A Spectral Theory of Normalized Corrected GNN Propagation

Anonymous authors
Paper under double-blind review

Abstract

We develop a spectral theory for *normalized corrected GNN propagation*. The object of study is the symmetric normalized adjacency with its degree-stationary component removed, matching the normalization used by standard GCN-style models while isolating the stationary direction most directly tied to oversmoothing. The central theoretical question is whether this corrected normalized operator preserves class-discriminative signal after many propagation layers. Our main result is a high-probability exact-recovery theorem for the binary Contextual Stochastic Block Model after $k = O(\log n)$ propagation steps in the dense polylogarithmic regime $p \geq C \log^B n/n$, for any fixed $B > 4$, under explicit graph-signal and feature-SNR conditions. We also establish a multi-class partial recovery theorem showing contraction toward class centers for most nodes. Synthetic and real node-classification experiments are included as empirical checks of the theory's predicted dependence on depth, graph signal, and feature noise.

1 Introduction

Deep GNNs repeatedly apply graph propagation operators to node features (Scarselli et al., 2009; Kipf & Welling, 2017; Hamilton et al., 2017; Wu et al., 2019; Gasteiger et al., 2019). For GCN-style architectures, the basic linear backbone is the symmetric normalized adjacency $D^{-1/2}AD^{-1/2}$ (Kipf & Welling, 2017; Chen et al., 2020; Wu et al., 2019). This normalization is essential in practice, but it also creates a precise spectral mechanism for oversmoothing: repeated propagation amplifies the degree-stationary component while attenuating directions that separate classes (Li et al., 2018; Oono & Suzuki, 2020; Keriven, 2022; Rusch et al., 2023; Jin & Zhu, 2025). Thus a central theoretical question is not only whether oversmoothing occurs, but whether the normalized propagation operator can be corrected in a way that provably preserves class information at logarithmic depth.

We study this question through the operator

$$\widehat{A} = D^{-1/2}AD^{-1/2} - \frac{1}{\mathbb{1}^\top D \mathbb{1}} D^{1/2}\mathbb{1}\mathbb{1}^\top D^{1/2},$$

which subtracts the degree-stationary rank-one component before propagation. This operator-level viewpoint follows the broader spectral and graph-signal processing perspective on graph filters (Shuman et al., 2013; Balcilar et al., 2021; Chien et al., 2021; Goksu, 2025). This is the normalized counterpart of the corrected convolution analyzed by (Wang et al., 2024) for unnormalized adjacency matrices. The normalized setting has three characteristic features: the stationary vector is degree-dependent and random, each entry contains nonlinear degree factors, and powers of the residual matrix are tracked entrywise in addition to their spectral-norm behavior.

This paper develops a spectral theory of this normalized corrected GNN propagation operator under the Contextual Stochastic Block Model (CSBM) (Deshpande et al., 2018; Lu & Sen, 2020; Baranwal et al., 2021; 2023a; Dalle & Thiran, 2024). The main theorem gives exact recovery by a linear classifier after $k = O(\log n)$ propagation steps when $p \geq C \log^B n/n$ for any fixed $B > 4$, provided the graph signal and feature signal-to-noise ratio exceed explicit thresholds. The theorem characterizes when the stationary correction is sufficient

to preserve separability through logarithmic depth. The proof reduces the normalized operator to a corrected rank-one signal plus a residual and then develops an entrywise Krylov bound for multi-layer residual powers. The key technical device is a degree-truncated atom expansion combined with decorated-walk counting, which separates true graph fluctuations from artifacts introduced by random normalization.

The experiments conducted in this paper examine the predicted dependencies on depth, graph signal, and feature noise. They are consistent with reduced depth-dependent degradation on synthetic CSBM instances and standard node-classification benchmarks. The main contributions are as follows.

- **A normalized corrected propagation model.** We formulate the corrected GCN backbone by subtracting the degree-stationary component of $D^{-1/2}AD^{-1/2}$, giving an explicit operator-level model of spectral correction in normalized GNNs.
- **Binary exact-recovery theorem.** We prove binary CSBM exact recovery after $k = O(\log n)$ normalized corrected propagation steps in the dense polylogarithmic regime $p \geq C \log^B n/n$, with explicit graph-signal and feature-SNR conditions.
- **Entrywise residual-power theory.** We prove worst-node multi-layer bounds for powers of the normalized residual despite random degree normalization, using atom expansions and decorated-walk counting.
- **Multi-class recovery theorem and empirical checks.** We prove partial recovery for fixed L-class CSBMs and include empirical checks of the predicted depth dependence on synthetic and real node-classification datasets.
- **Empirical validation.** Experiments on synthetic CSBM data and standard node-classification benchmarks support the theory and show that normalized spectral correction remains effective in deep GNN regimes.

2 Related Work

Message passing and normalized propagation. GNNs combine graph structure and node features through repeated local aggregation (Scarselli et al., 2009; Kipf & Welling, 2017; Hamilton et al., 2017). Several influential models can be viewed, at least in their linear propagation backbone, as applying normalized graph filters before or between learned feature transformations (Kipf & Welling, 2017; Wu et al., 2019; Gasteiger et al., 2019; Chen et al., 2020). This paper focuses on that propagation backbone rather than on architectural components such as nonlinear activations, attention, or optimization. The specific object of study is the symmetric normalized operator after subtracting its degree-stationary component.

Oversmoothing and its spectral mechanisms. Oversmoothing was identified as a depth-dependent loss of class information in message passing (Li et al., 2018); subsequent theory shows that repeated propagation can drive node representations toward low-dimensional or stationary subspaces (Oono & Suzuki, 2020; Cai & Wang, 2020; Hou et al., 2020; Keriven, 2022; Adam-Day et al., 2024; Rusch et al., 2023). Many methods attempt to mitigate this behavior through residual connections, normalization, edge dropping, diffusion design, or signed/high-frequency components (Xu et al., 2018; Zhao & Akoglu, 2020; Chen et al., 2020; Rong et al., 2020; Chamberlain et al., 2021; Scholkemper et al., 2025; WANG et al., 2026; He et al., 2025). Our contribution is complementary: we do not introduce a new trained architecture, but analyze an explicit correction of the propagation operator and prove when it preserves separability.

Graph spectral filters and correction. Graph signal processing interprets message passing as filtering over the graph spectrum (Shuman et al., 2013). Early spectral GNNs and localized polynomial filters established this operator viewpoint (Bruna et al., 2014; Defferrard et al., 2016; Balcilar et al., 2021). Learnable or adaptive filters, including generalized PageRank and polynomial constructions, have been proposed for heterophily or oversmoothing (Chien et al., 2021; Goksu, 2025). These works motivate the operator-level viewpoint, but most are algorithmic or empirical. Here the filter is fixed by a correction principle: remove the rank-one stationary component and analyze the resulting normalized residual through high-probability spectral and entrywise bounds.

CSBM theory and the closest prior work. The CSBM provides a controlled model in which graph structure and noisy node features jointly determine classification performance (Deshpande et al., 2018; Lu & Sen, 2020; Mehta et al., 2019; Baranwal et al., 2021; Maskey et al., 2022; Baranwal et al., 2023a;b; Maskey et al., 2024; Duranthon & Zdeborová, 2025; Dalle & Thiran, 2024; Ma et al., 2025). The closest result to ours is the corrected convolution theory of (Wang et al., 2024), which proves recovery guarantees for an *unnormalized* corrected adjacency operator. Our setting differs in the normalization used by standard GCNs: $D^{-1/2}AD^{-1/2}$ has a random degree-weighted stationary vector, and its entries contain nonlinear degree denominators. These features require the atom expansion and decorated-walk counting developed below; they are not consequences of the unnormalized analysis.

3 Preliminaries

We use the following notation throughout this paper. For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$. The all-ones vector is denoted by $\mathbb{1}$, and e_i denotes the i-th standard basis vector. For a vector x, let $\|x\|$ and $\|x\|_\infty$ denote its Euclidean and infinity norms. For a matrix M, let

$$\|M\| = \sup_{\|x\|=1} \|Mx\|, \qquad \|M\|_F = \left(\sum_{i,j} M_{ij}^2 \right)^{1/2}$$

denote the operator and Frobenius norms. We also use the induced $\ell_\infty \to \ell_\infty$ matrix norm

$$\|M\|_{\infty\to\infty} := \sup_{\|x\|_\infty=1} \|Mx\|_\infty = \max_i \sum_j |M_{ij}|,$$

that is, the maximum absolute row-sum norm.

Let $G = (V, E)$ be an undirected unweighted graph with $V = [n]$, adjacency matrix A, degrees d_i, degree matrix $D = \mathrm{diag}(d_1, \dots, d_n)$, and average degree

$$d := \frac{2|E|}{n}.$$

We use the standard conventions $(D^{-1/2})_{ii} = 0$ when $d_i = 0$, and $d = (\mathbb{1}^\top D \mathbb{1})/n$, so $d^{-1} = n/(\mathbb{1}^\top D \mathbb{1})$ when $\mathbb{1}^\top D \mathbb{1} > 0$, and $d^{-1} = 0$ when $d = 0$. For GCN-style propagation, the symmetric normalized adjacency is $D^{-1/2}AD^{-1/2}$. The dominant stationary component of this operator is not the uniform vector but the degree-weighted vector $D^{1/2}\mathbb{1}$. We therefore study the corrected normalized propagation matrix

$$\widehat{A} = D^{-1/2}AD^{-1/2} - \frac{1}{\mathbb{1}^\top D \mathbb{1}} D^{1/2}\mathbb{1}\mathbb{1}^\top D^{1/2}. \tag{1}$$

For comparison with the unnormalized theory of (Wang et al., 2024), we also write

$$\widetilde{A} = \frac{1}{d}A - \frac{1}{n}\mathbb{1}\mathbb{1}^\top.$$

Both operators remove the leading stationary direction before repeated propagation. The normalized operator $\widehat{A}$ is the one used in our main results and experiments.

4 Main Results

We now state the main spectral consequences of normalized corrected propagation. The CSBM (Duranthon & Zdeborová, 2025) adds Gaussian node features to a planted block model, making it possible to track three quantities in the same theorem: graph signal, feature noise, and propagation depth. Our results show that, after removing the degree-stationary component, the normalized operator behaves like a low-rank class-signal operator plus a residual whose multi-layer action remains controlled.

Definition 4.1. The CSBM, denoted $\mathcal{CSBM}(n, m, p, q, \mu, \nu, \sigma)$, is defined as follows. Let $V = [n]$ be the set of vertices (with n even), partitioned into two disjoint communities S and T of equal size $|S| = |T| = n/2$. Let $y \in \{-1, 1\}^n$ be the community indicator vector by setting $y_i = 1$ if $i \in S$ and $y_i = -1$ if $i \in T$. The model generates a random graph $G = (V, E)$ and a feature matrix $X \in \mathbb{R}^{n\times m}$ according to the following processes:

1. **Graph Generation:** For every pair of distinct vertices $i, j \in V$, the edge (i, j) is included in E independently with probability:

$$\mathbb{P}((i,j) \in E) = \begin{cases} p & \text{if } y_i = y_j \quad \text{(intra-class)}, \\ q & \text{if } y_i \neq y_j \quad \text{(inter-class)}. \end{cases}$$

2. **Feature Generation:** For each vertex $i \in V$, the feature vector $x_i \in \mathbb{R}^m$ (the i-th row of X) is drawn independently from a Gaussian mixture:

$$x_i \sim \begin{cases} \mathcal{N}(\mu, \sigma^2 I_m) & \text{if } i \in S, \\ \mathcal{N}(\nu, \sigma^2 I_m) & \text{if } i \in T. \end{cases}$$

For $p > q > 0$, set $\gamma(p, q) := \frac{p-q}{p+q}$. In Wang et al. (2024), proved that

Theorem 4.2 (Exact Recovery for $\widetilde{A}$ (Wang et al., 2024))**.** *Suppose we are given a 2-block m-dimensional CSBM with parameters $n, p > q, \mu, \nu, \sigma$ satisfying $\gamma(p, q) \geq \Omega\left(k\sqrt{\frac{\log n}{np}}\right)$ and $p \geq \frac{\log^3 n}{n}$. Then after $k = O(\log n)$ rounds of graph convolution with $\tilde{A}$, our data is linearly separable with probability $1 - n^{-\Omega(1)}$ if*

$$\frac{\|\mu - \nu\|}{\sigma} \geq \Omega\left(\max\left(\sqrt{\frac{\log n}{n}}, \left(\frac{C}{\gamma\sqrt{np}}\right)^k \sqrt{\log n}\right)\right)$$

where C is an absolute constant.

The first main theoretical result is the normalized analogue. It has the same structure, but the random degree normalization requires a dense polylogarithmic density assumption and a slightly stronger residual-control term.

Theorem 4.3 (Exact Recovery for $\widehat{A}$)**.** *Fix any constant $B > 4$. There exist constants $C_d, C_\gamma, C_0, C_*, C_{\text{snr}}, c_0 > 0$, depending only on B, such that the following holds. Consider a balanced two-block $\mathcal{CSBM}(n, m, p, q, \mu, \nu, \sigma)$ with $p > q > 0$, centered feature means $\mu + \nu = 0$. Let $\theta := C\sqrt{\frac{\log n}{np}}$ and $\rho := \frac{\theta}{\gamma}$, where $C > 0$ is a sufficiently large absolute constant. Assume $p \geq C_d \frac{\log^B n}{n}$ and suppose $1 \leq k \leq C_0 \log n$ and $\gamma \geq C_\gamma k\sqrt{\frac{\log n}{np}}$ (where C_0 is chosen no larger than the constant c in Lemma B.9), which equivalently implies $k\rho \leq c_0$. Then, after k rounds of normalized corrected graph convolution, the rows of $X^{(k)} = \widehat{A}^k X$ are linearly separable with probability at least $1 - n^{-\Omega(1)}$, provided*

$$\frac{\|\mu - \nu\|}{\sigma} \geq C_{\text{snr}} \max\left\{\sqrt{\frac{\log n}{n}}, \left(\frac{C_*}{\gamma}\sqrt{\frac{\log n}{np}}\right)^k \sqrt{\log n}\right\}.$$

Remark 4.4. Theorem 4.3 gives a normalized exact-recovery guarantee under the dense polylogarithmic condition $np \geq C \log^B n$ with fixed $B > 4$. This is a slightly stronger density assumption than the sharp $np \geq C \log^3 n$ target, but it makes Taylor remainders in the normalized degree expansion controllable by deterministic row-sum bounds.

