac{\gamma}{\alpha}Z_r \;\Big|\; \mathbf{A}\Big) = \beta - \frac{\gamma^2}{\alpha} = 0.$$

Therefore,

$$V_r = \frac{\gamma}{\alpha}Z_r \quad \text{almost surely,} \qquad Z_rV_r = \frac{\gamma}{\alpha}Z_r^2 \leq 0 \quad \text{almost surely.}$$

Since $\alpha > 0$, we have $\sum_{r=1}^d Z_r^2 > 0$ almost surely. It follows that $\widetilde{\Delta}_i^{(k)} = \sum_{r=1}^d Z_rV_r < 0$ almost surely, and hence

$$\mathbb{P}_{\mathbf{X}}\Big(\Delta_i^{(k)} \geq 0 \mid \mathbf{A}\Big) = 0.$$

Equivalently, in the definition below we set $R_i(k,\sigma(i);\mathbf{A}) = +\infty$ in this case.

Let

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

We use the conventions $x/0 := +\infty$ for $x > 0$ and $0/0 := 0$ for the pairwise quantity. Only graph-good pairs enter $R^*(\mathbf{A})$. Thus, on $\mathcal{G}_{i,k}$,

$$\mathbb{P}_{\mathbf{X}}\Big(\Delta_i^{(k)} \geq 0 \mid \mathbf{A}\Big) \leq \exp\Big[-\frac{d}{2}\log(1 + R_i(k,\sigma(i);\mathbf{A}))\Big].$$

For the broader-regime theorem, recall the following notation from the main text. In terms of the pairwise quantity above,

$$R^*(\mathbf{A}) = \min_{\substack{i\in[n],\ k\neq\sigma(i)\\ \gamma(i,k,\sigma(i))<0}} R_i(k,\sigma(i);\mathbf{A}), \qquad \Lambda_n = \max\Big\{\log n,\ \frac{n}{K}(\sqrt{p}-\sqrt{q})^2\Big\},$$

If no node–class pair satisfies the condition in the minimum defining $R^*(\mathbf{A})$, we regard $R^*(\mathbf{A})$ as infinite. Also recall that $\zeta_n \to \infty$ is the deterministic sequence appearing in that theorem. Whenever

$$d \geq \frac{2\zeta_n\Lambda_n}{\log\big(1+R^*(\mathbf{A})\big)},$$

the following bound holds uniformly on the graph-good events $\mathcal{G}_{i,k}$.

$$\mathbb{P}_{\mathbf{X}}\Big(\Delta_i^{(k)} \geq 0 \mid \mathbf{A}\Big) \leq \exp(-\zeta_n\Lambda_n).$$

The theorem assumes that the displayed dimension inequality fails with probability at most $\exp[-\omega(\Lambda_n)]$. Therefore, the unconditional feature-induced contribution for every pair satisfies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma},\mathcal{L},\mathbf{A}}\Big[\mathbf{1}_{\mathcal{G}_{i,k}}\mathbb{P}_{\mathbf{X}}\Big(\Delta_i^{(k)} \geq 0 \mid \mathbf{A}\Big)\Big]\\ \leq \exp(-\zeta_n\Lambda_n) + \exp[-\omega(\Lambda_n)] = \exp[-\omega(\Lambda_n)]\,.\end{aligned}$$

The exact-recovery specialization under $d = \omega(n)$ is established after the graph-only MGF bound below by a shifted-margin argument. Here this means splitting the graph margin at a small negative threshold rather than at zero. Weakly negative margins are included in the graph-tail event, and the resulting negative buffer on its complement is used to control the feature error uniformly.

Consider the randomness of P. Fix a node i with true label $\sigma(i) = r$, and consider any competing label $k \neq r$. Recall from the conditional feature analysis above that the graph-only score margin is

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

The graph-bad event is $\{\bar{\Delta}_i^{(k)} \geq 0\}$, and the feature analysis above gives

$$\mathbb{P}(\widehat{\sigma}(i) = k) \ \leq\ \mathbb{P}\Big(\bar{\Delta}_i^{(k)} \geq 0\Big) + \exp[-\omega(\Lambda_n)]\,.$$

It is enough first to treat $i \notin \mathcal{L}$. Indeed, if $i \in \mathcal{L}_r$, then

$$\bar{\Delta}_i^{(k)} = \left[\sum_{\ell\in\mathcal{L}_k}(\mathbf{A}\mathbf{A}^\top)_{i\ell} - \sum_{\ell\in\mathcal{L}_r\setminus\{i\}}(\mathbf{A}\mathbf{A}^\top)_{i\ell}\right] - (\mathbf{A}\mathbf{A}^\top)_{ii}.$$

The last term equals $-\deg(i) \leq 0$, so the upper-tail event is contained in the corresponding event with i removed from $\mathcal{L}_r$. The loss of one labeled node is absorbed by the estimates below.

Assume henceforth that $i \notin \mathcal{L}$ and define

$$s_u := \mathbf{1}_{\{u\in\mathcal{L}_k\}} - \mathbf{1}_{\{u\in\mathcal{L}_r\}}, \qquad B_v := \mathbf{A}_{iv}.$$

Then $s_i = 0$. Grouping the two directed appearances of every undirected edge gives the exact identity

$$\bar{\Delta}_i^{(k)} = \sum_{\substack{u<v\\ u,v\neq i}} \mathbf{A}_{uv}c_{uv}, \qquad c_{uv} := s_uB_v + s_vB_u.$$

Conditional on the full collection $B = (B_v)_{v\neq i}$ of edges incident to node i, the remaining edge variables in this sum are mutually independent. Write

$$\rho_{uv} := \mathbb{P}(\mathbf{A}_{uv} = 1 \mid \boldsymbol{\sigma}), \qquad \mu_v := \sum_{\substack{u\neq v\\ u\neq i}} s_u\rho_{uv}.$$

For $t = o(1)$, the elementary bound

$$\log(1 - \rho + \rho e^z) \le \rho z + C\rho z^2, \qquad |z| \le 1,$$

and $|c_{uv}| \le 2$ yield

$$\begin{aligned}\log \mathbb{E}\Big[e^{t\bar{\Delta}_i^{(k)}} \mid B, \boldsymbol{\sigma}, \mathcal{L}\Big] &\le t \sum_{\substack{u<v \\ u,v\neq i}} \rho_{uv} c_{uv} + Ct^2 \sum_{\substack{u<v \\ u,v\neq i}} \rho_{uv} c_{uv}^2 \\ &\le \sum_{v\neq i} B_v \left[t\mu_v + Ct^2 \frac{\eta n}{K} \max\{p, q\}\right].\end{aligned}$$

Here we used

$$\sum_{\substack{u<v \\ u,v\neq i}} \rho_{uv} c_{uv} = \sum_{v\neq i} \mu_v B_v, \qquad \sum_{\substack{u<v \\ u,v\neq i}} \rho_{uv} c_{uv}^2 \le C\frac{\eta n}{K} \max\{p, q\} \sum_{v\neq i} B_v.$$

Finally, the incident-edge variables B_v are mutually independent. Thus,

$$\mathbb{E}\Big[e^{t\bar{\Delta}_i^{(k)}} \mid \boldsymbol{\sigma}, \mathcal{L}\Big] \le \prod_{v\neq i} \left[1 - \rho_{iv} + \rho_{iv} \exp\Big(t\mu_v + Ct^2 \frac{\eta n}{K} \max\{p, q\}\Big)\right].$$

This is the required independent MGF factorization; it is obtained only after conditioning on all edges incident to node i, rather than by treating the original column contributions as independent.

We now evaluate this bound in the regime of the theorem. The assumptions $|p - q| = \Theta(p + q)$ and $\eta = \omega(1/(np))$ imply

$$\frac{\eta n(p - q)^2}{K(p + q)} \longrightarrow \infty.$$

We present the case $p > q$ and define the truncated standardized tilt

$$x_n := \min\left\{\frac{1}{2} \log \frac{p(1 - q)}{q(1 - p)}, \left[\frac{\eta n(p - q)^2}{K(p + q)}\right]^{1/4}, (p + q)^{-1/4}\right\}.$$

In the preceding MGF bound, choose

$$t := \frac{Kx_n}{\eta n(p - q)}.$$

When $p < q$, the same argument replaces the first term in the minimum by $\frac{1}{2} \log \frac{q(1-p)}{p(1-q)}$, chooses $t = Kx_n/[\eta n(q - p)]$, and interchanges p and q in the two community-specific factors below. The truncation implies $t = o(1)$ and

$$Ct^2 \frac{\eta n}{K} \max\{p, q\} = O\left(\frac{Kx_n^2(p + q)}{\eta n(p - q)^2}\right) = o(1).$$

For completeness, the random community-size fluctuations can be removed at a smaller order than the target large-deviation scale. Set

$$a_n := (p + q)^{1/4} x_n^{-1/2}$$

and let $\mathcal{E}_{\mathrm{bal}}$ be the event that

$$\max_{c\in[K]} \Big||\mathcal{V}_c| - \frac{n}{K}\Big| \le a_n n.$$

Since x_n is bounded away from zero and $x_n \le (p + q)^{-1/4}$,

$$a_n = o(1), \qquad x_n a_n = o(1), \qquad \frac{a_n^2}{p + q} \longrightarrow \infty.$$

A multinomial Chernoff bound therefore gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{bal}}^c) \le 2K \exp(-cna_n^2) = \exp\big[-\omega(n(p + q))\big].$$

On $\mathcal{E}_{\mathrm{bal}}$, the fixed class-wise sampling rule for $\mathcal{L}$ gives

$$|\mathcal{L}_c| = \frac{\eta n}{K}\big(1 + O(a_n)\big) + O(1) \qquad (c \in [K]).$$

Consequently, uniformly over $v \neq i$,

$$\mu_v = \begin{cases} -\dfrac{\eta n}{K}(p-q) + O\Big(\dfrac{\eta n}{K} a_n(p+q) + (p+q)\Big), & \sigma(v) = r, \\ \dfrac{\eta n}{K}(p-q) + O\Big(\dfrac{\eta n}{K} a_n(p+q) + (p+q)\Big), & \sigma(v) = k, \\ O\Big(\dfrac{\eta n}{K} a_n(p+q) + (p+q)\Big), & \sigma(v) \notin \{r, k\}. \end{cases}$$

Because $x_n a_n = o(1)$ and $K x_n/(\eta n) = o(1)$, substituting the displayed choice of t into the product MGF bound and grouping vertices by community yields the following estimate. In this step, the possibly growing exponential remains harmless because

$$q e^{x_n} \le q \exp\left(\frac{1}{2} \log \frac{p(1-q)}{q(1-p)}\right) = \sqrt{\frac{pq(1-q)}{1-p}} = O(p+q).$$

Indeed, the total contribution of the size, floor, diagonal, and quadratic remainders is bounded by

$$O\left(n(p+q)\left[a_n + x_n a_n + \frac{K x_n}{\eta n} + \frac{K x_n^2 (p+q)}{\eta n (p-q)^2}\right]\right) = o(n(p+q)).$$

Consequently,

$$\log \mathbb{E}\Big[e^{t\bar{\Delta}_i^{(k)}} \mid \boldsymbol{\sigma}, \mathcal{L}\Big] \le \frac{n}{K} \log\big(1 - p + p e^{-x_n}\big) + \frac{n}{K} \log(1 - q + q e^{x_n}) + o(n(p+q)).$$

If the minimum defining x_n is attained by its first term, the sum of the two logarithms equals

$$2 \log\Big(\sqrt{(1-p)(1-q)} + \sqrt{pq}\Big).$$

If the tilt is truncated, then necessarily $x_n \to \infty$ and $q/p \to 0$. Moreover, the definition of x_n gives $x_n \le \frac{1}{2} \log \frac{p(1-q)}{q(1-p)}$, so $q e^{x_n} = o(p)$, and hence the same sum is

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

Thus both cases, together with Markov's inequality, the negligible probability of $\mathcal{E}_{\mathrm{bal}}^c$, and $(\sqrt{p} - \sqrt{q})^2 = \Theta(p+q)$, give

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

Recalling the broader-regime feature contribution established above, combining it with this graph-only bound and then taking a union bound over $k \neq r$ yields the single-node misclassification probability

$$\begin{aligned} \mathbb{P}(\widehat{\sigma}(i) \neq \sigma(i)) &\le (K-1)\Bigg[\exp\Big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2 (1 + o(1))\Big) \\ &\qquad + \exp[-\omega(\Lambda_n)]\Bigg] \\ &= \exp\Big(-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2 (1 + o(1))\Big). \end{aligned}$$

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

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

This proves the stated upper bound on the expected misclassification rate in the regime $1/n \ll p, q \ll 1$.

