Aggregation Dispersion: An Information-Geometric Diagnostic of Oversmoothing in Graph Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Message-passing neural networks (MPNNs) learn node representations by iteratively aggregating information from neighbors, but stacking too many layers causes all representations to converge, a phenomenon known as oversmoothing. We propose a geometric perspective on this phenomenon. At each depth k, the propagation operator assigns to every node v a walk distribution $p_v^{(k)}$, a probability distribution over all nodes describing how v distributes its attention across the graph. These walk distributions form a point cloud on the probability simplex, and oversmoothing is the contraction of this cloud to a single point. To measure this contraction, we exploit the intrinsic geometry of the simplex: the Fisher–Rao metric, the unique Riemannian metric invariant under sufficient statistics, whose chordal distance under the square-root embedding $p \mapsto \sqrt{p}$ is the Hellinger distance. The resulting diagnostic, the *aggregation dispersion* $\mathcal{D}_k$, is the variance of the embedded point cloud on the unit sphere. We prove that: (i) $\mathcal{D}_k = 0$ if and only if the k-th step propagation matrix has rank one, providing a necessary and sufficient condition for representational collapse; (ii) $\mathcal{D}_k$ is monotonically non-increasing with depth, so each layer irreversibly spends a finite diversity budget; (iii) the spectral gap of the propagation matrix controls the rate of this decay; and (iv) for contractive MPNNs, the aggregation dispersion upper-bounds the representation dispersion up to architecture-dependent constants, connecting the geometry of the simplex to the geometry of the feature space. In experiments across 14 model configurations and 11 datasets, $\mathcal{D}_k$ achieves the highest average Pearson correlation with accuracy degradation among all training-free diagnostics ($|r| = 0.721$), outperforming the post-training effective rank ($|r| = 0.688$).

1 Introduction

Message-passing neural networks (MPNNs) [Gilmer et al., 2017] learn node representations by iteratively aggregating information from neighbouring nodes in a graph $G = (V, E)$ with $n = |V|$ nodes. This aggregation is governed by a *propagation matrix* $\mathbf{P} \in \mathbb{R}^{n \times n}$, a non-negative matrix whose entry P_{uv} prescribes the weight that node u assigns to the message received from node v. When every row of $\mathbf{P}$ sums to one, a property called *row-stochasticity*, each row is a probability distribution over V. After k layers of propagation, the effective influence of node v on the rest of the graph is captured by the v-th column of the k-th matrix power, $p_v^{(k)} = [\mathbf{P}^k]_{v,:}$, which we call the *walk distribution* from v at depth k. Each walk distribution is a point on the *probability simplex* $\Delta^{n-1} = \{p \in \mathbb{R}^n : p \geq 0,\ \mathbf{1}^\top p = 1\}$, the geometric space of all probability distributions over n outcomes.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

On a connected graph, the Perron–Frobenius theorem guarantees that every point on this simplex converges, as $k \to \infty$, to a unique *stationary distribution* π, the eigenvector of $\mathbf{P}$ associated with eigenvalue one [Horn and Johnson, 2012]. This convergence means that the n walk distributions $\{p_v^{(k)}\}_{v \in V}$ contract toward a single point π on Δ^{n-1}. At the point of collapse the matrix $\mathbf{P}^k$ has rank one, so every node receives an identical aggregated signal and no downstream architecture, however expressive, can recover distinct node representations. This phenomenon is known as *oversmoothing* [Li et al., 2018, Oono and Suzuki, 2020].

Detecting oversmoothing before it erases useful information is therefore a central challenge. Existing diagnostics, however, operate only *after* training: the Dirichlet energy [Cai and Wang, 2020, Giovanni et al., 2023] measures the smoothness of the learned node features, while the effective rank [Zhang et al., 2026] tracks the number of significant singular values of the representation matrix. These metrics capture the *effect* of oversmoothing on learned representations, not its *structural cause*, the contraction of walk distributions on the simplex. To expose this cause directly, one needs a training-free, depth-resolved measure defined entirely in terms of the propagation matrix $\mathbf{P}$.

The key observation is that $\mathbf{P}$ already endows the simplex with a rich geometric structure that can be exploited for exactly this purpose. The interior of the simplex carries a canonical Riemannian metric, the *Fisher–Rao metric* [Amari, 2016], singled out by Čencov's theorem as the unique (up to scale) metric invariant under sufficient statistics, so that distances reflect genuine statistical differences rather than artifacts of the parameterization. Under the square-root embedding $\varphi\colon p \mapsto (\sqrt{p_1}, \dots, \sqrt{p_n})$, the simplex $(\Delta^{n-1}, \frac{1}{4} g_{\mathrm{FR}})$ maps isometrically into the unit sphere $\mathbb{S}^{n-1}$, and the chordal distance between embedded points is the *Hellinger distance* $H(p,q) = \frac{1}{\sqrt{2}} \|\varphi(p) - \varphi(q)\|_2$, a numerically stable proxy for the Fisher–Rao geodesic that we develop in detail in §3.