Proof roadmap and key technical estimate. We include the proof roadmap here, and the main technical contribution of the theorem lies in the normalized residual-power analysis. First, the feature model is reduced

to the one-dimensional centered form $Xw = s + g$, where $g \sim N(0, \sigma'^2 I_n)$, by Lemma B.2. Second, degree concentration gives the signal-plus-residual decomposition

$$\widehat{A} = \eta s s^\top + R', \qquad \eta = \frac{(p-q)n}{2d}, \qquad \|R'\| \lesssim \sqrt{\frac{\log n}{np}}.$$

However, this spectral-norm control is not enough for exact recovery. We need worst-node control after $k = O(\log n)$ layers. The key technical estimate is the following short form of Lemma B.9: with high probability, uniformly over all vertices u and all $1 \le a \le k$,

$$\left|e_u^\top (R')^a s\right| \le \frac{1}{\sqrt{n}} \left(K \sqrt{\frac{\log n}{np}} \right)^a. \tag{Key}$$

This bound says that residual powers remain delocalized in the planted signal direction; the extra $n^{-1/2}$ factor is what allows a sign-based classifier to control every node rather than only an average error.

The proof of (Key) is the normalized part of the argument. Unlike the unnormalized case, each entry of R' contains random nonlinear degree denominators and a random stationary correction. We handle these terms by expanding normalized entries into positive-level atoms (Lemma B.5) and then bounding the surviving high-moment terms by decorated-walk counting (Lemma B.6). This yields the decoupled decorated-walk expansion and moment bound in Appendix B. Finally, Lemma B.10 combines (Key), the spectral norm bound, and Gaussian concentration to show that $\eta^{-k}\widehat{A}^k(s+g)$ stays within $o(n^{-1/2})$ of s in ℓ_∞, which gives linear separability.

The second theoretical result extends the analysis to L classes, with equal-sized graph blocks and Gaussian class-conditional features. The result is weaker than exact recovery, but it retains the same spectral conclusion: after normalization correction, most propagated representations contract toward their class centers when the graph signal dominates perturbation and feature noise.

Definition 4.5 (Multi-Class CSBM)**.** A multi-class Contextual Stochastic Block Model (multi-class CSBM) with L equal-sized classes, denoted as $\mathcal{CSBM}(n, m, p, q, L, \mu, \sigma)$, is generated as follows:

1. **Node Labels:** The n nodes are partitioned into L disjoint classes $\mathcal{C}_1, \ldots, \mathcal{C}_L$, each of size n/L.
2. **Graph Topology:** An undirected graph $G = (V, E)$ is generated where edges between nodes i and j are drawn independently. The probability of an edge is p if i and j belong to the same class, and q if they belong to different classes.
3. **Node Features:** For a node $i \in \mathcal{C}_l$, its m-dimensional feature vector is generated as $x_i \sim \mathcal{N}(\mu_l, \sigma^2 I_m)$, where $\mu_l \in \mathbb{R}^m$ is the specific center for class l. The parameter $\mu = \{\mu_1, \ldots, \mu_L\}$ denotes the set of all class centers. We assume the centers are zero-mean, i.e., $\sum_{l=1}^L \mu_l = 0$.

The expected average degree is

$$\bar{d} := \frac{pn}{L} + \frac{(L-1)qn}{L}.$$

Let $\lambda := (p-q)n/(\bar{d}L)$ be the graph signal strength and

$$\delta := C\bar{d}^{-1} \left(\sqrt{\frac{np(1-p)}{L}} + \sqrt{nq(1-q)} \right)$$

be the graph-noise scale. Let $U = \mathbb{E}[X]$, assume $U^\top \mathbb{1} = 0$, and set $\Delta := \min_{\ell \neq h} \|\mu_\ell - \mu_h\|$. We first recall the corresponding unnormalized result of Wang et al. (2024); the normalized analogue follows as the next theorem.

Theorem 4.6 (Multi-class Partial Recovery for $\widetilde{A}$ (Wang et al., 2024))**.** *Given the CSBM with parameters, p, q, L, n, m, σ, suppose $\min(p, q) \ge \Omega(\frac{\log^2 n}{n})$ and $|\lambda| > 4k\delta$. Let $X^{(k)} = \frac{1}{\lambda^k}\tilde{A}^k X$ be the feature matrix after*

k rounds of convolutions with scalaing factor $1/\lambda^k$. Let $x_i^{(k)}$ be the ith row of the matrix $X^{(k)}$. Then with probability $1 - n^{-\Omega(1)}$, at least $n - n_e$ nodes, i, satisfy $\left\|x_i^{(k)} - \mu_i\right\| < \Delta/2$ where

$$n_e = O\Big((k\delta/|\lambda|)^2 \frac{\|U\|_F^2}{\Delta^2} + (L + n(\delta/|\lambda|)^{2k}) \frac{\sigma^2 m \log n}{\Delta^2}\Big).$$

In particular, the quadratic classifer $x \mapsto softmax(\|x - c_l\|^2)_{l=1}^L$ will correctly classify at least $n - n_e$ points, and when $n_e = o(n)$, then we can correctly classify $1 - o(1)$ fraction of points.

The following theorem gives the corresponding partial-recovery guarantee for the normalized corrected operator $\widehat{A}$.

Theorem 4.7 (Multi-class Partial Recovery for $\widehat{A}$)**.** *Given the multi-class CSBM with parameters p, q, L, n, m, σ, assume that L is fixed, $L \mid n$, and all constants below may depend on L. Suppose the graph is sufficiently dense with $\bar{d} \geq C \log n$, the feature dimension satisfies $m \leq n^{C_m}$ for a fixed constant $C_m > 0$, and the signal dominates the noise such that $|\lambda| \geq C_0 k(\delta + \epsilon)$ for a sufficiently large absolute constant C_0, where $\epsilon := C_\epsilon \sqrt{\log n/\bar{d}}$. Let $X^{(k)} = \lambda^{-k} \widehat{A}^k X$ be the scaled feature matrix after k rounds of propagation. Let $x_i^{(k)}$ be the i-th row of $X^{(k)}$, and let $\ell(i)$ denote the class of node i. Then with high probability $1 - n^{-\Omega(1)}$, at least $n - n_e$ nodes satisfy $\|x_i^{(k)} - \mu_{\ell(i)}\| < \Delta/2$, where n_e is bounded by*

$$n_e = O\left(\underbrace{\Big[(\frac{\epsilon}{|\lambda|})^{2k} + 4k^2(\frac{\delta+\epsilon}{|\lambda|})^2\Big]\frac{\|U\|_F^2}{\Delta^2}}_{\textit{Structural \& Normalization Error}} + \underbrace{\Big[(1 + \frac{\delta+\epsilon}{|\lambda|})^{2k}(L-1) + n(\frac{\delta+\epsilon}{|\lambda|})^{2k}\Big]\frac{\sigma^2 m \log n}{\Delta^2}}_{\textit{Initial Variance Error}}\right).$$

In particular, the nearest-center classifier $\arg\min_{\ell \in [L]} \|x - \mu_\ell\|$ correctly classifies all nodes outside this exceptional set.

Theorem 4.7 gives an explicit partial-recovery bound for normalized corrected propagation. When graph perturbation, normalization error, and feature noise are small relative to Δ, the bound on n_e^{norm} is sublinear. The two error terms separate structural error from propagated feature variance:

- **Structural error.** The first term accumulates graph perturbation and degree-normalization error relative to class separation.
- **Feature-noise error.** The second term tracks propagated Gaussian noise and its dependence on m, σ, and Δ.

Compared with Theorem 4.6, the normalized result replaces the graph perturbation scale δ by $\delta + \epsilon$ and adds the purely normalization-driven term $(\epsilon/|\lambda|)^{2k}\|U\|_F^2/\Delta^2$.

The proof is deferred to Appendices B and C.2.

5 Experiments

The experiments are secondary to the theoretical results and are designed as empirical checks of the spectral predictions rather than as a full architecture benchmark. On synthetic CSBM graphs, we vary feature SNR, graph signal strength, and propagation depth to test the theorem's predicted dependencies. On standard node-classification benchmarks, we test whether the same normalized correction is associated with smaller depth-dependent degradation outside the idealized model. We compare against standard normalized propagation and three recent oversmoothing baselines:

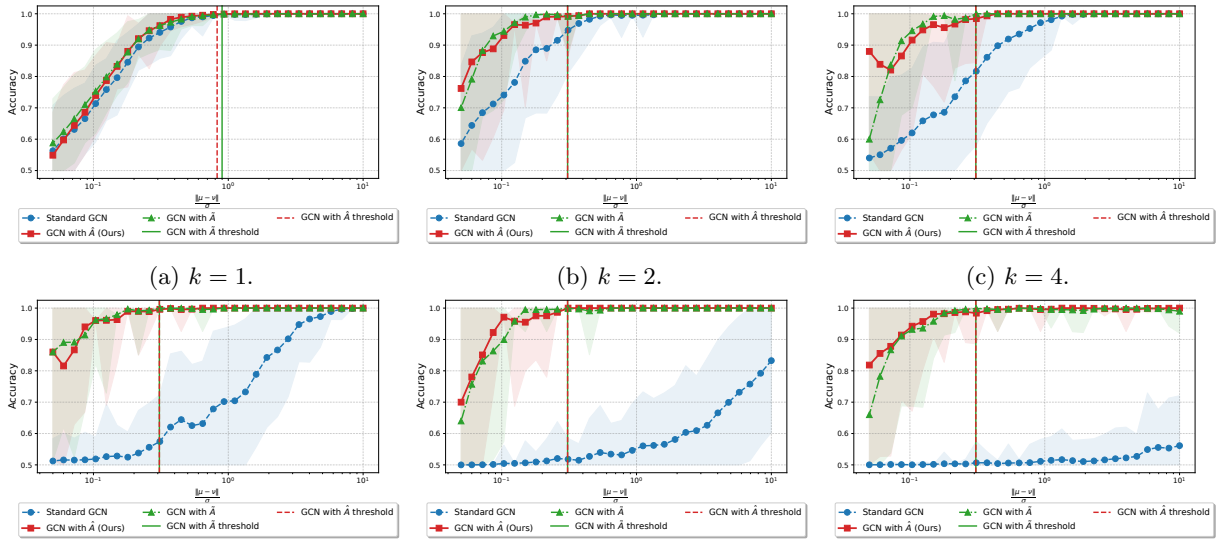
- **DropEdge** (Rong et al., 2020), which randomly removes edges during training.
- **GraphMamba** (He et al., 2025), which uses state-space modules for long-range propagation.
- **RevGNN** (Park et al., 2024), which uses reverse diffusion to recover high-frequency information.

5.1 Synthetic Data

We generate two-block CSBM graphs with $n = 2000$ and feature dimension $m = 20$. After k propagation steps, a linear classifier is trained on the resulting embeddings; all curves report average accuracy over 50 trials.

We test two regimes:

1. **Feature SNR.** We fix $\gamma = (p - q)/(p + q) = 2/3$ and vary $\|\mu - \nu\|/\sigma$. Consistent with the theorem, corrected propagation remains more stable at large k when feature SNR is sufficient (Figure 1).
2. **Graph signal.** We fix $\|\mu - \nu\|/\sigma = 1$ and vary γ. Corrected propagation retains higher accuracy as graph signal decreases, and the empirical transition follows the theoretical threshold trend (Figure 2).



(a) $k = 1$. (b) $k = 2$. (c) $k = 4$.

(d) $k = 8$. (e) $k = 12$. (f) $k = 16$.

Figure 1: Synthetic CSBM accuracy versus feature SNR for increasing propagation depth.

Baseline comparison. Against recent oversmoothing baselines, corrected propagation shows comparable accuracy trends while preserving the same stability pattern across both feature-SNR and graph-signal sweeps (Figures 3 and 4).

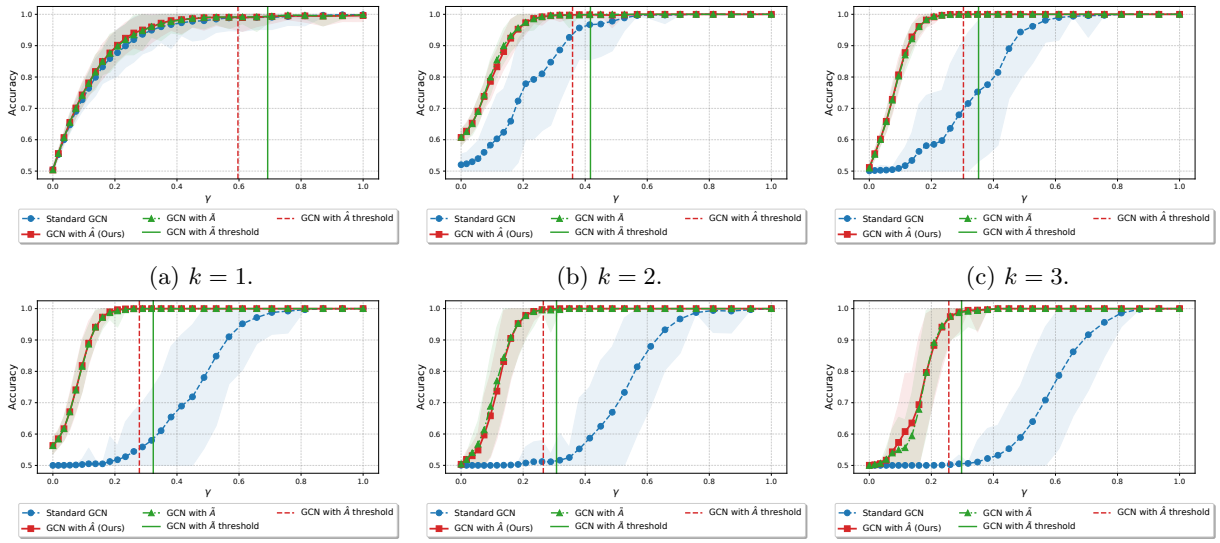
5.2 Real Data

We evaluate on *Cora*, *CiteSeer*, *PubMed*, *Reddit*, *ogbn-arxiv*, and *ogbn-products*. Table 1 reports dataset statistics. These experiments assess whether the correction is associated with reduced depth-dependent degradation on heterogeneous multi-class graphs.