For the exact-recovery result, continue to fix i, let $r = \sigma(i)$ and $k \neq r$, and consider the logarithmic regime $p = a \log n/n$, $q = b \log n/n$ with fixed $a > b > 0$. Define the shifted-margin scale

$$g_n^{\mathrm{ex}} := \eta \log^2 n \left(\frac{n}{d}\right)^{1/4}.$$

Under $d = \omega(n)$,

$$\left(\frac{n}{d}\right)^{1/4} \to 0, \qquad \sqrt{\frac{d}{n}} \longrightarrow \infty.$$

Moreover, in the present regime the three terms in the minimum defining x_n satisfy

$$\frac{1}{2}\log\frac{p(1-q)}{q(1-p)} \longrightarrow \frac{1}{2}\log\frac{a}{b} \in (0,\infty), \qquad \left[\frac{\eta n(p-q)^2}{K(p+q)}\right]^{1/4} = \Theta\Big((\eta \log n)^{1/4}\Big) \longrightarrow \infty, \qquad (p+q)^{-1/4} \longrightarrow \infty.$$

Thus $x_n = \frac{1}{2}\log\frac{p(1-q)}{q(1-p)}$ for all sufficiently large n, and the preceding choice of t becomes

$$t = \frac{K x_n}{\eta n(p-q)} = \Theta\left(\frac{1}{\eta \log n}\right).$$

Recall also that $\mathcal{E}_{\mathrm{bal}}$ is the community-size balance event in the preceding graph-MGF calculation and that

$$\mathbb{P}(\mathcal{E}_{\mathrm{bal}}^c) \le \exp[-\omega(n(p+q))] \le n^{-\omega(1)}$$

in the present logarithmic regime. On $\mathcal{E}_{\mathrm{bal}}$, recall also that the preceding estimate gives, uniformly over $(\boldsymbol{\sigma}, \mathcal{L})$,

$$\log \mathbb{E}\Big[e^{t\bar{\Delta}_i^{(k)}} \mid \boldsymbol{\sigma}, \mathcal{L}\Big] \le -\frac{(\sqrt{a}-\sqrt{b})^2}{K}\log n + o(\log n).$$

The same conditional MGF bound, now applied at the shifted threshold $-g_n^{\mathrm{ex}}$, gives

$$\begin{aligned}\mathbb{P}\Big(\bar{\Delta}_i^{(k)} \ge -g_n^{\mathrm{ex}}\Big) &\le \mathbb{P}(\mathcal{E}_{\mathrm{bal}}^c) + \exp(t g_n^{\mathrm{ex}})\,\mathbb{E}\Big[\mathbf{1}_{\mathcal{E}_{\mathrm{bal}}}\mathbb{E}\Big(e^{t\bar{\Delta}_i^{(k)}} \mid \boldsymbol{\sigma}, \mathcal{L}\Big)\Big] \\ &\le n^{-(\sqrt{a}-\sqrt{b})^2/K + o(1)},\end{aligned}$$

because

$$t g_n^{\mathrm{ex}} = O\left(\left(\frac{n}{d}\right)^{1/4}\log n\right) = o(\log n).$$

Here the $o(\log n)$ remainder in the preceding conditional MGF estimate is uniform over $(\boldsymbol{\sigma}, \mathcal{L})$ on $\mathcal{E}_{\mathrm{bal}}$. For $i \in \mathcal{L}_r$, the additional term $-\deg(i)$ identified above is favorable, and the removal of one labeled node changes the MGF exponent by only $o(\log n)$. Thus the same shifted upper-tail bound holds.

Recall that

$$\bar{\Delta}_i^{(k)} = \gamma(i,k,r;\mathbf{A}), \qquad R_i(k,r;\mathbf{A}) = \frac{\gamma^2(i,k,r)}{\alpha(i)\beta(k,r) - \gamma^2(i,k,r)}$$

in the non-degenerate case. Recall from Lemma D.3 that

$$\mathcal{E}_{\mathrm{str}}^{\mathrm{ex}} = \left\{\max_{i\in[n]} \alpha(i) \le C\log n, \quad \max_{\substack{k,r\in[K]\\ k\neq r}} \beta(k,r) \le C\eta^2 n \log^2 n\right\}.$$

The same lemma proves $\mathbb{P}((\mathcal{E}_{\mathrm{str}}^{\mathrm{ex}})^c) = o(1)$. On this event and $\gamma(i,k,r) < -g_n^{\mathrm{ex}}$, the non-degenerate case satisfies

$$\begin{aligned}R_i(k,r;\mathbf{A}) &\ge \frac{\gamma^2(i,k,r)}{\alpha(i)\beta(k,r)} \\ &\ge c\frac{\log n}{\sqrt{nd}}.\end{aligned}$$

If $\alpha(i)\beta(k,r) - \gamma^2(i,k,r) = 0$, the conditional feature error is zero by the degenerate-case argument above. Otherwise, uniformly on this shifted graph-good event, monotonicity and $\log(1+x) \ge x/2$ for sufficiently small $x > 0$ give

$$\log(1 + R_i(k,r;\mathbf{A})) \ge c\frac{\log n}{\sqrt{nd}},$$

where $\log n/\sqrt{nd} = o(1)$. Consequently,

$$\mathbb{P}_{\mathbf{X}}\Big(\Delta_i^{(k)} \geq 0 \mid \mathbf{A}\Big) \leq \exp\left[-c\frac{d\log n}{\sqrt{nd}}\right] = \exp\left[-c\sqrt{\frac{d}{n}}\log n\right] = n^{-\omega(1)}.$$

Consequently,

$$\begin{aligned}
&\mathbb{P}(\exists i \in [n] : \widehat{\sigma}(i) \neq \sigma(i)) \\
&\quad \leq \mathbb{P}((\mathcal{E}_{\mathrm{str}}^{\mathrm{ex}})^c) + \sum_{i=1}^{n}\sum_{k\neq\sigma(i)} \mathbb{P}(\gamma(i,k,\sigma(i)) \geq -g_n^{\mathrm{ex}}) + n(K-1)n^{-\omega(1)} \\
&\quad \leq o(1) + n(K-1)n^{-(\sqrt{a}-\sqrt{b})^2/K+o(1)} + n(K-1)n^{-\omega(1)}.
\end{aligned}$$

Thus, if $(\sqrt{a}-\sqrt{b})^2 > K$, the estimator 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

   $$\eta = \omega\left(\frac{1}{np}\right)$$

   and there is a deterministic sequence $\zeta_n \to \infty$ such that

   $$d \geq \frac{2\zeta_n \Lambda_n}{\log\big(1+R^*(\mathbf{A})\big)}$$

   with probability at least $1-\exp[-\omega(\Lambda_n)]$ over $(\boldsymbol{\sigma}, \mathcal{L}, \mathbf{A})$, then the expected misclassification rate of the one-step gradient GCN estimator satisfies

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

   which agrees with the classical minimax misclassification rate at the leading exponential order Zhang & Zhou (2016).
2. In the regime $p, q = \Theta(\log n/n)$ with $|p-q| = \Theta(p+q)$, if

   $$\eta = \omega\left(\frac{1}{\log n}\right), \qquad d = \omega(n),$$

   then the one-step gradient GCN estimator achieves exact recovery whenever

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

   which again coincides with the information-theoretic limit in Abbe (2018).

C.2 Proof of Theorem 4.2

The proof adapts the bad-vertex swapping argument for partially revealed labels in Saad & Nosratinia (2018, Lemmas 9–10). For general K, we fix two different communities and use the candidate-set construction from Jog & Loh (2015, Theorem 3.4 and Section 4.5). We give the details because our labeled sets have fixed class-wise sizes and the community sizes are random.

The assumption of the theorem implies that $1-\beta_n-(\sqrt{a}-\sqrt{b})^2/K$ is bounded below by a positive constant. Since $1-\eta = n^{-\beta_n}$, it follows that $(1-\eta)n = n^{1-\beta_n}$ diverges at a polynomial rate. Write

$$N_k := |\mathcal{V}_k|, \qquad \mathcal{U}_k := \mathcal{V}_k \setminus \mathcal{L}_k,$$

and consider the event

$$\mathcal{E}_{\mathrm{size}}^{\mathrm{swap}} := \left\{\max_{k\in[K]}\Big|N_k - \frac{n}{K}\Big| \leq \sqrt{n}\log n\right\}.$$

Since the labels are independent and uniform, Hoeffding's inequality gives

$$\mathbb{P}((\mathcal{E}_{\text{size}}^{\text{swap}})^c) \le 2K \exp(-2\log^2 n) = o(1).$$

Moreover,

$$|\mathcal{U}_k| = N_k - \lfloor \eta N_k \rfloor = \lceil (1-\eta) N_k \rceil.$$

Uniformly on $\mathcal{E}_{\text{size}}^{\text{swap}}$,

$$|\mathcal{U}_k| = \frac{(1-\eta)n}{K}(1+o(1)).$$

Fix two different communities r and s. Choose $\mathcal{C}_r \subseteq \mathcal{U}_r$ and $\mathcal{C}_s \subseteq \mathcal{U}_s$ by a fixed rule that depends only on $(\boldsymbol{\sigma}, \mathcal{L})$ and not on $\mathbf{A}$, with

$$|\mathcal{C}_r| = |\mathcal{C}_s| =: h_n = \left\lfloor \frac{\min\{|\mathcal{U}_r|, |\mathcal{U}_s|\}}{\log^3 n} \right\rfloor.$$

Let $\mathcal{C} := \mathcal{C}_r \cup \mathcal{C}_s$. The preceding estimate gives

$$\log h_n = (1-\beta_n)\log n - 3\log\log n + O(1).$$

We first record how the fixed labeled counts behave under a swap. For a candidate assignment $\boldsymbol{\sigma}'$, let

$$\mathcal{V}_k(\boldsymbol{\sigma}') := \{v : \sigma'_v = k\}, \qquad N_k(\boldsymbol{\sigma}') := |\mathcal{V}_k(\boldsymbol{\sigma}')|.$$

The likelihood of the observed labeled sets is given below. In this display, the sets $\mathcal{L}_k$ are fixed by the observed labels.

$$\mathbb{P}(\mathcal{L}, \boldsymbol{\sigma}_{\mathcal{L}} \mid \boldsymbol{\sigma}') = \prod_{k=1}^{K} \frac{\mathbb{1}\{\mathcal{L}_k \subseteq \mathcal{V}_k(\boldsymbol{\sigma}'),\ |\mathcal{L}_k| = \lfloor \eta N_k(\boldsymbol{\sigma}') \rfloor\}}{\binom{N_k(\boldsymbol{\sigma}')}{\lfloor \eta N_k(\boldsymbol{\sigma}') \rfloor}}.$$

If $i \in \mathcal{U}_r$ and $j \in \mathcal{U}_s$, let $\boldsymbol{\sigma}^{ij}$ be obtained by exchanging their labels. The two nodes are unlabeled, and the exchange preserves every N_k. It therefore preserves the likelihood above exactly. It also preserves the i.i.d. uniform prior, and the feature density is unchanged because $\mathbf{X}$ is independent of the labels.

Define the edge log-likelihood ratio

$$\varphi_n(x) := x \log \frac{p}{q} + (1-x)\log \frac{1-p}{1-q}.$$

For $i \in \mathcal{C}_r$, set

$$T_i^{r,s} := \sum_{u \in \mathcal{V}_r \setminus \mathcal{C}_r} \varphi_n(A_{iu}) - \sum_{v \in \mathcal{V}_s \setminus \mathcal{C}_s} \varphi_n(A_{iv}).$$

For $j \in \mathcal{C}_s$, define

$$T_j^{s,r} := \sum_{v \in \mathcal{V}_s \setminus \mathcal{C}_s} \varphi_n(A_{jv}) - \sum_{u \in \mathcal{V}_r \setminus \mathcal{C}_r} \varphi_n(A_{ju}).$$

Conditional on $(\boldsymbol{\sigma}, \mathcal{L})$, these variables use pairwise disjoint edges. They are mutually independent and are also independent of the graph induced by $\mathcal{C}$.