Measuring the Hellinger distance among all pairs of walk distributions yields a single scalar that we call the *aggregation dispersion* at depth k:

$$\mathcal{D}_k \;=\; \frac{1}{n^2} \sum_{u,v} H^2\big(p_u^{(k)}, p_v^{(k)}\big) \;=\; 1 - \Big\| \tfrac{1}{n} \textstyle\sum_v \sqrt{p_v^{(k)}} \Big\|_2^2. \tag{1}$$

The left-hand side is the mean pairwise squared Hellinger distance among all walk distributions. The right-hand side reformulates this mean as the variance of the Hellinger-embedded point cloud on $\mathbb{S}^{n-1}$, reducing the cost from $O(n^3)$ for the naïve pairwise computation to $O(n^2)$. Because the Hellinger distance is sensitive to differences in both support and probability mass, $\mathcal{D}_k$ captures all sources of diversity loss among the walk distributions in a single scalar.

Contributions. We develop the theory and practice of aggregation dispersion in three parts. *First* (§5.1), we prove that $\mathcal{D}_k = 0$ if and only if $\mathbf{P}^k$ has rank one (Theorem 5.1), establishing that aggregation dispersion is a tight characterisation of oversmoothing. We further show that $\mathcal{D}_k$ is monotonically non-increasing in k via the data-processing inequality (Theorem 5.2), confirming that repeated message passing can only destroy distributional diversity, never create it. *Second* (§5.2–5.3), we link this monotonic decay to the spectrum of $\mathbf{P}$: the spectral gap controls the rate at which $\mathcal{D}_k$ vanishes (Theorem 5.3). For contractive MPNNs, we further show that the aggregation dispersion upper-bounds the dispersion of the learned representations (Theorems 5.5 and 5.6), connecting the geometry of the simplex to the geometry of the feature space. *Third* (§6), we demonstrate empirically that our proposed metric $\mathcal{D}_k$ achieves the highest correlation with accuracy degradation ($|r| = 0.721$), outperforming post-training metrics including the effective rank of Zhang et al. [2026] ($|r| = 0.688$) across 14 model configurations and 11 datasets.

2 Related Work

Prior work on the depth limitations of message-passing neural networks has largely focused on two complementary phenomena: oversmoothing and oversquashing, along with corresponding diagnostics and mitigation strategies. We position our work within the "oversmoothing" literature and introduce a novel perspective by leveraging information geometry to diagnose it.

Oversmoothing. Li et al. [2018] first identified oversmoothing, observing that node representations in graph convolutional networks become indistinguishable as depth increases. Oono and Suzuki

[2020] formalized this observation by proving that, under standard assumptions, the distance between any two node representations contracts exponentially with depth. Subsequent work quantified the rate of this contraction through the *Dirichlet energy*, the sum of squared feature differences across edges, showing that it decays to zero as layers are added [Cai and Wang, 2020]. More recently, Rusch et al. [2023] provided a comprehensive survey of these asymptotic results, while Keriven [2022] showed that some smoothing is in fact beneficial for learning, and that performance degrades only when depth is pushed past an optimal point. A key insight shared by all these analyses is that oversmoothing is ultimately rooted in the topology of the input graph: the connectivity pattern dictates how fast walk distributions converge .

Post-training diagnostics. The diagnostic described above, Dirichlet energy, is computed from the feature matrices and is known to fail as oversmoothing indicator for separately trained GNNs of moderate depth. Zhang et al. [2026] proposed two alternatives, the *effective rank* "erank" and the *numeric rank* "nrank", both derived from the singular-value spectrum of the representation matrix. These rank-based measures track the effective dimensionality of the learned features and correlate strongly with accuracy degradation across diverse architectures and datasets, even in regimes where energy-based metrics show no visible change. Their evaluation protocol pairs GCN [Kipf and Welling, 2017] and GAT [Veličković et al., 2018] with LeakyReLU activations at eight depths across 7 datasets. We extend their protocol to 14 model configurations spanning 11 datasets and show that a training-free measure achieves higher average correlation, demonstrating that oversmoothing can be diagnosed from the propagation matrix alone, without ever computing a forward pass.