Balanced two-cluster subsets. To align with the balanced CSBM theory, we select the two largest classes in each dataset and downsample the majority class to obtain a 1:1 class ratio. In these balanced subsets, corrected propagation maintains more stable accuracy as depth increases (Figure 5), consistent with the predicted effect of removing the stationary component.

Full Dataset Evaluation. On the full datasets, we observe a smaller depth-dependent accuracy drop for corrected propagation than for standard GCN propagation (Figure 6). The effect is smaller than on balanced subsets but remains consistent with the spectral-collapse mechanism.

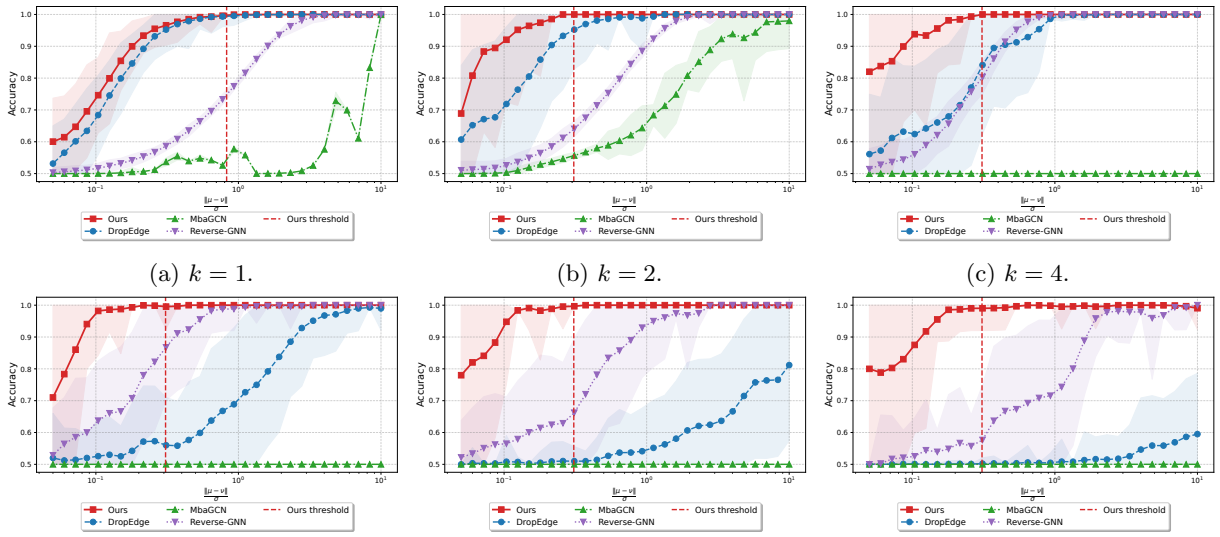
Baseline comparison. Corrected propagation has comparable accuracy to recent baselines on both balanced subsets and full datasets (Figures 7 and 8), while using a training-free operator-level correction.



(a) $k = 1$. (b) $k = 2$. (c) $k = 3$.

(d) $k = 4$. (e) $k = 5$. (f) $k = 6$.

Figure 2: Synthetic CSBM accuracy versus graph signal strength.



(a) $k = 1$. (b) $k = 2$. (c) $k = 4$.

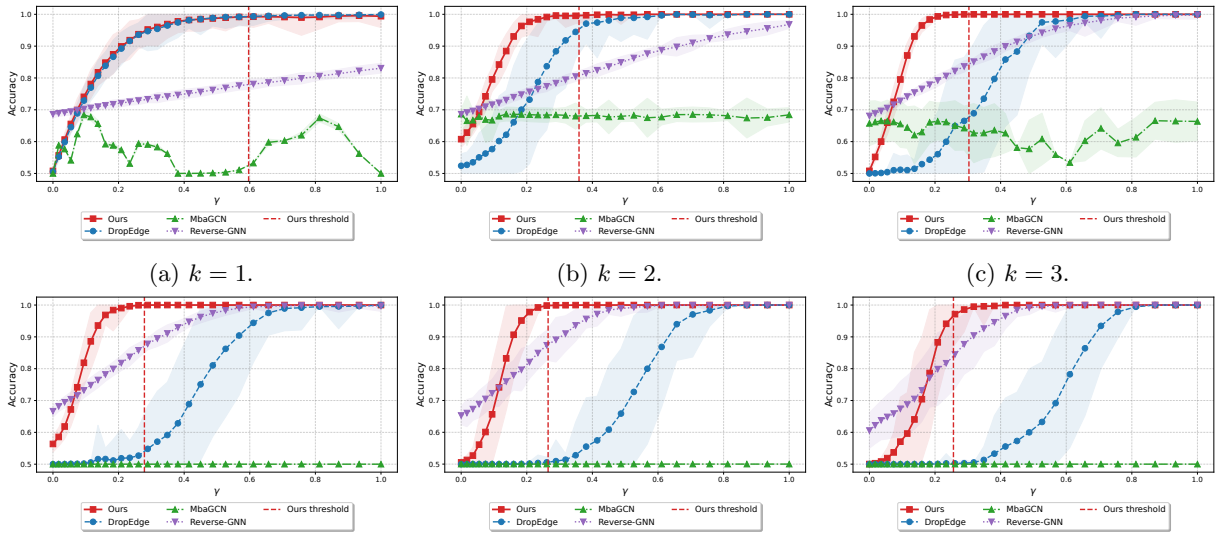
(d) $k = 8$. (e) $k = 12$. (f) $k = 16$.

Figure 3: Synthetic CSBM comparison with recent oversmoothing baselines under varying feature SNR.

6 Discussion and Limitations

The theoretical contribution is threefold. First, normalized corrected propagation is identified as a spectral object in its own right: subtracting the degree-stationary direction of $D^{-1/2}AD^{-1/2}$ yields an operator whose signal direction can remain separated for logarithmically many layers under explicit CSBM conditions. Second, the exact-recovery and partial-recovery theorems connect this spectral correction to node classification guarantees. Third, the proof develops normalized residual-power bounds that are entrywise rather than only spectral-norm bounds, which is the level needed for exact recovery.

The main technical message is that normalized correction requires tools beyond the unnormalized analysis. Random degrees enter through nonlinear factors and also appear inside the correction term. Our proof



(a) $k = 1$. (b) $k = 2$. (c) $k = 3$.

(d) $k = 4$. (e) $k = 5$. (f) $k = 6$.

Figure 4: Synthetic CSBM comparison with recent oversmoothing baselines under varying graph signal.

Table 1: Statistics of real-world datasets used in our experiments

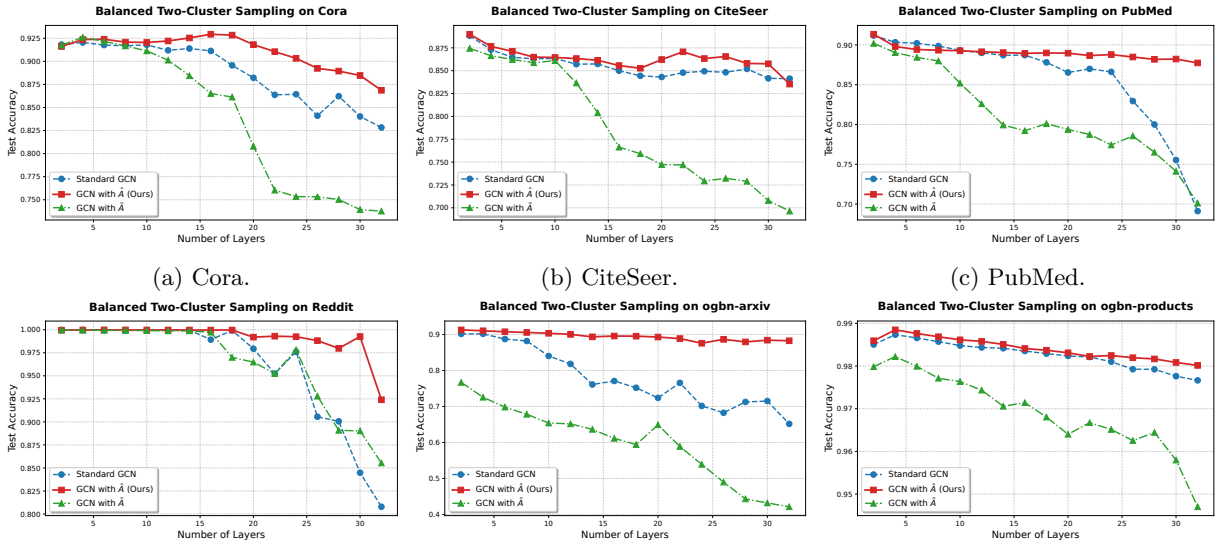
Dataset	Nodes	Edges	Features	Classes	Class Imbalance
Cora	2,708	10,556	1,433	7	Mild
CiteSeer	3,327	9,104	3,703	6	Mild
PubMed	19,717	88,648	500	3	Moderate
Reddit	232,965	114,615,892	602	41	High
ogbn-arxiv	169,343	1,166,243	128	40	High
ogbn-products	2,449,029	61,859,140	100	47	High

handles these effects by expanding each normalized residual entry into positive-level atoms and then counting the decorated walk patterns that survive in high moments. This is the main normalized-specific component of the theory.

Several limitations remain. First, the CSBM does not capture all degree heterogeneity, heterophily, or label semantics in real graphs. Second, the guarantees analyze the linear propagation backbone rather than nonlinear activations, attention, optimization, or learned feature transformations. Third, the binary theorem is proved in the dense polylogarithmic regime $np \geq C \log^B n$ for fixed $B > 4$, leaving the sharper $np \geq C \log^3 n$ normalized threshold open. Finally, very large graphs require scalable approximations of the stationary-component correction.

7 Conclusion and Future Work

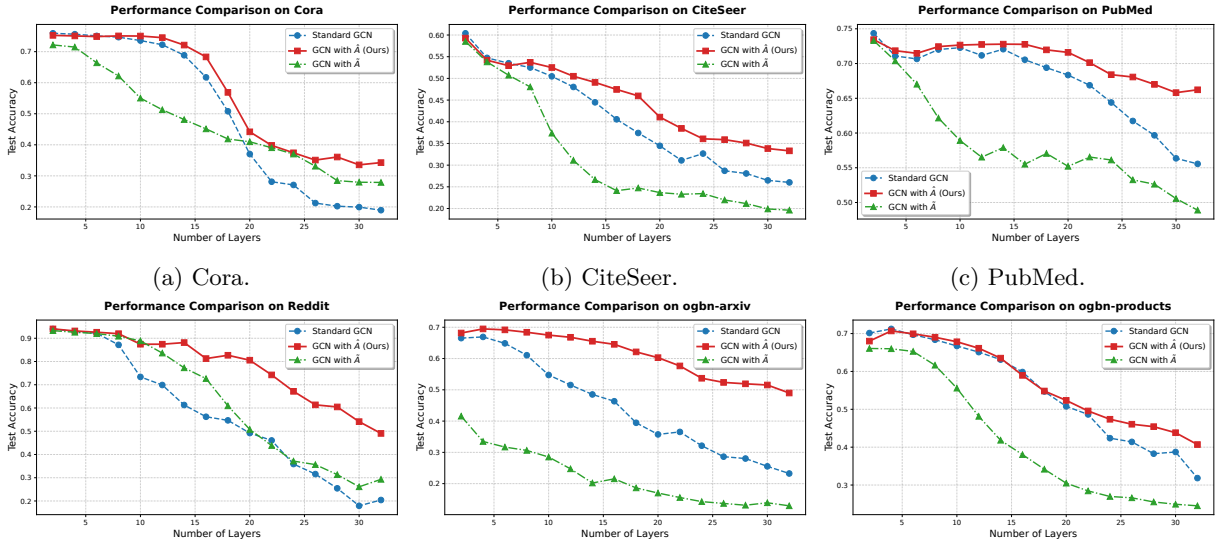
We developed a spectral theory of normalized corrected GNN propagation. The main result proves exact recovery in the binary CSBM after $k = O(\log n)$ layers under dense polylogarithmic graph density and explicit signal-to-noise assumptions. A multi-class theorem gives partial recovery for most nodes under analogous signal-dominance conditions. The central technical ingredient is entrywise control of normalized residual powers, obtained by atom expansion and decorated-walk counting. Future work includes sharpening the



(a) Cora. (b) CiteSeer. (c) PubMed.

(d) Reddit. (e) ogbn-arxiv. (f) ogbn-products.

Figure 5: Accuracy versus propagation depth on balanced two-cluster sampled subsets.



(a) Cora. (b) CiteSeer. (c) PubMed.

(d) Reddit. (e) ogbn-arxiv. (f) ogbn-products.