We next bound the probability that one of these local margins favors the wrong community. For $i \in \mathcal{C}_r$, write

$$E_i^{(r)} := \sum_{u \in \mathcal{V}_r \setminus \mathcal{C}_r} A_{iu}, \qquad E_i^{(s)} := \sum_{v \in \mathcal{V}_s \setminus \mathcal{C}_s} A_{iv}.$$

Conditional on $(\boldsymbol{\sigma}, \mathcal{L})$, these variables are independent and satisfy

$$E_i^{(r)} \sim \mathrm{Bin}(N_r - h_n, p), \qquad E_i^{(s)} \sim \mathrm{Bin}(N_s - h_n, q).$$

We have

$$T_i^{r,s} = \big(E_i^{(r)} - E_i^{(s)}\big) \log \frac{p(1-q)}{q(1-p)} + (N_r - N_s)\varphi_n(0).$$

On $\mathcal{E}_{\text{size}}^{\text{swap}}$,

$$(N_r - N_s)\varphi_n(0) = O\left(\frac{\log^2 n}{\sqrt{n}}\right) = o(1), \qquad \log\frac{p(1-q)}{q(1-p)} \longrightarrow \log\frac{a}{b} > 0.$$

Choose a fixed integer $c_{\text{sw}} \geq 1$ such that

$$c_{\text{sw}} \log\frac{a}{b} > 2.$$

For all sufficiently large n, the event $E_i^{(s)} = E_i^{(r)} + c_{\text{sw}}$ implies $T_i^{r,s} \leq -1$. Set

$$s_n := \left\lfloor \sqrt{(N_r - h_n)(N_s - h_n)pq} \right\rfloor.$$

The following calculation plays the same role as the local lower-tail bounds in Saad & Nosratinia (2018, Lemma 10 and its proof) and Jog & Loh (2015, Lemma 4.7). Here we derive the bound directly for Bernoulli edges and random community sizes. A direct Stirling expansion of the two binomial point probabilities gives, uniformly on $\mathcal{E}_{\text{size}}^{\text{swap}}$,

$$\begin{aligned}
&\log \mathbb{P}\Big(E_i^{(r)} = s_n,\ E_i^{(s)} = s_n + c_{\text{sw}} \mid \boldsymbol{\sigma}, \mathcal{L}\Big) \\
&\quad = -(N_r - h_n)p - (N_s - h_n)q + 2\sqrt{(N_r - h_n)(N_s - h_n)pq} + o(\log n) \\
&\quad = -\frac{(\sqrt{a} - \sqrt{b})^2}{K}\log n + o(\log n).
\end{aligned}$$

The same calculation applies to every $j \in \mathcal{C}_s$. Consequently, uniformly over all configurations in the support of the model for which $\mathcal{E}_{\text{size}}^{\text{swap}}$ holds,

$$\min\left\{ \min_{i \in \mathcal{C}_r} \mathbb{P}(T_i^{r,s} \leq -1 \mid \boldsymbol{\sigma}, \mathcal{L}),\, \min_{j \in \mathcal{C}_s} \mathbb{P}\big(T_j^{s,r} \leq -1 \mid \boldsymbol{\sigma}, \mathcal{L}\big) \right\} \geq n^{-(\sqrt{a}-\sqrt{b})^2/K + o(1)},$$

where the $o(1)$ term is deterministic and uniform over these configurations.

Let $\mathcal{C}_r^\circ$ contain the nodes in $\mathcal{C}_r$ having no edge to any other node in $\mathcal{C}$, and define $\mathcal{C}_s^\circ$ in the same way. Since $q < p$,

$$\mathbb{E}[|\mathcal{C}_r \setminus \mathcal{C}_r^\circ| \mid \boldsymbol{\sigma}, \mathcal{L}] \leq 2h_n^2 p.$$

Markov's inequality gives

$$\mathbb{P}\left(|\mathcal{C}_r^\circ| < \frac{h_n}{2} \mid \boldsymbol{\sigma}, \mathcal{L}\right) \leq 4h_n p = O(\log^{-2} n).$$

The same bound holds for $\mathcal{C}_s^\circ$. These two sets depend only on the graph induced by $\mathcal{C}$, so they are independent of all the local margins above.

Conditional on the induced graph and on $|\mathcal{C}_r^\circ| \geq h_n/2$, independence and the uniform bound above give

$$\begin{aligned}
&\mathbb{P}(T_i^{r,s} > -1 \text{ for every } i \in \mathcal{C}_r^\circ \mid \boldsymbol{\sigma}, \mathcal{L}, \mathbf{A}_{\mathcal{C}}) \\
&\qquad \leq \exp\left[-\frac{h_n}{2} n^{-(\sqrt{a}-\sqrt{b})^2/K + o(1)}\right].
\end{aligned}$$

The exponent diverges because

$$\log\left(h_n n^{-(\sqrt{a}-\sqrt{b})^2/K + o(1)}\right) \geq \left(1 - \beta_n - \frac{(\sqrt{a} - \sqrt{b})^2}{K} + o(1)\right)\log n - 3\log\log n \longrightarrow +\infty.$$

The same conclusion holds for $\mathcal{C}_s^\circ$. Combining these bounds with $\mathbb{P}((\mathcal{E}_{\text{size}}^{\text{swap}})^c) = o(1)$ and using a union bound, we find that, with probability $1 - o(1)$, there are nodes $i \in \mathcal{C}_r^\circ$ and $j \in \mathcal{C}_s^\circ$ such that

$$T_i^{r,s} \leq -1, \qquad T_j^{s,r} \leq -1.$$

This is the fixed-count counterpart of the pair-swapping argument in Saad & Nosratinia (2018, Lemma 9) and Jog & Loh (2015, Lemma 4.4). For this pair, the graph likelihood ratio is

$$\log\frac{\mathbb{P}(\mathbf{A} \mid \boldsymbol{\sigma})}{\mathbb{P}(\mathbf{A} \mid \boldsymbol{\sigma}^{ij})} = T_i^{r,s} + T_j^{s,r} \leq -2.$$

The edge (i, j) is a between-community edge under both assignments, and the edges to the other $K - 2$ communities are unchanged. Since i and j have no edges inside $\mathcal{C}$ and $|\mathcal{C}_r| = |\mathcal{C}_s|$, the remaining internal nonedge terms cancel. This proves the displayed identity.

Let

$$O_n := (\mathbf{A}, \mathbf{X}, \mathcal{L}, \boldsymbol{\sigma}_{\mathcal{L}}).$$

The invariance of the prior, feature density, and labeled-set likelihood now gives

$$\mathbb{P}\big(\boldsymbol{\sigma}^{ij} \mid O_n\big) > \mathbb{P}(\boldsymbol{\sigma} \mid O_n)$$

with probability $1 - o(1)$. Thus the true assignment is not a posterior maximizer with probability $1 - o(1)$.

Let $\widehat{\boldsymbol{\sigma}}_{\mathrm{MAP}}$ be a posterior maximizer with fixed tie breaking. Then

$$\mathbb{P}(\widehat{\boldsymbol{\sigma}}_{\mathrm{MAP}} = \boldsymbol{\sigma}) = o(1).$$

Conditional on O_n, the success probability of any possibly randomized estimator is a convex combination of the posterior probabilities and is therefore at most their maximum. This maximum is attained by the MAP estimator. Consequently,

$$\sup_{\widehat{\boldsymbol{\sigma}}} \mathbb{P}(\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}) = \mathbb{P}(\widehat{\boldsymbol{\sigma}}_{\mathrm{MAP}} = \boldsymbol{\sigma}) = o(1).$$

This proves the theorem.

C.3 Proof of Theorem 4.4

The proof is divided into three parts. The first part conditions the model on a high-probability balance event and uses symmetrization to reduce the risk to one representative node. The second part checks that the fixed labeled counts do not reveal this node label with high probability. The last part applies a genie-aided testing lower bound.

Prior and symmetrization. Define

$$I := I(p, q) := -2 \log\Big(\sqrt{pq} + \sqrt{(1-p)(1-q)}\Big).$$

Under the assumptions $p, q \ll 1$ and $|p - q| = \Theta(p + q)$,

$$I = (\sqrt{p} - \sqrt{q})^2(1 + o(1)) \asymp p + q.$$

Since $p, q \gg 1/n$, we have

$$nI \to \infty.$$

Let Π_0 denote the joint distribution of $(\boldsymbol{\sigma}, \mathcal{L})$ specified in the theorem. Fix a deterministic sequence $\varepsilon_n \downarrow 0$ such that $n\varepsilon_n^2 \to \infty$. Under Π_0,

$$\sigma_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([K]), \qquad i \in [n],$$

and define the regular event

$$\mathcal{E}_{\mathrm{bal}} = \left\{ \max_{k \in [K]} \Big| |\mathcal{V}_k| - \frac{n}{K} \Big| \le \varepsilon_n n \right\}.$$

For each fixed k, the count $|\mathcal{V}_k|$ follows $\mathrm{Bin}(n, 1/K)$ and has mean n/K. By the Chernoff bound, for some constant $c > 0$,

$$\mathbb{P}\Big(\Big| |\mathcal{V}_k| - \frac{n}{K} \Big| > \varepsilon_n n \Big) \le 2 \exp(-cn\varepsilon_n^2).$$

Since $K = O(1)$, a union bound over $k \in [K]$ gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{bal}}^c) \le 2K \exp(-cn\varepsilon_n^2) = o(1).$$

Hence $\mathbb{P}(\mathcal{E}_{\mathrm{bal}}) \to 1$. This will be used below to show that conditioning on the balanced event does not change the relevant probability estimates.

Given $\boldsymbol{\sigma}$, choose the sets $\mathcal{L}_k$ independently across $k \in [K]$, with each set chosen uniformly among all subsets of $\mathcal{V}_k$ with size

$$|\mathcal{L}_k| = \lfloor \eta |\mathcal{V}_k| \rfloor,$$

and set

$$\mathcal{L} = \bigcup_{k=1}^{K} \mathcal{L}_k, \qquad \mathcal{U} = \mathcal{V} \setminus \mathcal{L}.$$

Let

$$B_n := \inf_{\hat{\boldsymbol{\sigma}}} \mathbb{E}_{\Pi_0, \mathbf{A}, \mathbf{X}} \left[\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) \right]$$

be the optimal expected risk in the unconditioned model of the theorem. Let $\Pi := \Pi_0(\,\cdot \mid \mathcal{E}_{\mathrm{bal}})$ and define

$$B_n^{\mathrm{bal}} := \inf_{\hat{\boldsymbol{\sigma}}} \mathbb{E}_{\Pi, \mathbf{A}, \mathbf{X}} \left[\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) \right].$$

For every estimator, its risk under Π_0 is at least its conditional risk on $\mathcal{E}_{\mathrm{bal}}$ multiplied by $\mathbb{P}(\mathcal{E}_{\mathrm{bal}})$. Taking the infimum gives

$$B_n \geq \mathbb{P}(\mathcal{E}_{\mathrm{bal}}) B_n^{\mathrm{bal}} = (1 - o(1)) B_n^{\mathrm{bal}}.$$

Unless stated otherwise, all probabilities and expectations below are taken under the conditional law Π together with the graph and feature randomness.

The prior Π, the SBM likelihood, and the loss $\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma})$ are invariant under permutations of nodes and permutations of community labels. Since the expected risk is linear in the decision rule, allowing randomized estimators does not change its infimum.

For any estimator $\hat{\boldsymbol{\sigma}}$, let π be a uniformly random permutation of $[n]$, and let τ be a uniformly random permutation of $[K]$, independent of all data. Apply (π, τ) to the input

$$(\mathbf{A}, \mathbf{X}, \mathcal{L}, \{\sigma_i : i \in \mathcal{L}\}),$$

run $\hat{\boldsymbol{\sigma}}$ on the transformed input, and then invert (π, τ) on the output. Averaging over (π, τ) gives an equivariant randomized estimator with the same expected risk. Therefore, in proving the lower bound, we may restrict attention to estimators that are equivariant with respect to both node permutations and community-label permutations. Until the genie-aided comparison below, all infima are understood to be over equivariant randomized estimators.

Fix an arbitrary representative node $i_0 \in \mathcal{V}$. For any such equivariant estimator,

$$\mathbb{E}_{\Pi, \mathbf{A}, \mathbf{X}} \left[\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) \right] = \mathbb{P}_{\Pi, \mathbf{A}, \mathbf{X}} \left(\hat{\sigma}_{i_0} \neq \sigma_{i_0} \right).$$

Thus the symmetrization reduces the optimal expected risk for the global average misclassification rate to the error probability of one representative node. Thus

$$B_n^{\mathrm{bal}} = \inf_{\hat{\boldsymbol{\sigma}}} \mathbb{P} \left(\hat{\sigma}_{i_0} \neq \sigma_{i_0} \right).$$

Since node i_0 is unlabeled with probability $1 - \eta + o(1)$,

$$B_n^{\mathrm{bal}} \geq (1 - \eta + o(1)) \inf_{\hat{\boldsymbol{\sigma}}} \mathbb{P} \left(\hat{\sigma}_{i_0} \neq \sigma_{i_0} \mid i_0 \in \mathcal{U} \right).$$

Because $\eta = o(1)$, the prefactor $1 - \eta + o(1)$ is $1 + o(1)$ and hence does not affect the leading exponent.