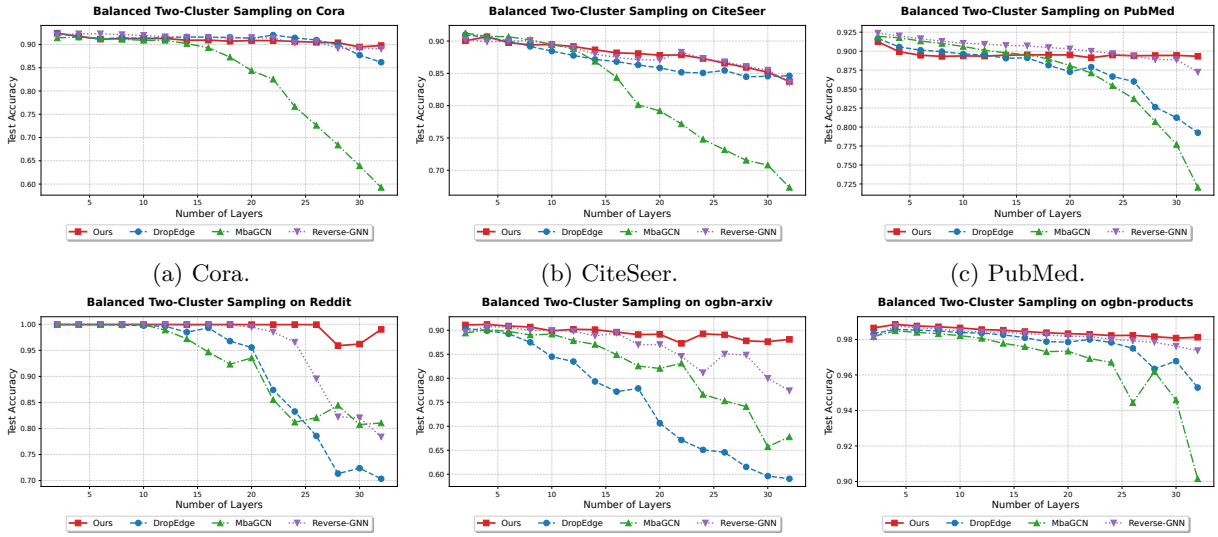
Figure 6: Accuracy versus propagation depth on full real-world datasets.

density threshold, extending the theory to heterophilous and degree-heterogeneous models, incorporating nonlinear trained GNN layers, and developing scalable approximations to the stationary correction.

References

Sam Adam-Day, Michael Benedikt, İsmail İlkan Ceylan, and Ben Finkelshtein. Almost surely asymptotically constant graph neural networks. In *Advances in Neural Information Processing Systems*, 2024.

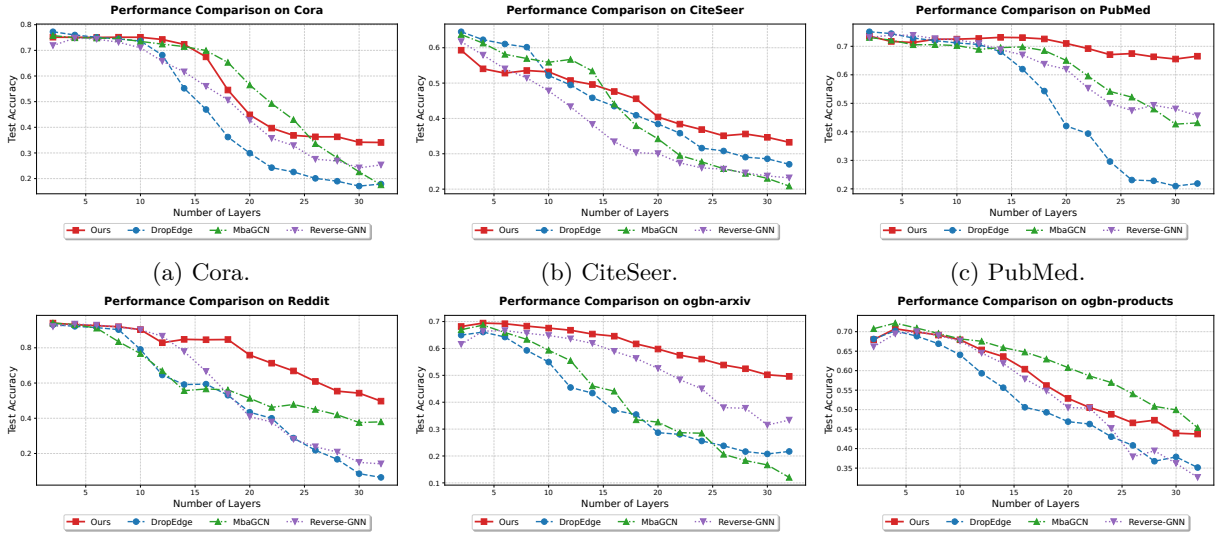
Muhammet Balcilar, Guillaume Renton, Pierre Héroux, Benoit Gaüzère, Sébastien Adam, and Paul Honeine. Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective. In *International*



(a) Cora. (b) CiteSeer. (c) PubMed.

(d) Reddit. (e) ogbn-arxiv. (f) ogbn-products.

Figure 7: Balanced-subset comparison with recent oversmoothing baselines.



(a) Cora. (b) CiteSeer. (c) PubMed.

(d) Reddit. (e) ogbn-arxiv. (f) ogbn-products.

Figure 8: Full-dataset comparison with recent oversmoothing baselines.

Conference on Learning Representations, 2021.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 684–693, 2021.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Effects of graph convolutions in multi-layer networks. In *The Eleventh International Conference on Learning Representations*, 2023a.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Optimality of message-passing architectures for sparse graphs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40320–40341,

2023b.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations*, 2014.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020.

Benjamin Paul Chamberlain, James Rowbottom, Maria I Gorinova, Stefan Webb, Emanuele Rossi, and Michael M Bronstein. Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pp. 1407–1418. PMLR, 2021.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and Deep Graph Convolutional Networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735. PMLR, 13–18 Jul 2020.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive Universal Generalized PageRank Graph Neural Network. In *International Conference on Learning Representations*, 2021.

Guillaume Dalle and Patrick Thiran. Optimal performance of graph convolutional networks on the contextual stochastic block model. In *The Third Learning on Graphs Conference*, 2024.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

O Duranthon and Lenka Zdeborová. Statistical physics analysis of graph neural networks: Approaching optimality in the contextual stochastic block model. *Physical Review X*, 15(4):041026, 2025.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations (ICLR)*, 2019.

Huseyin Goksu. Krawtchouknet: A unified gnn solution for heterophily and over-smoothing with adaptive bounded polynomials. *arXiv preprint arXiv:2511.15327*, 2025.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–1035, 2017.

Xin He, Yili Wang, Wenqi Fan, Xu Shen, Xin Juan, Rui Miao, and Xin Wang. Mamba-based graph convolutional networks: Tackling over-smoothing with selective state space. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5345–5353, 2025.

Yifan Hou, Jian Zhang, James Cheng, Kaili Ma, Richard T. B. Ma, Hongzhi Chen, and Ming-Chang Yang. Measuring and Improving the Use of Graph Information in Graph Neural Networks. In *International Conference on Learning Representations*, 2020.

Yufei Jin and Xingquan Zhu. Oversmoothing alleviation in graph neural networks: A survey and unified view. *Knowledge and Information Systems Journal*, 2025.

Nicolas Keriven. Not Too Little, Not Too Much: A theoretical Analysis of Graph (Over)Smoothing. In *Advances in Neural Information Processing Systems*, 2022.

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

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

Chen Lu and Subhabrata Sen. Contextual stochastic block model: Sharp thresholds and contiguity. *ArXiv*, 2020.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Sohir Maskey, Ron Levie, and Gitta Kutyniok. Generalization bounds for message passing networks on mixture of graphons. *SIAM Journal on Mathematics of Data Science*, 2024.

Nikhil Mehta, Lawrence Carin Duke, and Piyush Rai. Stochastic Blockmodels meet Graph Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4466–4474, 2019.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

MoonJeong Park, Jaeseung Heo, and Dongwoo Kim. Mitigating oversmoothing through reverse process of gnns for heterophilic graphs. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards Deep Graph Convolutional Networks on Node Classification. In *International Conference on Learning Representations*, 2020.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1), 2009.

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in GNNs. In *The Thirteenth International Conference on Learning Representations*, 2025.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Jiaqi WANG, Xinyi Wu, James Cheng, and Yifei Wang. A signed graph approach to understanding and mitigating oversmoothing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Robert Wang, Aseem Baranwal, and Kimon Fountoulakis. Analysis of corrected graph convolutions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871, 2019.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pp. 5453–5462. PMLR, 2018.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020.

A Mathematical Tools

To rigorously analyze the spectral properties of the corrected graph convolution and its effect on node features, we require a set of fundamental probabilistic and algebraic tools. These tools allow us to bound the spectral norms of random matrices, control eigenvalue perturbations, and manage the concentration of scalar sums that arise from the CSBM's stochastic generation process.

We begin with a foundational result from matrix perturbation theory, which guarantees that small changes to a symmetric matrix result in bounded shifts in its eigenvalues.

Theorem A.1. *Let A and B be $n \times n$ symmetric matrices. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ be the eigenvalues of A and B, respectively. If $\|A - B\| \leq \delta$, then $\max_i |\lambda_i - \mu_i| \leq \delta$.*

Furthermore, we often need to bound the spectral norm of specific sparse symmetric matrices encountered in our graph constructions. The following theorem provides a general upper bound based on row sparsity and entry magnitude.

Theorem A.2. *Let M be an $n \times n$ symmetric matrix such that each row of M has at most m non-zero entries and each entry of M has an absolute value at most ϵ. Then $\|M\| \leq \epsilon m$.*

While spectral bounds control the global operator behavior, local graph properties, such as node degrees and edge counts, are sums of independent Bernoulli random variables. To establish high-probability bounds on these quantities, we employ the standard Chernoff concentration inequality.

Theorem A.3. *Let $X_1, ...X_n$ be independent Bernoulli random variables with mean at most p, and let $S_n = \sum_{i=1}^n X_i$. Then for any $t > 0$,*

$$\mathbf{Pr}\left[|S_n - \mathbb{E}[S_n]| > t\right] \leq \exp\left(-\Omega\left(\frac{t^2}{\mathbb{E}[S_n]}\right)\right).$$

In the analysis of multi-layer graph convolutions, terms involving products and powers of $(1 + x)$ frequently emerge. To simplify these expressions while maintaining tight bounds, we utilize the following standard exponential inequalities.

Theorem A.4. *For all $x \in \mathbb{R}$, $1 + x \leq e^x$. Furthermore, if $x \leq 1$, then $e^x \leq 1 + 2x$.*

Finally, since the initial node features in our CSBM framework are drawn from Gaussian mixture models, we rely on the concentration of Gaussian random vectors to bound the noise propagation.

Theorem A.5. *Let $g \sim \mathcal{N}(0, \sigma^2 I_n)$ be a random Gaussian vector. Then for any $t > 0$,*

$$\mathbf{Pr}\left[\|g\| > t\right] \leq 2\exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

B Additional Preliminaries and Lemmas

We recall the spectral theorem for real symmetric matrices. If $M \in \mathbb{R}^{n\times n}$ is symmetric, it admits an eigendecomposition $M = \sum_{i=1}^n \lambda_i w_i w_i^\top$, where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the eigenvalues and $w_1, \ldots, w_n \in \mathbb{R}^n$ form an orthonormal basis of eigenvectors. In this case, the operator norm satisfies $\|M\| = \max_i |\lambda_i|$.

Additionally, we denote the multivariate Gaussian distribution with mean vector μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$. For a scalar random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we utilize the standard tail bound:

$$\mathbf{Pr}\left[|X - \mu| > t\sigma\right] \leq \exp\left(-\frac{t^2}{2}\right).$$

A key insight from Baranwal et al. (2021) establishes that the analysis of linear classifiers on general m-dimensional CSBMs can be effectively mapped to a simpler one-dimensional setting. To formalize this, we introduce the following definition:

Definition B.1 (One-Dimensional Centered CSBM)**.** A CSBM is said to be **one-dimensional and centered** with parameters n, p, q, σ, if the feature dimensionality is $m = 1$. In this model, the single feature vector $x \in \mathbb{R}^n$ is decomposed as $x = s + g$, where:

- $g \sim \mathcal{N}(0, \sigma^2 I_n)$ represents the Gaussian noise vector.
- $s \in \mathbb{R}^n$ is the *signal* vector defined by class membership:

$$s_i = \begin{cases} \frac{1}{\sqrt{n}} & \text{if } i \in S, \\ \frac{-1}{\sqrt{n}} & \text{if } i \in T. \end{cases}$$

This construction ensures that the signal vector s has a unit norm ($\|s\| = 1$), simplifying subsequent notation.

The lemma presented below enables the reduction of linear classifier analysis from the general CSBM to the centered 1-dimensional model. Accordingly, our proof strategy for the main theorems begins with the 1-dimensional case, subsequently applying Lemma B.2 to extend the results.

Lemma B.2. *Consider an m-dimensional balanced two-block* $\mathrm{CSBM}(n, m, p, q, \mu, \nu, \sigma)$ *with $p > q > 0$ with centered feature means $\mu + \nu = 0$ and $\mu \neq \nu$. Let $s_i = n^{-1/2}$ on S and $s_i = -n^{-1/2}$ on T. Then the projection*

$$w = \frac{2(\mu - \nu)}{\sqrt{n}\,\|\mu - \nu\|^2}$$

satisfies

$$Xw = s + g, \qquad g \sim N(0, \sigma'^2 I_n), \qquad \sigma' = \frac{2\sigma}{\sqrt{n}\,\|\mu - \nu\|}.$$

Consequently, for every $k \geq 1$,

$$\widehat{A}^k X w = \widehat{A}^k (s + g).$$

Thus any sign-separation result for the one-dimensional centered model transfers to the original CSBM through the linear classifier with normal vector w.