Labeled counts. The labeled set is sampled with fixed class-wise sizes $|\mathcal{L}_k| = \lfloor \eta |\mathcal{V}_k| \rfloor$. If all labels except σ_{i_0} are revealed, these counts may help identify σ_{i_0}, since they depend on the class sizes. We now condition on a high-probability event on which this does not happen.

For $k \in [K]$, define

$$N_k^{(-i_0)} := |\{j \neq i_0 : \sigma_j = k\}|.$$

Let

$$\mathcal{E}_{\mathrm{nl}} := \left\{ \lfloor \eta (N_k^{(-i_0)} + 1) \rfloor = \lfloor \eta N_k^{(-i_0)} \rfloor \text{ for all } k \in [K] \right\}.$$

On $\mathcal{E}_{\mathrm{nl}}$, if node i_0 is unlabeled, then adding node i_0 to any candidate community does not change the required number of labeled nodes in that community. Thus the fixed labeled counts do not distinguish among the candidate values of σ_{i_0}.

We also need to ensure that the almost-balance conditioning does not reveal σ_{i_0}. Define the inner-balance event

$$\mathcal{E}_{\text{int}} := \left\{ \max_{k\in[K]} \left| N_k^{(-i_0)} - \frac{n}{K} \right| \le \frac{1}{2}\varepsilon_n n \right\}.$$

For each fixed k, a Chernoff bound for multinomial counts gives

$$\mathbb{P}\left(\left| N_k^{(-i_0)} - \frac{n}{K} \right| > \frac{1}{2}\varepsilon_n n \right) \le 2\exp(-cn\varepsilon_n^2)$$

for some constant $c > 0$. Since $K = O(1)$ and $n\varepsilon_n^2 \to \infty$, a union bound over $k \in [K]$ gives

$$\mathbb{P}(\mathcal{E}_{\text{int}}^c) = o(1).$$

As $\mathbb{P}(\mathcal{E}_{\text{bal}}) \to 1$, it follows that

$$\mathbb{P}(\mathcal{E}_{\text{int}} \mid \mathcal{E}_{\text{bal}}) = 1 - o(1).$$

Moreover, since $\varepsilon_n n \to \infty$, on $\mathcal{E}_{\text{int}}$ every candidate value of σ_{i_0} remains compatible with $\mathcal{E}_{\text{bal}}$ after adding node i_0 to the corresponding community. Therefore, the balance constraint also does not distinguish among the candidate values of σ_{i_0} on $\mathcal{E}_{\text{int}}$.

We now show that

$$\mathbb{P}(\mathcal{E}_{\text{nl}} \mid i_0 \in \mathcal{U}) = 1 - o(1).$$

It is enough to verify this under the unconditioned i.i.d. label distribution. Indeed,

$$\mathbb{P}(\mathcal{E}_{\text{bal}}) \to 1$$

and, uniformly on $\mathcal{E}_{\text{bal}}$,

$$\mathbb{P}(i_0 \in \mathcal{U} \mid \boldsymbol{\sigma}) = 1 - \frac{\lfloor \eta |\mathcal{V}_{\sigma_{i_0}}| \rfloor}{|\mathcal{V}_{\sigma_{i_0}}|} = 1 - \eta + O(1/n) = 1 + o(1).$$

It follows that

$$\mathbb{P}(\mathcal{E}_{\text{bal}} \cap \{i_0 \in \mathcal{U}\}) = 1 - o(1).$$

Consequently,

$$\mathbb{P}(\mathcal{E}_{\text{nl}}^c \,|\, \mathcal{E}_{\text{bal}},\ i_0 \in \mathcal{U}) \le \frac{\mathbb{P}(\mathcal{E}_{\text{nl}}^c)}{\mathbb{P}(\mathcal{E}_{\text{bal}} \cap \{i_0 \in \mathcal{U}\})}.$$

Thus an unconditioned $o(1)$ bound implies the desired conditional $o(1)$ bound.

Fix $k \in [K]$. The event

$$\lfloor \eta(N_k^{(-i_0)} + 1) \rfloor \ne \lfloor \eta N_k^{(-i_0)} \rfloor$$

can occur only when $\eta N_k^{(-i_0)}$ lies within distance η of an integer. Choose a sequence $M_n \to \infty$ so slowly that

$$\eta M_n \to 0.$$

Since $N_k^{(-i_0)}$ is a binomial-type count with variance $O(n)$, there exists an interval $I_{n,k}$ of length $O(M_n\sqrt{n})$ around its mean such that

$$\mathbb{P}\left(N_k^{(-i_0)} \notin I_{n,k} \right) = o(1).$$

Inside this interval, the number of integers m for which

$$\lfloor \eta(m+1) \rfloor \ne \lfloor \eta m \rfloor$$

is at most $O(\eta M_n\sqrt{n} + 1)$. Let

$$B_{n,k} := \{ m \in I_{n,k} : \lfloor \eta(m+1) \rfloor \ne \lfloor \eta m \rfloor \}.$$

Then $|B_{n,k}| \le O(\eta M_n\sqrt{n} + 1)$. The maximum point mass of a binomial-type count satisfies

$$\max_m \mathbb{P}(N_k^{(-i_0)} = m) = O(n^{-1/2}),$$

since $N_k^{(-i_0)} \sim \mathrm{Bin}(n-1, 1/K)$ and $K = O(1)$. Hence

$$\begin{aligned}
&\mathbb{P}\left(\lfloor \eta(N_k^{(-i_0)}+1)\rfloor \neq \lfloor \eta N_k^{(-i_0)}\rfloor\right) \\
&\quad \le \mathbb{P}\Big(N_k^{(-i_0)} \notin I_{n,k}\Big) + \sum_{m\in B_{n,k}} \mathbb{P}(N_k^{(-i_0)} = m) \\
&\quad \le o(1) + O(\eta M_n\sqrt{n}+1)\, O(n^{-1/2}) \\
&\quad = o(1) + O(\eta M_n) + O(n^{-1/2}) = o(1).
\end{aligned}$$

Taking a union bound over the fixed number $K = O(1)$ of communities gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{nl}} \mid i_0 \in \mathcal{U}) = 1 - o(1).$$

Combining the two high-probability events, define

$$\mathcal{G}_0 := \{i_0 \in \mathcal{U}\} \cap \mathcal{E}_{\mathrm{nl}} \cap \mathcal{E}_{\mathrm{int}}.$$

Then

$$\mathbb{P}(\mathcal{G}_0 \mid i_0 \in \mathcal{U}) = 1 - o(1).$$

Recalling that B_n^{bal} denotes the conditional optimal risk defined in the preceding symmetrization part, we obtain

$$B_n^{\mathrm{bal}} \ge (1-o(1)) \inf_{\hat{\boldsymbol{\sigma}}} \mathbb{P}\left(\hat{\sigma}_{i_0} \neq \sigma_{i_0} \,|\, \mathcal{G}_0\right).$$

Genie-aided testing. Recall that the information quantity introduced at the beginning of this proof is

$$I = I(p,q) = -2\log\Big(\sqrt{pq} + \sqrt{(1-p)(1-q)}\Big), \qquad nI \longrightarrow \infty.$$

Condition on $\mathcal{G}_0$. Now give the estimator the genie information

$$\{\sigma_j : j \neq i_0\}.$$

This can only make the estimation problem easier. Therefore, the conditional error probability in the original semi-supervised problem is at least the optimal error probability in the genie-aided problem, where we may enlarge the class to all measurable genie-aided decision rules.

On $\mathcal{G}_0$, neither the class-wise labeled counts in

$$(\mathcal{L}, \{\sigma_i : i \in \mathcal{L}\})$$

nor the almost-balance constraint rules out any candidate value of σ_{i_0}. Indeed, by $\mathcal{E}_{\mathrm{nl}}$, the class-wise labeled counts are the same for all candidate values of σ_{i_0} after conditioning on $i_0 \in \mathcal{U}$. By $\mathcal{E}_{\mathrm{int}}$, all candidate values remain compatible with $\mathcal{E}_{\mathrm{bal}}$.

Moreover, under the prior construction, the labeled set is sampled independently of the graph and the features conditional on $\boldsymbol{\sigma}$. The features are non-informative. In particular,

$$\mathcal{L} \perp (\mathbf{A}, \mathbf{X}) \mid \boldsymbol{\sigma}, \qquad \mathbf{X} \perp (\mathbf{A}, \boldsymbol{\sigma}).$$

Thus the conditional distribution of $\mathbf{X}$ is identical under all hypotheses $\sigma_{i_0} = k$, $k \in [K]$, and $\mathbf{X}$ cancels from the likelihood ratio.

After the genie reveals $\{\sigma_j : j \neq i_0\}$, all edges not incident to node i_0 also have distributions independent of σ_{i_0}. Therefore, the genie-aided problem reduces to estimating σ_{i_0} from the incident edge vector

$$(\mathbf{A}_{i_0 j})_{j \neq i_0}$$

and the known community labels $\{\sigma_j : j \neq i_0\}$.

Fix two distinct communities $r, s \in [K]$. On $\mathcal{E}_{\mathrm{nl}}$, write

$$\ell_k := |\mathcal{L}_k| = \left\lfloor \eta N_k^{(-i_0)} \right\rfloor, \qquad k \in [K].$$

For any candidate $h \in [K]$, the conditional probability of the observed labeled set is

$$\mathbb{P}(\mathcal{L} \,|\, \sigma_{i_0} = h, \boldsymbol{\sigma}_{-i_0}) = \binom{N_h^{(-i_0)}+1}{\ell_h}^{-1} \prod_{k \neq h} \binom{N_k^{(-i_0)}}{\ell_k}^{-1}.$$

Since all candidates are compatible with $\mathcal{E}_{\mathrm{bal}}$ on $\mathcal{E}_{\mathrm{int}}$, it follows that

$$\frac{\Pi(\sigma_{i_0} = r \,|\, \boldsymbol{\sigma}_{-i_0}, \mathcal{L}, \mathcal{G}_0)}{\Pi(\sigma_{i_0} = s \,|\, \boldsymbol{\sigma}_{-i_0}, \mathcal{L}, \mathcal{G}_0)} = \frac{1 - \ell_r/(N_r^{(-i_0)} + 1)}{1 - \ell_s/(N_s^{(-i_0)} + 1)} = 1 + o(1),$$

where the last estimate is uniform on $\mathcal{E}_{\mathrm{int}}$ because $N_k^{(-i_0)} \asymp n$ and $\eta = o(1)$. Hence the K-class testing error is bounded from below, up to a constant depending only on K, by the minimum average error probability of the binary testing problem

$$H_r : \sigma_{i_0} = r \qquad \text{versus} \qquad H_s : \sigma_{i_0} = s.$$

Since $K = O(1)$ and $nI \to \infty$, this constant has no effect on the leading exponent.

Let

$$m_r = |\{j \neq i_0 : \sigma_j = r\}|, \qquad m_s = |\{j \neq i_0 : \sigma_j = s\}|.$$

On $\mathcal{E}_{\mathrm{int}}$,

$$m_r = \frac{n}{K}(1 + o(1)), \qquad m_s = \frac{n}{K}(1 + o(1)).$$

We first treat the case $p > q$. When $p < q$, the same binary testing argument applies after interchanging p and q and reversing the likelihood-ratio inequality. Both $I(p, q)$ and the final exponent are symmetric in p and q.

Under H_r, the edges from node i_0 to community r are independent $\mathrm{Bern}(p)$ variables, while the edges from node i_0 to community s are independent $\mathrm{Bern}(q)$ variables. Under H_s, the roles of p and q are reversed. Edges from node i_0 to communities other than r and s have distribution $\mathrm{Bern}(q)$ under both hypotheses and hence cancel from the likelihood ratio.