Proof. It follows from $\mu + \nu = 0$ that

$$\nu = -\mu, \qquad \mu - \nu = 2\mu.$$

For each vertex i, the feature vector can be written as

$$x_i = \begin{cases} \mu + \xi_i, & i \in S, \\ \nu + \xi_i, & i \in T, \end{cases}$$

where $\xi_i \sim N(0, \sigma^2 I_m)$ are independent. Recall that

$$w = \frac{2(\mu - \nu)}{\sqrt{n}\,\|\mu - \nu\|^2}.$$

Then, if $i \in S$,

$$\mu^\top w = \mu^\top \frac{2(\mu - \nu)}{\sqrt{n}\,\|\mu - \nu\|^2} = \frac{2\mu^\top (2\mu)}{\sqrt{n}\,\|2\mu\|^2} = \frac{1}{\sqrt{n}}.$$

Similarly, if $i \in T$, since $\nu = -\mu$, we also have

$$\nu^\top w = -\frac{1}{\sqrt{n}}.$$

Therefore

$$x_i^\top w = s_i + \xi_i^\top w.$$

Define

$$g_i := \xi_i^\top w.$$

Since the vectors ξ_i are independent Gaussian vectors with distribution $N(0, \sigma^2 I_m)$, the variables g_i are independent centered Gaussian random variables. Moreover,

$$\mathrm{Var}(g_i) = \sigma^2 \|w\|^2.$$

By the definition of w,

$$\|w\|^2 = \frac{4\|\mu - \nu\|^2}{n\|\mu - \nu\|^4} = \frac{4}{n\|\mu - \nu\|^2},$$

which implies that

$$g_i \sim N\left(0, \frac{4\sigma^2}{n\|\mu - \nu\|^2}\right).$$

Equivalently,

$$g_i \sim N(0, \sigma'^2), \qquad \sigma' = \frac{2\sigma}{\sqrt{n}\,\|\mu - \nu\|}.$$

Thus, in vector form,

$$Xw = s + g,$$

where $g = (g_1, \ldots, g_n)^\top$ has independent coordinates distributed as $N(0, \sigma'^2)$.

Since $\widehat{A}$ acts on the vertex coordinate and w is fixed in the feature coordinate, we have, for every integer $k \geq 1$,

$$\widehat{A}^k X w = \widehat{A}^k (Xw) = \widehat{A}^k (s + g).$$

If the coordinates of $\widehat{A}^k(s + g)$ have signs agreeing with the true community labels, then

$$\mathrm{sign}\big((\widehat{A}^k X)_i w\big) = \mathrm{sign}\big((\widehat{A}^k (s + g))_i\big)$$

correctly classifies every vertex i. Hence the rows of $\widehat{A}^k X$ are linearly separable by the hyperplane with normal vector w and threshold 0. □

In the 1-dimensional model, the presence of the signal s in the feature space is immediate. Furthermore, the convolution matrix effectively captures this signal, as it can be modeled as a perturbation of the signal matrix $ss^\top$. Following the framework of (Wang et al., 2024), the un-normalized convolution matrix $\tilde{A}$ admits the decomposition

$$\tilde{A} = \eta ss^\top + \frac{1}{d} R + d' \mathbb{1}\mathbb{1}^\top \tag{2}$$

where $\eta := \frac{(p-q)n}{2d}$ represents the *signal strength*, $d' := \frac{(p+q)n - 2d}{2nd}$ denotes the *average degree deviation*, and R is the "edge-deviation" matrix defined by $R_{i,j} = A_{i,j} - \mathbb{E}[A_{i,j}]$. The expected signal strength ratio is defined as $\gamma = \frac{|p-q|}{p+q}$. As established in (Wang et al., 2024), since R consists of i.i.d. zero-mean entries with bounded variance, and d' is controlled by degree concentration, the matrix $\tilde{A}$ concentrates around $\gamma ss^\top$. This behavior extends to the normalized convolution matrix $\widehat{A}$ as well. The relevant concentration bounds from Wang et al. (2024) are summarized in the following proposition.

Our analysis commences with a spectral norm bound for the corrected graph convolution matrix. Lemma B.3 bounds the spectral norm of the deviation between $\widehat{A}$ and $\tilde{A}$. Control over this spectral norm is fundamental to limiting error propagation across network layers.

Lemma B.3. *Suppose that $p = \Omega(\log n/n)$ and $p > q > 0$. Then, there exists a constant $C > 0$ such that with probability at least $1 - n^{-\Omega(1)}$,*

$$\|\widehat{A} - \widetilde{A}\| \leq C\sqrt{\frac{\log n}{np}}.$$

Proof. Let

$$\bar{d} := \frac{(p+q)n}{2}, \qquad \varepsilon := K\sqrt{\frac{\log n}{np}},$$

where $K > 0$ is a sufficiently large absolute constant. Since $q \le p$, we have $\bar{d} = \Theta(np)$. For each vertex v, its degree d_v is a sum of independent Bernoulli random variables and

$$\mathbb{E}d_v = \bar{d} + O(p).$$

The $O(p)$ term comes from excluding the vertex v itself from its own community and is negligible compared with $\bar{d} = \Theta(np)$. By Theorem A.3,

$$\mathbb{P}\big(|d_v - \mathbb{E}d_v| > \varepsilon\bar{d}/2\big) \le 2\exp\left(-c\varepsilon^2\bar{d}\right) \le 2\exp(-c'K^2\log n).$$

Taking K large enough and applying a union bound over all $v \in [n]$, we obtain, with probability at least $1 - n^{-\Omega(1)}$,

$$d_v \in [\bar{d}(1-\varepsilon), \bar{d}(1+\varepsilon)] \quad \text{for every } v \in [n]. \tag{3}$$

On this event,

$$d = \frac{1}{n}\sum_{v=1}^{n} d_v \in [\bar{d}(1-\varepsilon), \bar{d}(1+\varepsilon)]. \tag{4}$$

Now we bound $\|\widehat{A} - \widetilde{A}\|$. Recall that

$$\widehat{A} - \widetilde{A} = \left(D^{-1/2}AD^{-1/2} - \frac{1}{d}A\right) - \left(\frac{1}{1^\top D 1}D^{1/2}11^\top D^{1/2} - \frac{1}{n}11^\top\right).$$

It is enough to bound the two terms separately. Let

$$M_1 := D^{-1/2}AD^{-1/2} - \frac{1}{d}A.$$

Then M_1 is symmetric, and for every u, v,

$$(M_1)_{uv} = A_{uv}\left(\frac{1}{\sqrt{d_u d_v}} - \frac{1}{d}\right).$$

By (3) and (4), we have

$$\left|\frac{1}{\sqrt{d_u d_v}} - \frac{1}{d}\right| \le \frac{C\varepsilon}{\bar{d}}.$$

Moreover,

$$\Delta(G) := \max_v d_v \le 2\bar{d}$$

for all sufficiently large n. Therefore every row of M_1 has absolute row sum at most

$$\sum_v |(M_1)_{uv}| \le d_u \cdot \frac{C\varepsilon}{\bar{d}} \le C'\varepsilon.$$

Applying Theorem A.2 gives that

$$\|M_1\| \le C'\varepsilon. \tag{5}$$

Next let

$$M_2 := \frac{1}{1^\top D 1}D^{1/2}11^\top D^{1/2} - \frac{1}{n}11^\top.$$

Then M_2 is symmetric and its entries are

$$(M_2)_{uv} = \frac{\sqrt{d_u d_v}}{\sum_{w=1}^{n} d_w} - \frac{1}{n}.$$

Again (3) and the fact

$$\sum_{w=1}^{n} d_w \in [n\bar{d}(1-\varepsilon), n\bar{d}(1+\varepsilon)],$$

we obtain

$$\left| \frac{\sqrt{d_u d_v}}{\sum_w d_w} - \frac{1}{n} \right| \leq \frac{C\varepsilon}{n}.$$

Thus every row of M_2 has absolute row sum at most

$$\sum_v |(M_2)_{uv}| \leq n \cdot \frac{C\varepsilon}{n} = C\varepsilon.$$

Applying Theorem A.2 again gives that

$$\|M_2\| \leq C\varepsilon. \tag{6}$$

Combining (5) and (6) gives

$$\|\widehat{A} - \widetilde{A}\| \leq \|M_1\| + \|M_2\| \leq C\varepsilon = CK\sqrt{\frac{\log n}{np}}.$$

Absorbing K into the constant C, we conclude

$$\|\widehat{A} - \widetilde{A}\| \leq C\sqrt{\frac{\log n}{np}}.$$

This completes the proof. □

Lemma B.4. *Assume that $p \geq C\frac{\log n}{n}$ and $p > q > 0$. Let*

$$\gamma := \frac{p-q}{p+q}, \qquad \bar{d} := \frac{(p+q)n}{2}, \qquad \eta := \frac{(p-q)n}{2d}, \qquad d' := \frac{\bar{d} - d}{nd}.$$

Then, with probability at least $1 - n^{-\Omega(1)}$, the empirical average degree d satisfies

$$d = \bar{d}\left(1 + O\left(\sqrt{\frac{\log n}{n^2 p}}\right)\right).$$

Consequently, the deviations of η and d' are bounded by

$$|\eta - \gamma| \leq C\frac{\sqrt{\log n}}{n\sqrt{p}}, \qquad |d'| \leq C\frac{\sqrt{\log n}}{n^2\sqrt{p}}.$$

Moreover, let $R := A - \mathbb{E}A$ denote the centered adjacency fluctuation. Then the spectral norms are bounded as follows:

$$\|R\| \leq C\sqrt{np}, \qquad \left\|\widetilde{A} - \gamma ss^\top\right\| \leq \frac{C}{\sqrt{np}}.$$

Finally, combined with Lemma B.3, we have

$$\left\|\widehat{A} - \gamma ss^\top\right\| \leq C\sqrt{\frac{\log n}{np}}.$$

Consequently,

$$\left\|\widehat{A} - \eta ss^\top\right\| \leq C\sqrt{\frac{\log n}{np}}.$$

Proof. Note that $d = 2|E|/n$ is the average degree and $|E|$ is a sum of independent Bernoulli random variables with

$$\mathbb{E}|E| = \frac{n^2}{4}(p+q) + O(np).$$

We have

$$\mathbb{E}d = \bar{d} + O(p).$$

Note that the variance of $|E|$ is at most Cn^2p. By Theorem A.3, with probability at least $1 - n^{-\Omega(1)}$,

$$||E| - \mathbb{E}|E|| \leq Cn\sqrt{p\log n}.$$

Therefore

$$|d - \bar{d}| \leq C\sqrt{p\log n}.$$

Since $\bar{d} = \Theta(np)$, this implies

$$d = \bar{d}\left(1 + O\left(\frac{\sqrt{p\log n}}{np}\right)\right) = \bar{d}\left(1 + O\left(\sqrt{\frac{\log n}{n^2p}}\right)\right).$$

We now estimate $\eta - \gamma$. By definition,

$$\eta = \frac{(p-q)n}{2d}, \qquad \gamma = \frac{p-q}{p+q} = \frac{(p-q)n}{2\bar{d}}.$$

Hence

$$|\eta - \gamma| = \frac{(p-q)n}{2}\left|\frac{1}{d} - \frac{1}{\bar{d}}\right| = \frac{(p-q)n}{2}\frac{|d-\bar{d}|}{d\bar{d}}.$$

Using $d, \bar{d} = \Theta(np)$, $p - q \leq p + q = O(p)$, and

$$|d - \bar{d}| \leq C\sqrt{p\log n},$$

we obtain

$$|\eta - \gamma| \leq C\frac{np\sqrt{p\log n}}{n^2p^2} = C\frac{\sqrt{\log n}}{n\sqrt{p}}. \tag{7}$$

Similarly,

$$|d'| = \frac{|\bar{d} - d|}{nd} \leq \frac{C\sqrt{p\log n}}{n \cdot np} = C\frac{\sqrt{\log n}}{n^2\sqrt{p}}. \tag{8}$$

Next we control the centered adjacency matrix

$$R = A - \mathbb{E}A.$$

The entries of R are independent above the diagonal, centered, and bounded by 1, with variances at most p. By a standard spectral norm bound for sparse centered adjacency matrices, since $p \geq C\log n/n$, with probability at least $1 - n^{-\Omega(1)}$,

$$\|R\| \leq C\sqrt{np}. \tag{9}$$

We now decompose $\widetilde{A}$. Up to the harmless diagonal correction, which has operator norm $O(1/d) = O(1/(np))$, the expectation of A can be written as

$$\mathbb{E}A = \frac{p+q}{2}11^\top + \frac{p-q}{2}yy^\top,$$

where $y_i = 1$ on S and $y_i = -1$ on T. It follows from $s = y/\sqrt{n}$ that

$$yy^\top = nss^\top.$$

Thus

$$\frac{1}{d}\mathbb{E}A - \frac{1}{n}11^\top = \eta ss^\top + d'11^\top + E_{\text{diag}},$$

where

$$\|E_{\text{diag}}\| \le \frac{C}{d} \le \frac{C}{np}. \tag{10}$$

Therefore

$$\widetilde{A} = \frac{1}{d}A - \frac{1}{n}11^\top = \eta ss^\top + \frac{1}{d}R + d'11^\top + E_{\text{diag}}.$$

Consequently,

$$\widetilde{A} - \gamma ss^\top = (\eta - \gamma)ss^\top + \frac{1}{d}R + d'11^\top + E_{\text{diag}}.$$