To reduce the possibly unequal block sizes to an equal-size test, define

$$\overline{m}_n := \left\lceil \frac{n}{K} + \frac{1}{2}\varepsilon_n n \right\rceil.$$

On $\mathcal{E}_{\mathrm{int}}$, both m_r and m_s are at most $\overline{m}_n$, and

$$\overline{m}_n = \frac{n}{K}(1 + o(1)).$$

Augment the experiment by adding independent observations to the two informative blocks until both have size $\overline{m}_n$. Under H_r, the added observations in the r block have distribution $\mathrm{Bern}(p)$ and those in the s block have distribution $\mathrm{Bern}(q)$. Under H_s, these two distributions are interchanged. The original two-block binary experiment is a marginal of the augmented experiment, so the augmented experiment can only have a smaller minimum average error probability.

The augmented problem is an equal-size local testing problem. Let

$$X_u \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(q), \qquad Y_u \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p), \qquad \{X_u\}_{u=1}^{\overline{m}_n} \perp \{Y_u\}_{u=1}^{\overline{m}_n}.$$

Lemma 5.2 of Zhang & Zhou (2016), applied after replacing its ambient size and its two edge parameters by

$$K\overline{m}_n, \qquad K\overline{m}_n p, \qquad K\overline{m}_n q,$$

respectively, gives the following bound. Here $K\overline{m}_n = (1 + o(1))n$ and $\overline{m}_n I(p, q) \to \infty$.

$$\mathbb{P}\left(\sum_{u=1}^{\overline{m}_n} X_u \geq \sum_{u=1}^{\overline{m}_n} Y_u\right) \geq \exp\{-\overline{m}_n I(p, q)(1 + o(1))\}.$$

The $1+o(1)$ prior odds imply that the minimum average error probability in the augmented experiment is bounded below by $1 - o(1)$ times its value under equal priors. Under equal priors, randomization at a tie bounds this error below by one half of the displayed probability. Consequently, the minimum average error probability in the original binary experiment is at least

$$\exp\{-\overline{m}_n I(p, q)(1 + o(1))\} = \exp\Big\{-\frac{n}{K} I(p, q)(1 + o(1))\Big\}.$$

Combining the two-hypothesis reduction with this local testing lower bound yields

$$\inf_{\hat{\boldsymbol{\sigma}}} \mathbb{P}\left(\hat{\sigma}_{i_0} \neq \sigma_{i_0} \,|\, \mathcal{G}_0\right) \geq \exp\Big\{-\frac{n}{K} I(p,q)(1+o(1))\Big\}.$$

Substituting the last display into the lower bound above gives

$$B_n^{\mathrm{bal}} \geq (1-o(1)) \exp\Big\{-\frac{n}{K} I(p,q)(1+o(1))\Big\}.$$

Since $nI \to \infty$, the prefactor $1-o(1)$ is $\exp(o(nI))$. Hence

$$B_n^{\mathrm{bal}} \geq \exp\Big\{-\frac{n}{K} I(p,q)(1+o(1))\Big\}.$$

Returning to the unconditioned model and using $\mathbb{P}(\mathcal{E}_{\mathrm{bal}}) = 1-o(1)$, the comparison at the beginning of the proof gives

$$B_n \geq \exp\Big\{-\frac{n}{K} I(p,q)(1+o(1))\Big\}.$$

Finally, since $p, q \ll 1$,

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

Therefore,

$$\begin{aligned}\sqrt{pq} + \sqrt{(1-p)(1-q)} &= 1 - \frac{p+q}{2} + \sqrt{pq} + o(p+q) \\ &= 1 - \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + o(p+q).\end{aligned}$$

Because $|p-q| = \Theta(p+q)$,

$$(\sqrt{p} - \sqrt{q})^2 = \frac{(p-q)^2}{(\sqrt{p}+\sqrt{q})^2} \asymp p+q.$$

Thus

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

Using $-\log(1-x) = x(1+o(1))$ as $x \to 0$, we obtain

$$\begin{aligned} I(p,q) &= -2\log\Big(\sqrt{pq} + \sqrt{(1-p)(1-q)}\Big) \\ &= -2\log\Big[1 - \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + o\big((\sqrt{p} - \sqrt{q})^2\big)\Big] \\ &= (\sqrt{p} - \sqrt{q})^2(1+o(1)).\end{aligned}$$

Consequently,

$$\inf_{\hat{\boldsymbol{\sigma}}} \mathbb{E}_{\boldsymbol{\sigma},\mathcal{L},\mathbf{A},\mathbf{X}}\left[\mathrm{err}(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma})\right] \geq \exp\Big\{-\frac{n}{K}(\sqrt{p} - \sqrt{q})^2(1+o(1))\Big\}.$$

The proof is complete.

C.4 Proof of Theorem 4.5

Recall from Section 3 that

$$\widehat{\mathbf{S}}^{(L)} = \mathbf{G}_{:\mathcal{L}}^{(L)} \mathbf{Y}_{\mathcal{L}}.$$

The assumptions of the theorem imply the conditions used in the auxiliary lemmas below. Indeed, they give

$$\delta_n \to \infty, \qquad \delta_n^{2L} = o(n), \qquad \delta_n \gamma_n^{2L} \to \infty, \qquad \eta n \gamma_n^{2L} \to \infty.$$

Moreover, $\delta_n^2 = o(n)$, and hence

$$n\gamma_n^{4L} = \frac{n}{\delta_n^2}\left(\delta_n \gamma_n^{2L}\right)^2 \to \infty.$$

Fix $i \in [n]$ and write $r = \sigma(i)$. Fix a class $k \neq r$ and define

$$\boldsymbol{v}_{k,r} := \mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r},$$
$$\Delta_{i,k}^{(L)} := \widehat{\mathbf{S}}_{i,k}^{(L)} - \widehat{\mathbf{S}}_{i,r}^{(L)}.$$

If an unlabeled node i is misclassified, then $\Delta_{i,k}^{(L)} \geq 0$ for at least one class $k \neq r$. It is therefore enough to bound these pairwise events. The graph quantities used below are

$$\bar{\Delta}_{i,k}^{(L)} := \boldsymbol{e}_i^\top \mathbf{P}^{2L} \boldsymbol{v}_{k,r} = \sum_{\ell \in \mathcal{L}_k} (\mathbf{P}^{2L})_{i\ell} - \sum_{\ell \in \mathcal{L}_r} (\mathbf{P}^{2L})_{i\ell},$$

$$\alpha_L(i; \mathbf{A}) := \boldsymbol{e}_i^\top \mathbf{P}^{2L} \boldsymbol{e}_i, \qquad \beta_L(k, r; \mathbf{A}) := \boldsymbol{v}_{k,r}^\top \mathbf{P}^{2L} \boldsymbol{v}_{k,r}.$$

Here $\boldsymbol{e}_i$ denotes the ith standard basis vector. Recall also from Lemma D.6 that the explicit remainder is

$$\mathcal{R}_{n,L} := 2K \exp\{-cn\gamma_n^{4L}\} + 2\exp\left(-c_D \frac{n\delta_n}{(\log n)^2}\right).$$

Here $c, c_D > 0$ are the constants appearing in that lemma. First observe that the explicit remainder in Lemma D.6 is absorbed by the graph term. Indeed,

$$\frac{n\gamma_n^{4L}}{\mathcal{W}_{n,L}} = \frac{n}{\eta\delta_n} \longrightarrow \infty, \qquad \frac{n\delta_n/(\log n)^2}{\mathcal{W}_{n,L}} = \frac{n}{\eta\gamma_n^{4L}(\log n)^2} \longrightarrow \infty.$$

The first limit follows from $\delta_n^{2L} = o(n)$ and $\eta \leq 1$, while the second follows from $\eta\gamma_n^{4L} \leq 1$. Thus, after changing constants,

$$\mathcal{R}_{n,L} \leq C \exp\{-c\mathcal{W}_{n,L}\}.$$

The negative of $\bar{\Delta}_{i,k}^{(L)}$ is the normalized true-class margin in Lemma D.6. By that lemma and Lemma D.7, outside an event of probability at most

$$Ce^{-c\mathcal{W}_{n,L}} + Ce^{-c\mathcal{V}_{n,L}},$$

we have

$$\bar{\Delta}_{i,k}^{(L)} \leq -c\eta\gamma_n^{2L},$$

as well as

$$\alpha_L(i; \mathbf{A}) \leq C\delta_n^{-L}, \qquad \beta_L(k, r; \mathbf{A}) \leq C\eta n\delta_n^{-L}(1 + \eta\delta_n^L).$$

For each feature coordinate a, let $\mathbf{X}_{\cdot a}$ denote the ath column of $\mathbf{X}$, and define

$$Z_a := (\mathbf{P}^L \mathbf{X}_{\cdot a})_i,$$

and

$$V_a := \boldsymbol{v}_{k,r}^\top \mathbf{P}^L \mathbf{X}_{\cdot a} = \sum_{\ell \in \mathcal{L}_k} (\mathbf{P}^L \mathbf{X}_{\cdot a})_\ell - \sum_{\ell \in \mathcal{L}_r} (\mathbf{P}^L \mathbf{X}_{\cdot a})_\ell.$$

Conditional on $(\mathbf{A}, \boldsymbol{\sigma}, \mathcal{L})$, the pairs (Z_a, V_a) are i.i.d. centered jointly Gaussian and

$$\Delta_{i,k}^{(L)} = \sum_{a=1}^{d} Z_a V_a.$$

Define their conditional variance and covariance parameters by

$$a_A := \mathrm{Var}(Z_a \mid \mathbf{A}, \boldsymbol{\sigma}, \mathcal{L}) = \alpha_L(i; \mathbf{A}),$$
$$b_A := \mathrm{Var}(V_a \mid \mathbf{A}, \boldsymbol{\sigma}, \mathcal{L}) = \beta_L(k, r; \mathbf{A}),$$

and

$$c_A := \mathrm{Cov}(Z_a, V_a \mid \mathbf{A}, \boldsymbol{\sigma}, \mathcal{L}) = \bar{\Delta}_{i,k}^{(L)} \leq -c\eta\gamma_n^{2L},$$

where the identities follow from the symmetry of $\mathbf{P}$. On the same good event, their variance product satisfies

$$a_A b_A \leq C\eta n\delta_n^{-2L}(1 + \eta\delta_n^L).$$

If $a_A b_A - c_A^2 > 0$, the Gaussian-product MGF in Lemma D.2, applied with a positive Chernoff parameter and optimized at

$$t^* = -\frac{c_A}{a_A b_A - c_A^2},$$

gives

$$\mathbb{P}_{\mathbf{X}}\left(\Delta_{i,k}^{(L)} \geq 0 \mid \mathbf{A}, \boldsymbol{\sigma}, \mathcal{L}\right) \leq \exp\left[-\frac{d}{2}\log\left(\frac{a_A b_A}{a_A b_A - c_A^2}\right)\right].$$

Since $-\log(1-x) \geq x$ for $0 \leq x < 1$,

$$\mathbb{P}_{\mathbf{X}}\left(\Delta_{i,k}^{(L)} \geq 0 \mid \mathbf{A}, \boldsymbol{\sigma}, \mathcal{L}\right) \leq \exp\left(-cd\frac{c_A^2}{a_A b_A}\right) \leq \exp\{-c\mathcal{F}_{n,L}\}.$$

If $a_A b_A - c_A^2 = 0$, then $a_A > 0$ and V_a is almost surely a negative scalar multiple of Z_a because $c_A < 0$. More precisely,

$$V_a = \frac{c_A}{a_A} Z_a \quad \text{almost surely,} \qquad \Delta_{i,k}^{(L)} = \frac{c_A}{a_A}\sum_{a=1}^{d} Z_a^2 < 0 \quad \text{almost surely.}$$

Thus the conditional sign-error probability is zero.

Combining the graph, variance, and feature events gives

$$\mathbb{P}\left(i \in \mathcal{U}, \quad k \neq r, \quad \Delta_{i,k}^{(L)} \geq 0\right) \leq Ce^{-c\mathcal{W}_{n,L}} + Ce^{-c\mathcal{V}_{n,L}} + Ce^{-c\mathcal{F}_{n,L}}.$$

The preceding pairwise bound is uniform over every realized $i \in \mathcal{U}$ and every admissible collection $\mathcal{L}$; labeled nodes have zero error by definition. A union bound over the fixed number $K-1$ of wrong classes, followed by averaging over nodes, proves the expected-error bound. Indeed,

$$\mathbb{E}\Big[\mathrm{err}\left(\widehat{\boldsymbol{\sigma}}^{(L)}, \boldsymbol{\sigma}\right)\Big] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\Big(i \in \mathcal{U}, \widehat{\sigma}_i^{(L)} \neq \sigma_i\Big).$$

Therefore,

$$\mathbb{E}\Big[\mathrm{err}\left(\widehat{\boldsymbol{\sigma}}^{(L)}, \boldsymbol{\sigma}\right)\Big] \leq (K-1)\left(Ce^{-c\mathcal{W}_{n,L}} + Ce^{-c\mathcal{V}_{n,L}} + Ce^{-c\mathcal{F}_{n,L}}\right),$$

and the fixed factor $K-1$ is absorbed into C. If all three exponent terms dominate $\log n$, then the expected misclassification rate is $o(1/n)$; Markov's inequality gives exact recovery.

Remark C.1 (Depth ratios and the crossover in Theorem 4.5). This remark records the derivation underlying the depth-dependence discussion in the main text. For two consecutive depths satisfying the assumptions of the theorem, the graph-margin exponent contracts as

$$\frac{\mathcal{W}_{n,L+1}}{\mathcal{W}_{n,L}} = \gamma_n^4 < 1,$$

while the feature-noise exponent satisfies

$$\frac{\mathcal{F}_{n,L+1}}{\mathcal{F}_{n,L}} = \delta_n^2\gamma_n^4\,\frac{1+\eta\delta_n^L}{1+\eta\delta_n^{L+1}},$$

which simplifies to $(1+o(1))\delta_n\gamma_n^4$ when $\eta\delta_n^L \gg 1$. Hence $\mathcal{F}_{n,L}$ increases across layers whenever $\delta_n\gamma_n^4 > 1$. The two exponents are comparable when $\mathcal{F}_{n,L} \asymp \mathcal{W}_{n,L}$, i.e. when

$$\frac{\mathcal{F}_{n,L}}{\mathcal{W}_{n,L}} = \frac{d\delta_n^{2L-1}}{n(1+\eta\delta_n^L)} \asymp 1,$$

which in the regime $\eta\delta_n^L \gg 1$ reduces to the crossover condition $d\delta_n^{L-1} \asymp \eta n$ stated in the main text. In the logarithmic-degree regime $p = a\log n/n$, $q = b\log n/n$ with fixed $a > b > 0$, we have $\delta_n = \Theta(\log n)$ and $\gamma_n = \Theta(1)$, so this crossover becomes $d(\log n)^{L-1} \asymp \eta n$. Before the crossover, additional depth mainly averages out feature noise; after it, the graph term $\mathcal{W}_{n,L}$ becomes the bottleneck and further depth worsens the bound by reducing community separation. The graph-induced variance term $\mathcal{V}_{n,L}$ imposes a separate constraint: if it is smaller than both $\mathcal{W}_{n,L}$ and $\mathcal{F}_{n,L}$, the present bound does not distinguish this crossover.

Remark C.2 (Why the sharp $L = 1$ results are not recovered). The single-layer proofs of Theorems 4.1 and 4.3 exploit the length-two structure of $\mathbf{A}\mathbf{A}^\top$: after conditioning on the edges incident to the node being classified, the remaining graph margin decomposes into contributions from independent edges, giving an exact conditional MGF and preserving the sharp graph large-deviation exponent. For $L > 1$ the score instead depends on length-$2L$ walks, which may share edges and vertices and traverse the same edge more than once; the resulting walk monomials are strongly dependent, and the number of overlap patterns grows quickly with L. Even after conditioning on the incident edges, the graph-side conditional MGF no longer factorizes edgewise. The Gaussian feature MGF conditional on the graph still applies exactly; the loss is entirely on the graph side. We therefore bound the walk polynomials via a general concentration inequality for fixed-degree Bernoulli polynomials (Lemma D.4), controlling the overlapping paths uniformly over walk patterns. This method cannot exploit the exact cancellations in the signed class difference or the sharp edgewise calculation available at $L = 1$, and it introduces the conservative constraint $\mathcal{V}_{n,L} \le \delta_n^{1/2}$. In particular, at $L = 1$ with $p, q = \Theta(\log n/n)$, this generic bound gives only $\mathcal{V}_{n,1} = O(\sqrt{\log n})$, which cannot recover the sharp exact-recovery threshold or the optimal misclassification rate obtained by the dedicated length-two argument.

D Auxiliary Lemmas

Lemma D.1. *In the symmetric K-community stochastic block model with* uniform *community assignment, assume $2 \le K = O(1)$ 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, let $|\mathcal{V}_k|$ denote the number of nodes in community k, and set

$$\varepsilon_n := \frac{1}{\log n}.$$

Then there exists an absolute constant $c > 0$ such that, with probability at least

$$1 - 2K\exp\left(-\frac{n}{3K(\log n)^2}\right) - 2\exp\left(-c\frac{n^2}{(\log n)^2}\frac{p + (K-1)q}{K}\right),$$

the following statements hold simultaneously.

1. $$\max_{k\in[K]} \left||\mathcal{V}_k| - \frac{n}{K}\right| \le \varepsilon_n \frac{n}{K}.$$
   In particular, $|\mathcal{V}_k| = \frac{n}{K}(1 + o(1))$ for every $k \in [K]$.
2. $$\mathbf{D}_{ii} = \left(1 + O\left(\frac{1}{\log n}\right)\right)\left[\frac{n}{K}p + n\left(1 - \frac{1}{K}\right)q\right].$$

Proof. Since each node independently chooses a community with probability $1/K$, $|\mathcal{V}_k| \sim \mathrm{Bin}(n, 1/K)$. The Chernoff bound with relative deviation ε_n yields

$$\mathbb{P}\left(\left||\mathcal{V}_k| - \frac{n}{K}\right| \ge \varepsilon_n \frac{n}{K}\right) \le 2\exp\left(-\frac{\varepsilon_n^2 n}{3K}\right) = 2\exp\left(-\frac{n}{3K(\log n)^2}\right).$$

Define the community-size event

$$\mathcal{E}^{(2)}_{\text{size}} := \left\{\max_{k\in[K]} \left||\mathcal{V}_k| - \frac{n}{K}\right| \le \varepsilon_n \frac{n}{K}\right\}.$$

A union bound gives

$$\mathbb{P}\left(\left(\mathcal{E}^{(2)}_{\text{size}}\right)^c\right) \le 2K\exp\left(-\frac{n}{3K(\log n)^2}\right). \tag{8}$$

This proves 1.

Let

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

Conditional on the community assignment $\boldsymbol{\sigma}$, the variables $\{\mathbf{A}_{ij}\}_{i<j}$ are independent Bernoulli variables with parameters in $\{p, q\}$. Hence

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

Set

$$t := \varepsilon_n \mathbb{E}[S \mid \boldsymbol{\sigma}].$$

Bernstein's inequality and $\varepsilon_n < 1$ give, for some absolute constant $c > 0$,

$$\mathbb{P}(|S - \mathbb{E}[S \mid \boldsymbol{\sigma}]| \ge \varepsilon_n \mathbb{E}[S \mid \boldsymbol{\sigma}] \mid \boldsymbol{\sigma}) \le 2 \exp\left(-c\varepsilon_n^2 \mathbb{E}[S \mid \boldsymbol{\sigma}]\right). \tag{9}$$

On $\mathcal{E}^{(2)}_{\mathrm{size}}$, using $\sum_{k=1}^{K} (|\mathcal{V}_k| - n/K) = 0$, we have

$$\begin{aligned}
\sum_{k=1}^{K} \binom{|\mathcal{V}_k|}{2} &= \frac{1}{2} \sum_{k=1}^{K} |\mathcal{V}_k|^2 - \frac{n}{2} \\
&= \frac{n^2}{2K} + \frac{1}{2} \sum_{k=1}^{K} \left(|\mathcal{V}_k| - \frac{n}{K}\right)^2 - \frac{n}{2} \\
&= \frac{n^2}{2K} \left(1 + O\left(\varepsilon_n^2 + \frac{1}{n}\right)\right).
\end{aligned}$$

Consequently, uniformly over all community assignments in $\mathcal{E}^{(2)}_{\mathrm{size}}$,

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

Substituting this estimate into (9) gives, uniformly over the same assignments,

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

after adjusting $c > 0$.

Since $\mathbf{D}_{ii} = 2S/n$, the same relative-deviation estimate holds for $\mathbf{D}_{ii}$. Moreover, (10) gives

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

Combining the last two estimates and using $\varepsilon_n = 1/\log n$ proves 2. Finally, taking a union bound with (8) gives the stated simultaneous probability. □

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

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

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,

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

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 D.3. *Consider the logarithmic-degree regime*

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

where a, b are fixed, $K = O(1)$, and $\eta \log n \to \infty$. Conditional on the community assignment, suppose that each $\mathcal{L}_k$ is sampled independently of $\mathbf{A}$ as in the model. Recall from the main text that

$$\alpha(i) = (\mathbf{A}\mathbf{A}^\top)_{ii}, \qquad \beta(k, r) = \big\|\mathbf{A}\big(\mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r}\big)\big\|_2^2,$$

where the expression for $\beta(k, r)$ uses the symmetry of $\mathbf{A}$. There is a constant $C > 0$, depending only on a, b, and K, such that

$$\mathcal{E}_{\mathrm{str}}^{\mathrm{ex}} := \left\{ \max_{i \in [n]} \alpha(i) \le C \log n, \quad \max_{\substack{k, r \in [K] \\ k \neq r}} \beta(k, r) \le C\eta^2 n \log^2 n \right\}$$

satisfies

$$\mathbb{P}\big(\big(\mathcal{E}_{\mathrm{str}}^{\mathrm{ex}}\big)^c\big) = o(1).$$

Proof. Since $\alpha(i) = (\mathbf{A}\mathbf{A}^\top)_{ii} = \deg(i)$ and $p > q$, each $\alpha(i)$ is stochastically dominated by $\mathrm{Bin}(n-1, p)$. A Chernoff upper-tail bound with a sufficiently large constant C, followed by a union bound over $i \in [n]$, gives

$$\mathbb{P}\left(\max_{i\in[n]} \alpha(i) > C\log n\right) = o(1).$$

It remains to control $\beta(k, r)$. Fix $k \neq r$ and write

$$\boldsymbol{s}^{(k,r)} := \mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r}, \qquad \mathbf{M} := \mathbb{E}[\mathbf{A} \mid \boldsymbol{\sigma}, \mathcal{L}] = \mathbb{E}[\mathbf{A} \mid \boldsymbol{\sigma}], \qquad \mathbf{W}_A := \mathbf{A} - \mathbf{M}.$$

The second equality uses the conditional independence of $\mathcal{L}$ and $\mathbf{A}$ given $\boldsymbol{\sigma}$. Since

$$\left\|\boldsymbol{s}^{(k,r)}\right\|_1 = \left\|\boldsymbol{s}^{(k,r)}\right\|_2^2 = |\mathcal{L}_k| + |\mathcal{L}_r| \le 2\eta n$$

and $\max\{p, q\} = O(\log n/n)$, every coordinate of $\mathbf{M}\boldsymbol{s}^{(k,r)}$ has absolute value at most $2\eta n \max\{p, q\}$. Hence

$$\left\|\mathbf{M}\boldsymbol{s}^{(k,r)}\right\|_2^2 \le C\eta^2 n \log^2 n.$$

Moreover, conditional on $(\boldsymbol{\sigma}, \mathcal{L})$, the centered variables in each fixed row of $\mathbf{W}_A$ correspond to distinct edges. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{W}_A\boldsymbol{s}^{(k,r)}\right\|_2^2 \,\middle|\, \boldsymbol{\sigma}, \mathcal{L}\right] &= \sum_{v=1}^{n}\sum_{\substack{u=1\\u\neq v}}^{n} \left(s_u^{(k,r)}\right)^2 \mathrm{Var}(\mathbf{A}_{vu} \mid \boldsymbol{\sigma}) \\ &\le n\max\{p, q\}\left\|\boldsymbol{s}^{(k,r)}\right\|_2^2 \\ &\le C\eta n \log n.\end{aligned}$$

Markov's inequality and $\eta \log n \to \infty$ now imply

$$\mathbb{P}\left(\left\|\mathbf{W}_A\boldsymbol{s}^{(k,r)}\right\|_2^2 > \eta^2 n\log^2 n \,\middle|\, \boldsymbol{\sigma}, \mathcal{L}\right) \le \frac{C}{\eta\log n} = o(1).$$

Finally,

$$\begin{aligned}\beta(k, r) &= \left\|\mathbf{A}\boldsymbol{s}^{(k,r)}\right\|_2^2 \\ &\le 2\left\|\mathbf{M}\boldsymbol{s}^{(k,r)}\right\|_2^2 + 2\left\|\mathbf{W}_A\boldsymbol{s}^{(k,r)}\right\|_2^2.\end{aligned}$$

Here the equality uses the symmetry of $\mathbf{A}$. The conditional bounds above are uniform over the realized $(\boldsymbol{\sigma}, \mathcal{L})$. Since there are only $K(K-1) = O(1)$ ordered class pairs, a union bound completes the proof. □

Lemma D.4. *Let $\{\xi_e : e \in \mathcal{E}_0\}$ be independent Bernoulli random variables, let $Q \ge 1$ be a fixed integer, and let*

$$f(\boldsymbol{\xi}) = \sum_{\substack{S\subseteq\mathcal{E}_0\\|S|\le Q}} w_S \prod_{e\in S} \xi_e$$

be a multilinear polynomial with nonnegative coefficients $w_S \ge 0$. For $r = 0, \ldots, Q$, define

$$\mu_r(f) := \max_{\substack{F\subseteq\mathcal{E}_0\\|F|=r}} \sum_{S\supseteq F} w_S \prod_{e\in S\setminus F} \mathbb{E}[\xi_e].$$

In particular,

$$\mu_0(f) = \mathbb{E}[f(\boldsymbol{\xi})].$$

Then there exists a constant $c_Q > 0$, depending only on Q, such that for every $\lambda > 0$,

$$\mathbb{P}(|f(\boldsymbol{\xi}) - \mathbb{E}f(\boldsymbol{\xi})| \ge \lambda) \le 2e^2 \exp\left[-c_Q \min_{1\le r\le Q}\left\{\frac{\lambda^2}{\mu_0(f)\mu_r(f)}, \left(\frac{\lambda}{\mu_r(f)}\right)^{1/r}\right\}\right].$$

Here and below, a quotient with zero denominator is interpreted as $+\infty$; the case in which f is almost surely constant is trivial.