Taking operator norms gives

$$\left\|\widetilde{A} - \gamma ss^\top\right\| \le |\eta - \gamma| + \frac{1}{d}\|R\| + |d'|\|11^\top\| + \|E_{\text{diag}}\|.$$

It follows from the fact $p \ge C\log n/n$ that

$$\frac{\sqrt{\log n}}{n\sqrt{p}} \le \frac{C}{\sqrt{np}}.$$

This together with (7), (8), (9) and (10) yields that

$$\left\|\widetilde{A} - \gamma ss^\top\right\| \le \frac{C}{\sqrt{np}}.$$

By Lemma B.3, we have that

$$\left\|\widehat{A} - \widetilde{A}\right\| \le C\sqrt{\frac{\log n}{np}}.$$

Therefore,

$$\left\|\widehat{A} - \gamma ss^\top\right\| \le \left\|\widehat{A} - \widetilde{A}\right\| + \left\|\widetilde{A} - \gamma ss^\top\right\| \le C\sqrt{\frac{\log n}{np}},$$

and then together with (7), this implies

$$\left\|\widehat{A} - \eta ss^\top\right\| \le \left\|\widehat{A} - \gamma ss^\top\right\| + |\gamma - \eta| \le C\sqrt{\frac{\log n}{np}}.$$

This completes the proof. □

For the next lemma, let $y_i = \sqrt{n}\, s_i \in \{\pm 1\}$. For an unordered edge $e = \{r, t\}$, write $\xi_e = A_{rt} - \mathbb{E}A_{rt}$. An *atom* is one monomial term produced by the finite Taylor expansion of a single normalized-residual entry and has the form

$$\mathfrak{a}(i,j) = c_{\mathfrak{a}}\, \chi_{\mathfrak{a}}(i,j)\, n^{-b_{\mathfrak{a}}}\, \bar{d}^{-q_{\mathfrak{a}}} \sum_{\phi \in \Phi_{\mathfrak{a}}(i,j)} \prod_{h=1}^{m_{\mathfrak{a}}} \xi_{e_h(\phi)},$$

where $\bar{d} = (p+q)n/2$, $\chi_{\mathfrak{a}}(i,j) \in \{1, y_i, y_j, y_i y_j\}$, and $\Phi_{\mathfrak{a}}(i,j)$ is the finite set of auxiliary vertex choices generated by local and global degree sums. For $\phi \in \Phi_{\mathfrak{a}}(i,j)$, $e_h(\phi)$ is the h-th centered edge variable selected by that auxiliary choice. The integer $m_{\mathfrak{a}}$ counts centered edge variables, $q_{\mathfrak{a}}$ is the power of $\bar{d}^{-1}$, and $b_{\mathfrak{a}}$ records powers of n^{-1} from global degree averages. The atom level is

$$\lambda_{\mathfrak{a}} := 2q_{\mathfrak{a}} - m_{\mathfrak{a}}.$$

Lemma B.5. *Let $R' = \widehat{A} - \eta ss^\top$, where $\eta = (p-q)n/(2d)$, and fix $B > 4$. Assume $p \geq C_d \log^B n/n$. Fix $A_0 > 0$, let $M \leq A_0 \log n$, and set*

$$L_* := K_0 M, \qquad \tau := \left(C \frac{\log n}{\sqrt{np}} \right)^{L_*+1},$$

where K_0 is sufficiently large depending only on A_0. There is an event $\mathcal{G}$ with $\mathbb{P}(\mathcal{G}^c) \leq \exp(-c \log^2 n)$, a clipped residual R_{cl}, and a decomposition

$$R_{\mathrm{cl}} = R_{\mathrm{pol}} + R_{\mathrm{rem}}, \qquad R_{\mathrm{cl}} = R' \quad \text{on } \mathcal{G},$$

such that, with the atom notation above,

$$(R_{\mathrm{pol}})_{ij} = \sum_{\mathfrak{a} \in \mathcal{A}} \mathfrak{a}(i,j) \qquad (i,j \in [n]),$$

for a finite deterministic atom family $\mathcal{A}$ satisfying

$$\lambda_{\mathfrak{a}} \geq 1, \qquad m_{\mathfrak{a}} \leq C\lambda_{\mathfrak{a}}, \qquad \#\{\mathfrak{a} \in \mathcal{A} : \lambda_{\mathfrak{a}} = \lambda\} \leq C^{\lambda}.$$

Deterministic diagonal/no-self atoms have $q = 1, m = 0$ and are contracted before the centered-edge skeleton is counted. On $\mathcal{G}$,

$$\|R_{\mathrm{rem}}\|_{\infty\to\infty} \leq C\tau, \qquad \|R_{\mathrm{cl}}\|_{\infty\to\infty} + \|R_{\mathrm{pol}}\|_{\infty\to\infty} \leq C.$$

Proof. **Off-diagonal decomposition.** We first isolate the degree-normalization factors in one off-diagonal entry. Write

$$\beta_i := \frac{d_i - \bar{d}}{\bar{d}}, \qquad \beta := \frac{d - \bar{d}}{\bar{d}}.$$

Then $d_i = \bar{d}(1+\beta_i)$ and $d = \bar{d}(1+\beta)$. For $i \neq j$, define

$$F_{ij} := (1+\beta_i)^{-1/2}(1+\beta_j)^{-1/2} = \frac{\bar{d}}{\sqrt{d_i d_j}},$$

and

$$G_{ij} := (1+\beta_i)^{1/2}(1+\beta_j)^{1/2}(1+\beta)^{-1} = \frac{\sqrt{d_i d_j}}{\bar{d}(1+\beta)}.$$

It is easy to see that

$$(D^{-1/2}AD^{-1/2})_{ij} = \frac{A_{ij}}{\bar{d}} F_{ij}, \quad \text{and} \quad \frac{\sqrt{d_i d_j}}{\mathbb{1}^\top D \mathbb{1}} = \frac{1}{n} G_{ij}.$$

For $i \neq j$,

$$\mathbb{E}A_{ij} = \frac{\bar{d}}{n}(1 + \gamma y_i y_j), \qquad \eta = \gamma(1+\beta)^{-1}, \qquad s_i s_j = \frac{y_i y_j}{n}.$$

Substituting $A_{ij} = \mathbb{E}A_{ij} + \xi_{ij}$ into $R' = \widehat{A} - \eta ss^\top$ gives the off-diagonal decomposition

$$R'_{ij} = \frac{\xi_{ij}}{\bar{d}} F_{ij} + \frac{1}{n}\big(F_{ij} - G_{ij}\big) + \frac{\gamma y_i y_j}{n}\big(F_{ij} - (1+\beta)^{-1}\big). \tag{11}$$

The right-hand side vanishes at the idealized deterministic point $\xi = 0$, $\beta_i = \beta_j = \beta = 0$. The actual no-self offset $\mathbb{E}d_i - \bar{d} = O(p)$ is deterministic of size $O(1/n)$ after normalization and is charged as a deterministic atom with $q = 1, m = 0$.

We next expand the degree-normalization factors. This is where clipping is used: it makes the Taylor expansion globally defined while leaving the matrix unchanged on the high-probability event below. Define the event $\mathcal{G}$ by

$$\max_i |d_i - \mathbb{E}d_i| \leq C\sqrt{np}\log n, \qquad \left| \sum_v d_v - n\bar{d} \right| \leq Cn\sqrt{p}\,\log n.$$

Chernoff bounds and a union bound give $\mathbb{P}(\mathcal{G}^c) \le \exp(-c\log^2 n)$, and on $\mathcal{G}$ all degree fluctuations in (11) have absolute value $O(\log n/\sqrt{np}) < c_*$ after increasing C_d. In particular, $1+\beta_i$, $1+\beta_j$, and $1+\beta$ are bounded away from zero on $\mathcal{G}$, so the factors F_{ij} and G_{ij} are well-defined and uniformly bounded there. Replace $x^{-1/2}$, $x^{1/2}$, and x^{-1} by smooth clipped functions agreeing with them on $[1-c_*, 1+c_*]$. Taylor expand each clipped function to depth L_* around leave-one-out degrees $d_i^{(-j)}$, $d_j^{(-i)}$, and the global average d.

Atom expansion. The polynomial Taylor terms give R_{pol}. Each exposed edge contributes $\xi_{ij}/\bar{d}$, each local degree decoration contributes $\bar{d}^{-1}\sum_{\ell\neq i,j}\xi_{i\ell}$ or $\bar{d}^{-1}\sum_{\ell\neq i,j}\xi_{j\ell}$, and each global decoration contributes $n^{-1}\bar{d}^{-1}\sum_{r<s}\xi_{rs}$. In the atom notation, the auxiliary map ϕ records the choices of vertices in these sums. Thus an atom with $m_{\mathfrak{a}}$ centered edge variables and $q_{\mathfrak{a}}$ powers of $\bar{d}^{-1}$ has level $2q_{\mathfrak{a}} - m_{\mathfrak{a}}$. The cancellation in (A) removes all level-zero deterministic terms. Exposed-edge and local/global degree atoms have level at least one, deterministic no-self atoms have level two, and the number of Taylor multi-indices of level λ is at most C^λ. This proves the atom bounds.

Remainder bound. The Taylor remainders form R_{rem}. On $\mathcal{G}$, every normalized degree fluctuation W appearing in a remainder, such as a local leave-one-out degree fluctuation, a global average-degree fluctuation, or the exposed-edge term $A_{ij}/\bar{d}$, satisfies $|W| \le C\log n/\sqrt{np}$, so

$$|\rho_{\alpha,L_*}(W)| \le \left(C\frac{\log n}{\sqrt{np}}\right)^{L_*+1} = \tau.$$

Remainder terms in the adjacency part have row sums at most $C\tau$, and remainder terms in the stationary correction have entries at most $C\tau/n$. The displayed $\ell_\infty \to \ell_\infty$ bounds follow from the degree concentration on $\mathcal{G}$. □

Lemma B.6. *Consider a product of $2r$ path monomials obtained from the atoms in Lemma B.5. Suppose the product has $2M$ residual positions and total level $\Lambda = 2M + h$. After summing over all canonical decorated patterns of this level and over all vertex assignments, the total contribution to the $2r$-th moment is at most*

$$n^{-r}(CM)^M(np)^{-M}\left(\frac{CM}{\sqrt{np}}\right)^h.$$

Proof. For a fixed product, let m be the total number of centered edge variables before identification, let ℓ be the number of distinct centered edges after identification, and let Q be the total $\bar{d}^{-1}$-charge. By the definition of level,

$$Q - \frac{m}{2} = \frac{\Lambda}{2}.$$

The expectation is zero unless every centered edge appears at least twice; hence $\ell \le m/2$. For a nonzero pattern,

$$\left|\mathbb{E}\prod_e \xi_e^{\nu_e}\right| \le C^m p^\ell, \qquad \nu_e \ge 2.$$

Global decorations have factors $n^{-1}\bar{d}^{-1}\sum_{r<s}\xi_{rs}$. If a global-only edge is not identified with the local path skeleton, its two free endpoints give at most n^2 choices and the two corresponding global normalizations give n^{-2}. Hence global-only decorations do not create a positive power of n. After these cancellations, reveal the remaining canonical skeleton from the fixed root u. If F is the number of new free vertices, each first appears through a centered edge that is paired later, so $F \le \ell$. The $2r$ terminal factors s_{v_a} contribute n^{-r}, and the total vertex factor is at most $n^{\ell-r}$.

For one fixed canonical pattern, the contribution is therefore bounded by

$$n^{\ell-r}p^\ell(np)^{-Q} = n^{-r}(np)^{\ell-Q} \le n^{-r}(np)^{-\Lambda/2}.$$

It remains to count canonical patterns. The $2M$ residual positions, their paired first occurrences, and the choices of atom sources give at most $(CM)^M$ possibilities at base level $2M$. The extra level h, coming

from additional Taylor decorations or deterministic charges, contributes at most $(CM)^h$ choices. Thus the contribution of all patterns of level $\Lambda = 2M + h$ is at most

$$n^{-r}(CM)^M(np)^{-M}\left(\frac{CM}{\sqrt{np}}\right)^h,$$

as claimed. □

Lemma B.7. *Let $R' = \widehat{A} - \eta ss^\top$, where $\eta = (p-q)n/(2d)$. Fix $B > 4$, and assume $p \geq C_d \log^B n/n$. Fix $A_0 > 0$. For every $M \leq A_0 \log n$, every pair (a, r) with $ar = M$, and every $u \in [n]$, there is a decomposition*

$$e_u^\top R'^a s = Z_{u,a}^{(M)} + \mathcal{E}_{u,a}^{(M)}$$

such that

$$\mathbb{E}|\mathcal{E}_{u,a}^{(M)}|^{2r} \leq n^{-r}(CM)^M(np)^{-M}, \qquad \mathbb{E}|Z_{u,a}^{(M)}|^{2r} \leq n^{-r}(CM)^M(np)^{-M}.$$

The constant C depends only on A_0 and B.

Proof. Apply Lemma B.5 with this value of M. On the event $\mathcal{G}$, we have

$$R' = R_{\mathrm{cl}} = R_{\mathrm{pol}} + R_{\mathrm{rem}}, \qquad \|R_{\mathrm{rem}}\|_{\infty\to\infty} \leq C\tau, \qquad \|R_{\mathrm{pol}}\|_{\infty\to\infty} + \|R_{\mathrm{cl}}\|_{\infty\to\infty} \leq C,$$

where

$$\tau = \left(C\frac{\log n}{\sqrt{np}}\right)^{K_0 M+1}.$$

Expand

$$e_u^\top R_{\mathrm{pol}}^a s = \sum_{v_1,\ldots,v_a}\prod_{t=1}^{a}(R_{\mathrm{pol}})_{v_{t-1}v_t}s_{v_a}, \qquad v_0 = u.$$

The level of a path monomial is the sum of the levels of its a atoms. Since all atoms have level at least one, every path monomial has level at least a. Let $Z_{u,a}^{(M)}$ be the sum of all level-a path monomials and define

$$\mathcal{E}_{u,a}^{(M)} := e_u^\top R'^a s - Z_{u,a}^{(M)}.$$

We first bound $Z_{u,a}^{(M)}$. Every monomial appearing in $|Z_{u,a}^{(M)}|^{2r}$ has $2M = 2ar$ residual positions and total level $2M$. Applying Lemma B.6 with $h = 0$ gives

$$\mathbb{E}|Z_{u,a}^{(M)}|^{2r} \leq n^{-r}(CM)^M(np)^{-M}.$$

On $\mathcal{G}$, decompose

$$\mathcal{E}_{u,a}^{(M)} = \mathcal{E}_{\mathrm{pol}} + \mathcal{E}_{\mathrm{Tay}},$$

where $\mathcal{E}_{\mathrm{pol}}$ is the sum of polynomial path monomials of level at least $a+1$, and $\mathcal{E}_{\mathrm{Tay}}$ contains all paths with at least one R_{rem} factor. In $|\mathcal{E}_{\mathrm{pol}}|^{2r}$, terms of total level $2M + h$ with $h \geq 1$ contribute at most

$$n^{-r}(CM)^M(np)^{-M}\left(\frac{CM}{\sqrt{np}}\right)^h$$

by Lemma B.6. Since $M \leq A_0 \log n$ and $np \geq C_d \log^B n$ with $B > 4$, increasing C_d gives

$$\frac{CM}{\sqrt{np}} \leq \frac{1}{2}.$$

Summing the geometric series over $h \geq 1$ yields

$$\mathbb{E}|\mathcal{E}_{\mathrm{pol}}|^{2r}\mathbf{1}_{\mathcal{G}} \leq n^{-r}(CM)^M(np)^{-M}.$$

For the Taylor-remainder part, the row-sum bounds from Lemma B.5 imply

$$\mathcal{E}_{\mathrm{Tay}} = e_u^\top (R_{\mathrm{cl}}^a - R_{\mathrm{pol}}^a)s = \sum_{t=0}^{a-1} e_u^\top R_{\mathrm{cl}}^t R_{\mathrm{rem}} R_{\mathrm{pol}}^{a-1-t} s,$$

and therefore, using $\|s\|_\infty = n^{-1/2}$,

$$|\mathcal{E}_{\mathrm{Tay}}|\mathbf{1}_{\mathcal{G}} \le \frac{aC^a\tau}{\sqrt{n}}.$$

Hence

$$\mathbb{E}\big[|\mathcal{E}_{\mathrm{Tay}}|^{2r}\mathbf{1}_{\mathcal{G}}\big] \le n^{-r}(Ca)^{2r}\tau^{2r}.$$

Because $M = ar$, $(Ca)^{2r} \le (CM)^M$. Choosing K_0 large and then C_d large gives

$$\tau^2 \le (np)^{-M},$$

so

$$\mathbb{E}\big[|\mathcal{E}_{\mathrm{Tay}}|^{2r}\mathbf{1}_{\mathcal{G}}\big] \le n^{-r}(CM)^M(np)^{-M}.$$

It remains to handle $\mathcal{G}^c$. Under the conventions before (1), the true and clipped residual matrices have absolute row sums at most n^C. Thus, after taking the $2r$-th power,

$$\big|e_u^\top R'^a s\big|^{2r} + \big|e_u^\top R_{\mathrm{cl}}^a s\big|^{2r} \le \exp(Car\log n) = \exp(CM\log n).$$

Since $M \le A_0 \log n$, $r \le M$, and $np \le n$, the target scale is at least $\exp(-C_{A_0}\log^2 n)$. Taking the constant in the definition of $\mathcal{G}$ sufficiently large gives

$$\exp(CM\log n)\mathbb{P}(\mathcal{G}^c) \le n^{-r}(CM)^M(np)^{-M}.$$

Combining the Z, polynomial-error, Taylor-error, and bad-event bounds, and absorbing fixed factors from $|x+y+z|^{2r}$ into $(CM)^M$, proves

$$\mathbb{E}|\mathcal{E}_{u,a}^{(M)}|^{2r} \le n^{-r}(CM)^M(np)^{-M}.$$

Together with the bound for $Z_{u,a}^{(M)}$, this proves the lemma. □

Lemma B.8. *Let*

$$R' := \widehat{A} - \eta ss^\top, \qquad s_i = \begin{cases} n^{-1/2}, & i \in S, \\ -n^{-1/2}, & i \in T, \end{cases} \qquad \eta := \frac{(p-q)n}{2d}.$$

Assume

$$p \ge C_d \frac{\log^B n}{n}$$

for a fixed $B > 4$ and sufficiently large $C_d = C_d(B)$. Let $A_0 > 0$ be fixed. Then there exists a constant $C = C(A_0, B) > 0$ such that, for every $u \in [n]$ and every pair of integers $a, r \ge 1$ satisfying $ar \le A_0 \log n$,

$$\mathbb{E}\big|e_u^\top R'^a s\big|^{2r} \le n^{-r}\left(\frac{Car}{np}\right)^{ar}.$$

Proof. Set $M := ar$. By Lemma B.7,

$$e_u^\top R'^a s = Z_{u,a}^{(M)} + \mathcal{E}_{u,a}^{(M)},$$

with

$$\mathbb{E}|\mathcal{E}_{u,a}^{(M)}|^{2r} \le n^{-r}(CM)^M(np)^{-M}, \qquad \mathbb{E}|Z_{u,a}^{(M)}|^{2r} \le n^{-r}(CM)^M(np)^{-M}.$$

After increasing C,

$$n^{-r}(CM)^M(np)^{-M} \le n^{-r}\left(\frac{Car}{np}\right)^{ar}.$$

The elementary inequality $|x+y|^{2r} \le 2^{2r-1}(|x|^{2r} + |y|^{2r})$ and the fact that $r \le M \le A_0 \log n$ allow the factor 2^{2r} to be absorbed into the constant C. This gives the claimed bound. □

Lemma B.9. *Let*

$$R' := \widehat{A} - \eta s s^\top, \qquad \eta := \frac{(p-q)n}{2d}.$$

Assume

$$p \geq C_d \frac{\log^B n}{n}, \qquad 1 \leq k \leq c \log n,$$

where $B > 4$ is fixed, $C_d = C_d(B)$ is sufficiently large, and $c > 0$ is sufficiently small. Then there exists a constant $K > 0$ such that, with probability at least $1 - n^{-\Omega(1)}$,

$$\forall u \in [n], \quad \forall 1 \leq a \leq k : \qquad \left| e_u^\top R'^a s \right| \leq \frac{(K\sqrt{\log n/(np)})^a}{\sqrt{n}}.$$

Proof. Fix $u \in [n]$ and $1 \leq a \leq k$. Choose

$$r_a := \left\lceil \frac{A_1 \log n}{a} \right\rceil,$$

where A_1 is a sufficiently large absolute constant. Since $a \leq k \leq c \log n$, by taking $c \leq A_1$ we have

$$A_1 \log n \leq a r_a \leq (A_1 + c) \log n \leq 2A_1 \log n.$$

Applying Lemma B.8 with its fixed parameter $A_0 = 2A_1$ and with $r = r_a$ gives

$$\mathbb{E}|e_u^\top R'^a s|^{2r_a} \leq n^{-r_a} \left(\frac{C a r_a}{np} \right)^{a r_a}.$$

Let

$$T_a := \frac{(K\sqrt{\log n/(np)})^a}{\sqrt{n}}.$$

By Markov's inequality,

$$\mathbb{P}(|e_u^\top R'^a s| > T_a) \leq \left(\frac{C a r_a}{K^2 \log n} \right)^{a r_a}.$$

Since $a r_a \asymp A_1 \log n$, choosing A_1 fixed and then K sufficiently large in terms of A_1 yields

$$\mathbb{P}(|e_u^\top R'^a s| > T_a) \leq n^{-30}.$$

A union bound over all $u \in [n]$ and all $1 \leq a \leq k \leq c \log n$ gives failure probability at most $n \log n \cdot n^{-30} \leq n^{-28}$. This proves the claimed uniform bound. □

Lemma B.10. *Let $\eta > 0$ and $\theta := C\sqrt{\frac{\log n}{np}}$. Assume the decomposition*

$$\widehat{A} = \eta s s^\top + R',$$

satisfies the spectral norm bound $\|R'\| \leq \theta$ and the entrywise Krylov bound

$$|e_u^\top R'^a s| \leq \frac{\theta^a}{\sqrt{n}}$$

for all $u \in [n]$ and $1 \leq a \leq k$. Let $g \sim N(0, \sigma'^2 I_n)$ be independent of the graph. If $1 \leq k \leq C_0 \log n$ and $k\theta/\eta \leq c$ for a sufficiently small constant $c > 0$, then, with probability at least $1 - n^{-\Omega(1)}$ conditional on the graph event above,

$$\left\| \eta^{-k} \widehat{A}^k (s + g) - s \right\|_\infty \leq C \left(\frac{k\theta/\eta}{\sqrt{n}} + \frac{\sigma' \sqrt{\log n}}{\sqrt{n}} + \left(\frac{\theta}{\eta} \right)^k \sigma' \sqrt{\log n} \right).$$

Proof. Write $P := ss^\top$, so that $P^2 = P$, $Ps = s$, and $\widehat{A} = \eta P + R'$. We first control the signal part. Expand $(\eta P + R')^k s$ into words in the two letters ηP and R'. The all-projector word equals $\eta^k s$. We claim that every other word with exactly $j \geq 1$ copies of R' contributes, after division by η^k, at most

$$\frac{1}{\sqrt{n}} \left(\frac{\theta}{\eta}\right)^j$$

in ℓ_∞-norm. If the word has no P, then it is $R'^k s$, and the claim is exactly the assumed Krylov bound.

Now suppose the word contains at least one P. After deleting consecutive projectors and using $Ps = s$, the action of the word on s is a product of blocks of the form R'^a separated by projectors P. Each interior projector turns the vector to a multiple of s:

$$PR'^a s = s(s^\top R'^a s).$$

The scalar is bounded by

$$|s^\top R'^a s| \leq \|R'^a s\|_2 \leq \theta^a,$$

and the last block, if it is not followed by a projector on the left, is controlled coordinatewise by

$$\|R'^a s\|_\infty \leq \frac{\theta^a}{\sqrt{n}}$$

from the Krylov assumption. Thus a word with j total R'-factors has coordinate size at most $\eta^{k-j}\theta^j/\sqrt{n}$. Dividing by η^k proves the claim. Summing over the at most $\binom{k}{j}$ choices of the R'-positions gives

$$\left\|\eta^{-k}\widehat{A}^k s - s\right\|_\infty \leq \frac{C}{\sqrt{n}} \sum_{j=1}^{k} \binom{k}{j} \left(\frac{\theta}{\eta}\right)^j \leq \frac{Ck\theta/\eta}{\sqrt{n}}.$$

Next we control the Gaussian part. Conditional on the graph, the pure projector word gives

$$\eta^{-k}(\eta P)^k g = Pg = s(s^\top g).$$

Since $s^\top g \sim N(0, \sigma'^2)$, a standard Gaussian tail bound gives

$$\|Pg\|_\infty \leq C\frac{\sigma'\sqrt{\log n}}{\sqrt{n}}$$

with probability at least $1 - n^{-\Omega(1)}$.

For the pure R'-word, for every u,

$$e_u^\top R'^k g \sim N\left(0, \sigma'^2 \|e_u^\top R'^k\|^2\right), \qquad \|e_u^\top R'^k\| \leq \|R'\|^k \leq \theta^k.$$

A union bound over $u \in [n]$ yields

$$\max_u |e_u^\top R'^k g| \leq C\sigma'\theta^k\sqrt{\log n}$$

with probability at least $1 - n^{-\Omega(1)}$. Hence the pure R'-word contributes at most

$$C\left(\frac{\theta}{\eta}\right)^k \sigma'\sqrt{\log n}.$$

It remains to check the mixed words, namely those containing at least one projector and at least one R'. Fix such a word W with j copies of R'. Let ηP be the rightmost projector letter in W. Then

$$Wg = U\,(\eta P)R'^a g = \eta\, Us\,(s^\top R'^a g),$$

where $0 \le a \le j$ is the number of consecutive R'-letters to the right of this projector, and U is the prefix to its left. The prefix U contains exactly $j-a$ copies of R' and $k-j-1$ copies of ηP. For the scalar,

$$s^\top R'^a g \sim N\big(0, \sigma'^2 \|s^\top R'^a\|^2\big), \qquad \|s^\top R'^a\| \le \theta^a$$

with the convention $\theta^0 = 1$. Hence, by a union bound over $0 \le a \le k$,

$$|s^\top R'^a g| \le C\sigma'\theta^a \sqrt{\log n}$$

with probability at least $1 - n^{-\Omega(1)}$. Applying the signal-word bound proved above, with k replaced by the length of the prefix, gives

$$\|Us\|_\infty \le \frac{\eta^{k-j-1}\theta^{j-a}}{\sqrt{n}},$$

because U has $j-a$ residual letters and $k-j-1$ projector letters. Therefore, after division by η^k, the word contributes at most

$$C\frac{\sigma'\sqrt{\log n}}{\sqrt{n}}\left(\frac{\theta}{\eta}\right)^j.$$

There are at most $\binom{k}{j}$ such words with j copies of R'. Consequently the total mixed contribution is bounded by

$$C\frac{\sigma'\sqrt{\log n}}{\sqrt{n}}\sum_{j=1}^{k-1}\binom{k}{j}\left(\frac{\theta}{\eta}\right)^j \le C\frac{\sigma'\sqrt{\log n}}{\sqrt{n}},$$

where the last inequality follows from $(1+\theta/\eta)^k - 1 \le 2k\theta/\eta \le C$ under $k\theta/\eta \le c$. Combining the signal estimate, the pure projector noise estimate, the mixed Gaussian estimate, and the pure $R'^k g$ estimate proves the lemma. □

To prove our multi-class results (Theorem 4.7), we first characterize the unnormalized convolution matrix $\tilde{A}$ in terms of its expected behavior and a bounded random perturbation. These fundamental characterizations for the unnormalized setting were established by Wang et al. (2024).