Remark D.5. Lemma D.4 is the specialization of the polynomial concentration inequality of Schudy & Sviridenko (2012) to independent Bernoulli variables and nonnegative multilinear polynomials of fixed degree. It provides both upper- and lower-tail control and therefore replaces the generally invalid upper-tail version of a dependency-graph Bernstein inequality. Bernoulli variables are moment bounded with parameter one in their terminology, and the remaining degree-dependent factors are absorbed into c_Q because Q is fixed.

Lemma D.6. *Consider the symmetric K-community SBM with $2 \le K = O(1)$. Assume*

$$p, q = o(1), \qquad p, q \gg \frac{1}{n},$$

and assume that p and q are comparable, meaning that there exists a constant $c_{\mathrm{pq}} \in (0,1)$ such that

$$c_{\mathrm{pq}} \le \frac{p}{q} \le c_{\mathrm{pq}}^{-1}.$$

Define

$$\delta_n := \frac{n}{K}\big(p + (K-1)q\big), \qquad \gamma_n := \frac{|p-q|}{p + (K-1)q}.$$

Fix an integer $L \ge 1$ that does not depend on n, and set

$$m := 2L.$$

Let $0 < \eta \le 1$, possibly depending on n. Assume

$$\delta_n \to \infty, \qquad \delta_n^m = o(n),$$

and the signal-dominance conditions

$$\delta_n \gamma_n^m \to \infty, \qquad \eta n \gamma_n^m \to \infty, \qquad n\gamma_n^{2m} \to \infty.$$

Conditional on $\boldsymbol{\sigma}$, let the sets $\mathcal{L}_k$, $k \in [K]$, be chosen independently of $\mathbf{A}$ and satisfy

$$|\mathcal{L}_k| = \lfloor \eta |\mathcal{V}_k| \rfloor, \qquad \mathcal{L}_k \subseteq \mathcal{V}_k.$$

All assertions below hold uniformly over every collection of subsets with these cardinalities. In particular, they hold when the sets are sampled uniformly conditional on $\boldsymbol{\sigma}$, as in the semi-supervised model of the main text.

Fix $i \in [n]$ and $k \in [K]$, and write $r = \sigma(i)$. Define

$$D_{i,k}^{(L)} := \sum_{\ell \in \mathcal{L}_r} (\mathbf{A}^m)_{i\ell} - \sum_{\ell \in \mathcal{L}_k} (\mathbf{A}^m)_{i\ell}.$$

Define

$$\mathcal{W}_{n,L} := \eta \delta_n \gamma_n^{2m} = \eta \delta_n \gamma_n^{4L}.$$

Then there exist constants $C, c > 0$, depending only on K, L and c_{pq}, such that

$$\mathbb{P}_{\mathbf{A},\boldsymbol{\sigma},\mathcal{L}}\left(i \in \mathcal{U}, \quad k \ne r, \quad D_{i,k}^{(L)} \le c\,\eta \delta_n^m \gamma_n^m \right) \le C \exp\{-c\mathcal{W}_{n,L}\} + \mathcal{R}_{n,L}^{\mathrm{size}},$$

where

$$\mathcal{R}_{n,L}^{\mathrm{size}} := 2K \exp\{-cn\gamma_n^{2m}\} = 2K \exp\{-cn\gamma_n^{4L}\}.$$

Let

$$\mathbf{D}_{ii} = \frac{1}{n} \sum_{u,v=1}^{n} \mathbf{A}_{uv}, \qquad \mathbf{P} := \mathbf{D}^{-1}\mathbf{A}.$$

On the event $\sum_{u,v=1}^n \mathbf{A}_{uv} = 0$, set $\mathbf{P} = \mathbf{0}$. There is also a constant $c_D > 0$ such that

$$\mathbb{P}\left(i \in \mathcal{U}, \quad k \ne r, \quad \sum_{\ell \in \mathcal{L}_r} (\mathbf{P}^m)_{i\ell} - \sum_{\ell \in \mathcal{L}_k} (\mathbf{P}^m)_{i\ell} \le c\,\eta\gamma_n^m \right) \le C \exp\{-c\mathcal{W}_{n,L}\} + \mathcal{R}_{n,L},$$

where

$$\mathcal{R}_{n,L} := 2K \exp\{-cn\gamma_n^{4L}\} + 2\exp\left(-c_D \frac{n\delta_n}{(\log n)^2} \right).$$

Proof. For $s \in [K]$, define

$$T_{i,s}^{(m)} := \sum_{\ell \in \mathcal{L}_s} (\mathbf{A}^m)_{i\ell}.$$

Then

$$D_{i,k}^{(L)} = T_{i,r}^{(m)} - T_{i,k}^{(m)}.$$

Choose a sufficiently small constant $c_{\mathrm{b}} > 0$, depending only on K and L, and define

$$\mathcal{E}_{\mathrm{size}}^{(L)} := \left\{ \max_{s \in [K]} \left| |\mathcal{V}_s| - \frac{n}{K} \right| \le c_{\mathrm{b}} n \gamma_n^m \right\}.$$

By the multinomial Chernoff bound,

$$\mathbb{P}\Big((\mathcal{E}_{\mathrm{size}}^{(L)})^c \Big) \le 2K \exp\{-cn\gamma_n^{2m}\}.$$

On $\mathcal{E}_{\mathrm{size}}^{(L)}$,

$$\left| |\mathcal{L}_s| - \eta \frac{n}{K} \right| \le c_{\mathrm{b}} \eta n \gamma_n^m + 1, \qquad s \in [K].$$

The assumption $\eta n \gamma_n^m \to \infty$ makes the rounding error negligible relative to the signal scale $\eta n \gamma_n^m$.

We next compare the population scores. Let

$$\mathbf{Q} := \mathbb{E}[\mathbf{A} \mid \boldsymbol{\sigma}].$$

Let $\mathbf{\Pi} \in \{0,1\}^{n \times K}$ be the membership matrix and set

$$\mathbf{N} := \mathbf{\Pi}^\top \mathbf{\Pi} = \mathrm{diag}(|\mathcal{V}_1|, \dots, |\mathcal{V}_K|).$$

Writing $\mathbf{B}$ for the SBM connectivity matrix, define

$$b_0 := p + (K-1)q, \qquad \mathbf{H}_0 := \frac{\mathbf{B}}{b_0}, \qquad \mathbf{H} := \frac{\mathbf{NB}}{\delta_n}.$$

Since $\delta_n = (n/K) b_0$, the size event gives

$$\|\mathbf{H} - \mathbf{H}_0\|_{\mathrm{op}} \le C c_{\mathrm{b}} \gamma_n^m.$$

It is convenient first to include the diagonal entries of the population matrix and write

$$\widetilde{\mathbf{Q}} := \mathbf{\Pi B \Pi}^\top.$$

Then

$$\widetilde{\mathbf{Q}}^m = \mathbf{\Pi B} (\mathbf{NB})^{m-1} \mathbf{\Pi}^\top,$$

and, for $i \in \mathcal{V}_r$,

$$\delta_n^{-m} \sum_{\ell \in \mathcal{L}_s} (\widetilde{\mathbf{Q}}^m)_{i\ell} = \frac{K|\mathcal{L}_s|}{n} \boldsymbol{e}_r^\top \mathbf{H}_0 \mathbf{H}^{m-1} \boldsymbol{e}_s.$$

The balanced matrix $\mathbf{H}_0$ has eigenvalue 1 in the all-one direction and eigenvalue

$$\vartheta := \frac{p-q}{p+(K-1)q}$$

on the community-contrast subspace. Hence, since m is even,

$$\boldsymbol{e}_r^\top \mathbf{H}_0^m \boldsymbol{e}_r - \boldsymbol{e}_r^\top \mathbf{H}_0^m \boldsymbol{e}_k = \vartheta^m = \gamma_n^m.$$

The fixed-dimensional telescoping identity for matrix powers, together with $K|\mathcal{L}_s|/n = \eta + O(c_{\mathrm{b}} \eta \gamma_n^m) + O(1/n)$, now gives

$$\sum_{\ell \in \mathcal{L}_r} (\widetilde{\mathbf{Q}}^m)_{i\ell} - \sum_{\ell \in \mathcal{L}_k} (\widetilde{\mathbf{Q}}^m)_{i\ell} \ge c \eta \delta_n^m \gamma_n^m,$$

provided c_{b} is sufficiently small. Here the rounding term is lower order because $\eta n \gamma_n^m \to \infty$.

The zero diagonal of the adjacency matrix means that

$$\mathbf{Q} = \widetilde{\mathbf{Q}} - p\mathbf{I}_n.$$

On $\mathcal{E}^{(L)}_{\text{size}}$, the row sums of $\widetilde{\mathbf{Q}}$ are $O(\delta_n)$. The binomial expansion and $p \le C\delta_n/n$ therefore imply, for every s,

$$\sum_{\ell \in \mathcal{L}_s} \left| (\widetilde{\mathbf{Q}}^m)_{i\ell} - (\mathbf{Q}^m)_{i\ell} \right| \le C_L \frac{\delta_n^m}{n} = o(\eta \delta_n^m \gamma_n^m)\,.$$

Consequently,

$$\sum_{\ell \in \mathcal{L}_r} (\mathbf{Q}^m)_{i\ell} - \sum_{\ell \in \mathcal{L}_k} (\mathbf{Q}^m)_{i\ell} \ge c\,\eta \delta_n^m \gamma_n^m.$$

The matrix power $\mathbf{Q}^m$ treats repeated occurrences of the same edge as independent occurrences. We therefore bound the correction caused by walks using fewer than m distinct edges. A length-m walk from the fixed unlabeled node i to a labeled endpoint and using $h < m$ distinct edges has at most $C_L \eta n^h$ realizations, and each realization has expectation at most

$$C_L \left(\frac{\delta_n}{n}\right)^h.$$

Thus the total absolute expected repeated-edge contribution is at most

$$C_L \sum_{h=1}^{m-1} \eta n^h \left(\frac{\delta_n}{n}\right)^h \le C_L \eta \delta_n^{m-1}.$$

Since $\delta_n \gamma_n^m \to \infty$, this is

$$o(\eta \delta_n^m \gamma_n^m)\,.$$

Consequently, on $\mathcal{E}^{(L)}_{\text{size}}$,

$$\mathbb{E}[D^{(L)}_{i,k} \mid \boldsymbol{\sigma}, \mathcal{L}] \ge c\,\eta \delta_n^m \gamma_n^m.$$

We now bound the fluctuations of the walk polynomials. Conditional on $(\boldsymbol{\sigma}, \mathcal{L})$, index the potential unordered edges by

$$\mathcal{E}_0 := \big\{\{u, v\} : 1 \le u < v \le n\big\}, \qquad \xi_{\{u,v\}} := A_{uv} = A_{vu}.$$

The variables $\{\xi_e : e \in \mathcal{E}_0\}$ are mutually independent Bernoulli variables, with success probability p for a within-community edge and q for a between-community edge. Moreover, $\xi_e^r = \xi_e$ for every integer $r \ge 1$. Thus, the product of adjacency entries along a walk equals the product of the ξ_e's over the distinct edges used by that walk.

For fixed s, group all walk sequences yielding the same set of distinct edges. Then $T^{(m)}_{i,s}$ is a nonnegative multilinear polynomial of degree at most m in the independent edge variables.

Its expectation satisfies

$$\mu_0(T^{(m)}_{i,s}) \le C_L \eta \delta_n^m.$$

For $1 \le a \le m$, fix any set F of a distinct edges. Every monomial contributing to the corresponding derivative is the edge set of a connected walk containing F. If that monomial contains h distinct edges in total, then at most $h - a$ new vertices remain free after F is fixed. Hence the number of completions is at most $C_L n^{h-a}$, and the expectation of the remaining edge product is at most

$$C_L \left(\frac{\delta_n}{n}\right)^{h-a}.$$

Summing over $a \le h \le m$ gives

$$\mu_a(T^{(m)}_{i,s}) \le C_L \delta_n^{m-a}.$$

Set

$$\lambda := c_0 \eta \delta_n^m \gamma_n^m$$

with $c_0 > 0$ sufficiently small. Lemma D.4 gives

$$\mathbb{P}\Big(\Big|T_{i,s}^{(m)} - \mathbb{E}[T_{i,s}^{(m)} \mid \boldsymbol{\sigma}, \mathcal{L}]\Big| \geq \lambda \;\Big|\; \boldsymbol{\sigma}, \mathcal{L}\Big)$$
$$\leq 2e^2 \exp\Big[-c_L \min_{1\leq a\leq m} \Big\{\eta\delta_n^a \gamma_n^{2m}, \delta_n(\eta\gamma_n^m)^{1/a}\Big\}\Big].$$

Because $0 < \eta\gamma_n^m \leq 1$ and $\delta_n \geq 1$, both expressions in the minimum are minimized at $a = 1$, up to constants. Moreover,

$$\eta\delta_n\gamma_n^{2m} \leq \eta\delta_n\gamma_n^m.$$

Therefore

$$\mathbb{P}\Big(\Big|T_{i,s}^{(m)} - \mathbb{E}[T_{i,s}^{(m)} \mid \boldsymbol{\sigma}, \mathcal{L}]\Big| \geq \lambda \;\Big|\; \boldsymbol{\sigma}, \mathcal{L}\Big) \leq C\exp\{-c\mathcal{W}_{n,L}\}.$$

Applying this bound to $s = r$ and $s = k$, and combining it with the population separation above, proves the unnormalized score-margin bound.

It remains to pass to the normalized matrix. Write

$$|\mathcal{V}_s| = \frac{n}{K} + x_s, \qquad \sum_{s=1}^{K} x_s = 0.$$

On $\mathcal{E}_{\text{size}}^{(L)}$, $\sum_s x_s^2 \leq Cc_{\text{b}}^2 n^2 \gamma_n^{2m}$. It follows directly from the conditional mean of the total edge count that

$$\mathbb{E}[\mathbf{D}_{ii} \mid \boldsymbol{\sigma}] = \delta_n\left(1 + O(c_{\text{b}}^2\gamma_n^{2m}) + O\left(\frac{1}{n}\right)\right).$$

Bernstein's inequality around this conditional mean gives

$$\mathbf{D}_{ii} = \delta_n\left(1 + O(c_{\text{b}}^2\gamma_n^{2m}) + O\left(\frac{1}{\log n}\right) + O\left(\frac{1}{n}\right)\right)$$

with failure probability at most

$$2\exp\left(-c_D\frac{n\delta_n}{(\log n)^2}\right).$$

Choosing c_{b} sufficiently small, there are constants $0 < c_{\deg} < C_{\deg} < \infty$ such that, on this event,

$$c_{\deg}\delta_n \leq \mathbf{D}_{ii} \leq C_{\deg}\delta_n.$$

Since all diagonal entries of $\mathbf{D}$ are equal, there is a scalar $\rho_{n,m} \in [c, C]$ for which

$$\mathbf{P}^m = \mathbf{D}^{-m}\mathbf{A}^m = \rho_{n,m}\delta_n^{-m}\mathbf{A}^m.$$

Multiplication by this common positive factor preserves the score ordering and changes the margin only by a constant factor. This yields the normalized statement. □

Lemma D.7. *Under the assumptions of Lemma D.6, define*

$$B_{n,L} := \eta n\delta_n^L + \eta^2 n\delta_n^{2L},$$

and

$$\widetilde{\mathcal{V}}_{n,L} := \min_{1\leq a\leq 2L}\left\{\frac{B_{n,L}}{\delta_n^{2L-a}}, \left(\frac{B_{n,L}}{\delta_n^{2L-a}}\right)^{1/a}\right\}.$$

Set

$$\mathcal{V}_{n,L} := \min\left\{\delta_n^{1/2}, \widetilde{\mathcal{V}}_{n,L}\right\}.$$

Fix $i \in [n]$ and $k \in [K]$, and write $r = \sigma(i)$. Define

$$\boldsymbol{v}_{k,r} := \mathbf{1}_{\mathcal{L}_k} - \mathbf{1}_{\mathcal{L}_r},$$
$$\alpha_L(i; \mathbf{A}) := \boldsymbol{e}_i^\top \mathbf{P}^{2L}\boldsymbol{e}_i, \qquad \beta_L(k, r; \mathbf{A}) := \boldsymbol{v}_{k,r}^\top \mathbf{P}^{2L}\boldsymbol{v}_{k,r}.$$

Then there exist constants $C, c > 0$ such that

$$\mathbb{P}\Big(i \in \mathcal{U}, \quad k \neq r, \quad \Big[\alpha_L(i; \mathbf{A}) > C\delta_n^{-L} \text{ or } \beta_L(k, r; \mathbf{A}) > C\eta n\delta_n^{-L}(1 + \eta\delta_n^L)\Big]\Big)$$
$$\leq C\exp\{-c\mathcal{V}_{n,L}\} + \mathcal{R}_{n,L}.$$

Proof. We work with unnormalized walk polynomials.

For fixed L, there are only C_L abstract length-$2L$ walk types, and the coefficient of any distinct-edge monomial is at most C_L. Moreover, if a specified set F of a edges is contained in a connected support with h distinct edges, then the edges outside F introduce at most $h-a$ new vertices. Thus, after F is fixed, there are at most $C_L n^{h-a}$ possible completions. We use these facts together with the uniform bound $\mathbb{P}(A_{uv}=1 \mid \boldsymbol{\sigma}) \le C\delta_n/n$.

We first bound the diagonal term. Let

$$C_i^{(2L)} := (\mathbf{A}^{2L})_{ii}.$$

It is a nonnegative multilinear polynomial of degree at most $2L$ after walks with the same distinct edge set are grouped together.

To bound its expectation, consider a rooted closed walk with h distinct edges. If its connected support is a tree, every support edge is a bridge and must be traversed at least twice, so $h \le L$. Such supports contribute at most $C_L\delta_n^h$. If the support contains a cycle, it has at most h vertices and hence, with the root fixed, at most $h-1$ free vertices. The cyclic contribution is therefore at most $C_L\delta_n^h/n$. Summing over $h \le 2L$ and using $\delta_n^{2L} = o(n)$ gives

$$\mathbb{E}[C_i^{(2L)} \mid \boldsymbol{\sigma}] \le C_L\left(\delta_n^L + \frac{\delta_n^{2L}}{n}\right) \le C_L\delta_n^L.$$

Fix a set F of a distinct edges and consider a monomial with full edge set $S \supseteq F$. Put $u := |S \setminus F|$, and expose the vertices of F together with the fixed root i. The remaining u edges introduce at most u new vertices.

If they introduce at most $u-1$ new vertices, the corresponding derivative contribution has an extra factor $1/n$ and is bounded by

$$C_L\frac{\delta_n^u}{n} \le C_L\frac{\delta_n^{2L}}{n} = o(1).$$

If they introduce exactly u new vertices, every edge in $S \setminus F$ is a bridge of the full edge graph. A closed walk traverses each such bridge at least twice, while it traverses each of the a fixed edges at least once. The length constraint therefore gives

$$a + 2u \le 2L, \qquad u \le L - \left\lceil \frac{a}{2} \right\rceil.$$

There are only C_L possible walk types because L is fixed. Summing both classes of completions gives

$$\mu_a(C_i^{(2L)}) \le C_L \max\left\{1, \delta_n^{L-\lceil a/2\rceil}\right\}, \qquad 1 \le a \le 2L.$$

Apply Lemma D.4 with $\lambda = C_L\delta_n^L$. For every $1 \le a \le 2L$,

$$\frac{\lambda^2}{\mu_0(C_i^{(2L)})\mu_a(C_i^{(2L)})} \ge c_L\delta_n^{\min\{\lceil a/2\rceil, L\}},$$

and

$$\left(\frac{\lambda}{\mu_a(C_i^{(2L)})}\right)^{1/a} \ge c_L\delta_n^{1/2}.$$

Hence

$$\mathbb{P}_{\mathbf{A}}\left(C_i^{(2L)} > C\delta_n^L \,\middle|\, \boldsymbol{\sigma}\right) \le C\exp\{-c\delta_n^{1/2}\}.$$

We next bound the labeled-to-labeled term. Since $\mathbf{P}^{2L}$ has nonnegative entries,

$$\beta_L(k, r; \mathbf{A}) \le (\mathbf{1}_{\mathcal{L}_r} + \mathbf{1}_{\mathcal{L}_k})^\top \mathbf{P}^{2L}(\mathbf{1}_{\mathcal{L}_r} + \mathbf{1}_{\mathcal{L}_k}).$$

Define the corresponding unnormalized polynomial

$$B_{r,k}^{(2L)} := \sum_{u,v \in \mathcal{L}_r \cup \mathcal{L}_k} (\mathbf{A}^{2L})_{uv}.$$

For the terms with $u = v$, the rooted closed-walk estimate above and $|\mathcal{L}_r \cup \mathcal{L}_k| \le \eta n$ give a contribution at most $C_L \eta n \delta_n^L$. For $u \neq v$, a connected support with h edges has at most $h+1$ vertices. Choosing its two ordered endpoints from the labeled set and then its remaining vertices gives at most $C_L \eta^2 n^{h+1}$ embeddings. Multiplication by the edge-probability factor $(C\delta_n/n)^h$ and summation over $h \le 2L$ give a contribution at most $C_L \eta^2 n \delta_n^{2L}$. Consequently,

$$\mathbb{E}[B_{r,k}^{(2L)} \mid \boldsymbol{\sigma}, \mathcal{L}] \le C_L B_{n,L}.$$

For $1 \le a \le 2L$, fix any a distinct edges. The fixed edge set anchors every connected walk monomial containing it. If the full monomial has h distinct edges, there are at most $C_L n^{h-a}$ completions after the fixed edges are specified; ignoring the labeled-endpoint restrictions only enlarges this number. Therefore,

$$\mu_a(B_{r,k}^{(2L)}) \le C_L \delta_n^{2L-a}.$$

Applying Lemma D.4 with $\lambda = C_L B_{n,L}$ yields

$$\mathbb{P}_{\mathbf{A}}\Big(B_{r,k}^{(2L)} > C B_{n,L} \,\Big|\, \boldsymbol{\sigma}, \mathcal{L}\Big) \le C \exp\{-c\widetilde{\mathcal{V}}_{n,L}\}.$$

It remains to normalize these bounds. On the degree event used in Lemma D.6, there is a scalar $\rho_{n,2L} \in [c, C]$ such that

$$\mathbf{P}^{2L} = \rho_{n,2L} \delta_n^{-2L} \mathbf{A}^{2L}.$$

Consequently,

$$\alpha_L(i; \mathbf{A}) \le C\delta_n^{-2L} \delta_n^L = C\delta_n^{-L},$$

and

$$\beta_L(k, r; \mathbf{A}) \le C\delta_n^{-2L} B_{n,L} = C\eta n \delta_n^{-L} + C\eta^2 n.$$

The latter expression equals

$$C\eta n \delta_n^{-L}(1 + \eta \delta_n^L).$$

The two conditional polynomial bounds are uniform in $\boldsymbol{\sigma}$ and in every admissible collection $\mathcal{L}$. Averaging them and combining them with the community-size and degree-normalization events proves the lemma. □