Lemma B.11 (Multi-class Characterization for $\tilde{A}$ (Wang et al., 2024))**.** *In the multi-class setting, the unnormalized convolution matrix, $\tilde{A}$, can be decomposed as $\tilde{A} = M + R'$ where:*

- *M is symmetric and has rank $L-1$: its nonzero eigenvalues are all equal to $\lambda = \frac{(p-q)n}{dL}$, and the remaining eigenvalues are zero. Also, $MU = \lambda U$.*
- *R' is a random matrix such that with probability at least $1 - n^{-\Omega(1)}$, $\|R'\| \le \delta$ where $\delta = C(\frac{1}{d}(\sqrt{\frac{np(1-p)}{L}} + \sqrt{nq(1-q)}))$.*

We also need to bound the operator norm distance between the kth convolution and M^k, which directly follows from the properties of $\tilde{A}$.

Lemma B.12 (Operator Norm Bound for $\tilde{A}^k$ (Wang et al., 2024))**.** *Suppose $|\lambda| > 4k\delta$. Then with high probability, we have*

$$\left\|\frac{1}{\lambda^k}(\tilde{A}^k - M^k)\right\| \le \frac{2k\delta}{|\lambda|}.$$

Lemma B.13 (Normalization perturbation for multi-class CSBM)**.** *In the multi-class CSBM, let*

$$\bar{d} = \frac{pn}{L} + \frac{(L-1)qn}{L}$$

and assume $\bar{d} \ge C\log n$. Then, with probability at least $1 - n^{-\Omega(1)}$,

$$\|\widehat{A} - \widetilde{A}\| \le C\sqrt{\frac{\log n}{\bar{d}}}.$$

Proof. Let

$$\epsilon = C_0\sqrt{\frac{\log n}{\bar{d}}}.$$

For every vertex v,

$$\mathbb{E}d_v = (n/L-1)p + (n-n/L)q = \bar{d} - p,$$

so $\mathbb{E}d_v = \bar{d} + O(p)$. Since $\bar{d} \geq C\log n$, Chernoff's inequality and a union bound give, with probability $1 - n^{-\Omega(1)}$,

$$d_v \in [\bar{d}(1-\epsilon), \bar{d}(1+\epsilon)] \quad \text{for all } v, \qquad d = \frac{1}{n}\sum_v d_v \in [\bar{d}(1-\epsilon), \bar{d}(1+\epsilon)]$$

after increasing C_0.

On this event,

$$\widehat{A} - \widetilde{A} = \left(D^{-1/2}AD^{-1/2} - \frac{1}{d}A\right) - \left(\frac{D^{1/2}\mathbb{1}\mathbb{1}^\top D^{1/2}}{\mathbb{1}^\top D\mathbb{1}} - \frac{1}{n}\mathbb{1}\mathbb{1}^\top\right).$$

For the first term,

$$\left|\frac{1}{\sqrt{d_u d_v}} - \frac{1}{d}\right| \leq \frac{C\epsilon}{\bar{d}}.$$

Since $d_u \leq 2\bar{d}$, every row has absolute row sum at most $C\epsilon$. The matrix is symmetric, hence its operator norm is at most $C\epsilon$. For the second term,

$$\left|\frac{\sqrt{d_u d_v}}{\sum_w d_w} - \frac{1}{n}\right| \leq \frac{C\epsilon}{n},$$

so every row has absolute row sum at most $C\epsilon$, and again the operator norm is at most $C\epsilon$. Combining the two estimates proves the lemma. □

C Proofs of Main Theorems

C.1 Proof of Theorem 4.3

Proof. Let

$$s_i = \begin{cases} n^{-1/2}, & i \in S, \\ -n^{-1/2}, & i \in T, \end{cases} \qquad \eta = \frac{(p-q)n}{2d}, \qquad R' = \widehat{A} - \eta ss^\top.$$

Then $\widehat{A} = \eta ss^\top + R'$. By Lemma B.4, and by choosing the constant hidden in $p \geq \Omega_B(\log^B n/n)$ sufficiently large, with probability $1 - n^{-\Omega(1)}$,

$$|\eta - \gamma| \leq C\frac{\sqrt{\log n}}{n\sqrt{p}}, \qquad \|R'\| \leq C\sqrt{\frac{\log n}{np}} =: \theta.$$

By the graph-signal assumption, with the constant hidden in $\Omega_B(\cdot)$ chosen sufficiently large, $k\theta/\gamma$ is smaller than the absolute constant required below. Increasing constants if necessary gives $\eta \asymp \gamma$ and $\theta/\eta \leq 2\theta/\gamma$.

By Lemma B.2, there exists $w \in \mathbb{R}^m$ such that

$$Xw = s + g, \qquad g \sim N(0, \sigma'^2 I_n), \qquad \sigma' = \frac{2\sigma}{\sqrt{n}\,\|\mu - \nu\|}.$$

Hence $\widehat{A}^k Xw = \widehat{A}^k(s+g)$. Lemma B.9 gives, on the same high-probability graph event,

$$|e_u^\top R'^a s| \leq \frac{\theta^a}{\sqrt{n}}, \qquad \forall u \in [n],\ 1 \leq a \leq k,$$

after increasing θ by an absolute factor. The use of Lemma B.9 is legitimate because the constant hidden in $k = O_B(\log n)$ is chosen no larger than the constant c in that lemma. Applying Lemma B.10 therefore yields

$$\left\|\eta^{-k}\widehat{A}^k(s+g) - s\right\|_\infty \le C\left(\frac{k\theta/\eta}{\sqrt{n}} + \frac{\sigma'\sqrt{\log n}}{\sqrt{n}} + \left(\frac{\theta}{\eta}\right)^k \sigma'\sqrt{\log n}\right).$$

The first term is at most $1/(10\sqrt{n})$ by the graph-signal assumption. Since $\eta \asymp \gamma$, the constant C_B in the SNR condition is chosen larger than the absolute constants hidden in θ/η. The two remaining terms are therefore at most $1/(10\sqrt{n})$ each by the assumed lower bound on $\|\mu - \nu\|/\sigma$, together with

$$\sigma' = \frac{2\sigma}{\sqrt{n}\,\|\mu - \nu\|}.$$

Thus

$$\left\|\eta^{-k}\widehat{A}^k(s+g) - s\right\|_\infty < \frac{1}{2\sqrt{n}}.$$

Every coordinate of s equals $\pm n^{-1/2}$, so the signs of $\eta^{-k}\widehat{A}^k(s+g)$, and hence of $\widehat{A}^k(s+g)$, agree with the true labels. Since $\widehat{A}^k X w = \widehat{A}^k(s+g)$, the hyperplane with normal vector w and threshold zero separates the rows of $\widehat{A}^k X$. A union bound over the graph and Gaussian events gives failure probability $n^{-\Omega(1)}$. □

C.2 Proof of Theorem 4.7

Proof. We work on the high-probability graph event on which

$$\widetilde{A} = M + R, \qquad \|R\| \le \delta,$$

from Lemma B.11. Because L is fixed, the probability losses and constants in the multi-class spectral estimates are absorbed into $n^{-\Omega(1)}$ and the displayed absolute constants. By Lemma B.13, on the same event after a union bound, the degree-normalization perturbation satisfies

$$E_{\text{norm}} := \widehat{A} - \widetilde{A}, \qquad \|E_{\text{norm}}\| \le \epsilon.$$

Hence

$$\widehat{A} = M + H, \qquad H := R + E_{\text{norm}}, \qquad \|H\| \le \delta + \epsilon.$$

Since M has rank $L-1$ and satisfies $MU = \lambda U$, we also have $M^k U = \lambda^k U$.

Let $X = U + G$, where G has independent $N(0, \sigma^2)$ entries. Then

$$X^{(k)} - U = \lambda^{-k}\widehat{A}^k U - U + \lambda^{-k}\widehat{A}^k G.$$

Using $\|A+B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$, it is enough to control the signal term and the Gaussian term separately.

Signal term. For the signal part,

$$\lambda^{-k}\widehat{A}^k U - U = \lambda^{-k}(\widehat{A}^k - M^k)U.$$

We decompose

$$\widehat{A}^k - M^k = (\widehat{A}^k - \widetilde{A}^k) + (\widetilde{A}^k - M^k).$$

Lemma B.12 gives

$$\left\|\lambda^{-k}(\widetilde{A}^k - M^k)\right\| \le \frac{2k\delta}{|\lambda|}.$$

For the normalization part, use the telescoping identity

$$\widehat{A}^k - \widetilde{A}^k = \sum_{t=0}^{k-1} \widehat{A}^t(\widehat{A} - \widetilde{A})\widetilde{A}^{k-1-t}.$$

Moreover,

$$\|\widehat{A}\| \le |\lambda| + \delta + \epsilon, \qquad \|\widetilde{A}\| \le |\lambda| + \delta.$$

Therefore

$$\left\|\lambda^{-k}(\widehat{A}^k - \widetilde{A}^k)\right\| \le \frac{k\epsilon}{|\lambda|}\left(1 + \frac{\delta+\epsilon}{|\lambda|}\right)^{k-1} \le C\frac{k\epsilon}{|\lambda|},$$

where the last inequality uses $k(\delta+\epsilon)/|\lambda| \le C_0^{-1}$. Consequently,

$$\left\|\lambda^{-k}(\widehat{A}^k - M^k)\right\| \le C\frac{k(\delta+\epsilon)}{|\lambda|}.$$

This bound is dominated by

$$C\left[\left(\frac{\epsilon}{|\lambda|}\right)^k + 2k\frac{\delta+\epsilon}{|\lambda|}\right],$$

so

$$\left\|\lambda^{-k}\widehat{A}^k U - U\right\|_F^2 \le C\left[\left(\frac{\epsilon}{|\lambda|}\right)^{2k} + 4k^2\left(\frac{\delta+\epsilon}{|\lambda|}\right)^2\right]\|U\|_F^2.$$

Gaussian term. Conditional on the graph, set $B := \lambda^{-k}\widehat{A}^k$. Since G has independent Gaussian columns $g_1, \dots, g_m$,

$$\|BG\|_F^2 = \sum_{r=1}^{m} g_r^\top B^\top B g_r.$$

A standard Gaussian quadratic-form concentration bound gives, with probability at least $1 - n^{-\Omega(1)}$,

$$\|BG\|_F^2 \le C\sigma^2 m \log n \ \mathrm{Tr}(B^\top B).$$

Here the assumption $m \le n^{C_m}$ allows the Gaussian tail probability to be union bounded over the m feature coordinates while keeping a $\log n$ factor. Because $\widehat{A}$ is symmetric,

$$\mathrm{Tr}(B^\top B) = \frac{1}{|\lambda|^{2k}}\mathrm{Tr}(\widehat{A}^{2k}).$$

Since M is symmetric with rank $L-1$ and $\|M\| = |\lambda|$, Weyl's inequality for singular values gives

$$s_j(\widehat{A}) \le s_j(M) + \|H\|.$$

More explicitly, for $1 \le j \le L-1$,

$$s_j(\widehat{A}) \le |\lambda| + \|H\| \le |\lambda| + \delta + \epsilon,$$

while the rank inequality

$$s_{j+L-1}(M+H) \le s_j(H)$$

implies $s_j(\widehat{A}) \le \delta + \epsilon$ for every $j \ge L$. Thus at most $L-1$ singular values are bounded by $|\lambda| + \delta + \epsilon$, and all remaining singular values are bounded by $\delta + \epsilon$. Because $\widehat{A}$ is symmetric, $\mathrm{Tr}(\widehat{A}^{2k}) = \sum_{j=1}^{n} s_j(\widehat{A})^{2k}$. Hence

$$\frac{1}{|\lambda|^{2k}}\mathrm{Tr}(\widehat{A}^{2k}) \le \left(1 + \frac{\delta+\epsilon}{|\lambda|}\right)^{2k}(L-1) + n\left(\frac{\delta+\epsilon}{|\lambda|}\right)^{2k}.$$

Thus

$$\left\|\lambda^{-k}\widehat{A}^k G\right\|_F^2 \le C\left[\left(1 + \frac{\delta+\epsilon}{|\lambda|}\right)^{2k}(L-1) + n\left(\frac{\delta+\epsilon}{|\lambda|}\right)^{2k}\right]\sigma^2 m \log n.$$

Counting bad vertices. Let n_e be the number of vertices satisfying $\|x_i^{(k)} - \mu_{\ell(i)}\| \geq \Delta/2$. Then

$$n_e \frac{\Delta^2}{4} \leq \|X^{(k)} - U\|_F^2.$$

Combining the two bounds above and absorbing absolute constants gives

$$\begin{aligned} n_e \leq O\Bigg(& \left[\left(\frac{\epsilon}{|\lambda|} \right)^{2k} + 4k^2 \left(\frac{\delta + \epsilon}{|\lambda|} \right)^2 \right] \frac{\|U\|_F^2}{\Delta^2} \\ & + \left[\left(1 + \frac{\delta + \epsilon}{|\lambda|} \right)^{2k} (L-1) + n \left(\frac{\delta + \epsilon}{|\lambda|} \right)^{2k} \right] \frac{\sigma^2 m \log n}{\Delta^2} \Bigg). \end{aligned}$$

Every vertex outside this exceptional set is within distance $\Delta/2$ of its true class center, and hence is closer to its own center than to any other class center. This proves the claimed partial recovery bound. □