Information geometry in machine learning. Our diagnostic rests on objects from information geometry: the Fisher–Rao metric, the Hellinger distance, and the data-processing inequality [Amari, 2016]. The Fisher–Rao metric underpins natural gradients [Amari and Douglas, 1998, Martens, 2020], loss-landscape and data-manifold analyses [Sun and Nielsen, 2025, Ed-dib et al., 2025], and geometrically principled generative models [Arvanitidis et al., 2018]. To our knowledge, this is the first application of Hellinger geometry on the probability simplex to diagnosing oversmoothing.

Oversmoothing mitigation. Complementary work proposes architectural fixes: initial residuals and identity mappings (GCNII, Chen et al., 2020b), random edge removal (DropEdge, Rong et al., 2020), pairwise-distance normalization (PairNorm, Zhao and Akoglu, 2020), and multi-layer aggregation (JK-Nets, Xu et al., 2018). We evaluate $\mathcal{D}_k$ on several of these in Appendix F.3.

3 Setup

We now formalise the objects introduced informally in §1 (graphs, propagation operators, walk distributions, and the Hellinger distance) and establish the notation used throughout the paper.

3.1 Graphs and Propagation

Let $G = (V, E)$ be a connected non-bipartite undirected graph with $n = |V|$ nodes, $m = |E|$ edges, adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$, and degree matrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$. To allow self-loops we set $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\tilde{\mathbf{D}} = \mathbf{D} + \mathbf{I}$, and write $D = \mathrm{diam}(G)$ for the diameter. Every standard MPNN architecture defines a *propagation matrix* $\mathbf{P} \in \mathbb{R}^{n\times n}$ that governs how information flows along edges: GCN uses $\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ [Kipf and Welling, 2017], GraphSAGE uses $\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}$ [Hamilton et al., 2017], and GAT uses learned attention coefficients [Veličković et al., 2018]. All share a *locality property*: $[\mathbf{P}^k]_{vu} \neq 0$ only if $d(v,u) \leq k$, so after k layers each node aggregates information from exactly its k-hop neighborhood.

3.2 The Canonical Walk Operator

To study oversmoothing as a structural property of message passing, we abstract away architectural details and analyze the *lazy random walk*

$$\mathbf{P} \;=\; \tfrac{1}{2}\big(\mathbf{I} + \mathbf{D}^{-1}\mathbf{A}\big), \tag{2}$$

a row-stochastic, reversible, aperiodic operator with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > 0$ [Chung, 1997, Ch. 1], spectral gap $\gamma = 1 - \lambda_2$, and stationary distribution $\pi_v = \frac{d_v}{2m}$) [Levin et al.,

2017]. We adopt $\mathbf{P}$ as a parameter-free baseline; Appendix E shows the analysis transfers to GCN, GIN, and GraphSAGE-mean up to architecture-dependent constants, and §6 confirms the diagnostic remains effective on non-reversible operators such as GAT.

3.3 Walk Distributions on the Probability Simplex

Since $\mathbf{P}$ is row-stochastic, each row of $\mathbf{P}^k$ defines a probability distribution over nodes. For any node $v \in V$, we define the *walk distribution*

$$p_v^{(k)} = [\mathbf{P}^k]_{v,:} \in \Delta^{n-1}, \qquad \Delta^{n-1} = \{p \in \mathbb{R}^n_{\geq 0} : \textstyle\sum_i p_i = 1\},$$

the endpoint distribution of a k-step lazy random walk initialized at v. This induces a trajectory on the probability simplex, where each application of $\mathbf{P}$ evolves the distribution from local concentration toward global equilibrium. At $k = 0$ the walk is degenerate, $p_v^{(0)} = \delta_v$, concentrated entirely at the starting node; as k grows, repeated application of $\mathbf{P}$ redistributes mass over nodes reachable via multi-hop connectivity. Since the induced Markov chain is finite, connected, and aperiodic, the ergodic theorem guarantees convergence to the unique stationary distribution, $p_v^{(k)} \to \boldsymbol{\pi}$ for every $v \in V$. Oversmoothing is therefore the contraction of the cloud $\{p_v^{(k)}\}_{v \in V}$ toward a single point on the simplex; to quantify the rate of this contraction, we next equip the simplex with a distance.

3.4 The Hellinger Distance

The choice of distance is not arbitrary: different metrics on Δ^{n-1} reflect different notions of dissimilarity between distributions, and not all are well-suited to the oversmoothing setting. A principled selection criterion comes from Čencov's theorem [Čencov, 1982], which establishes that the *Fisher–Rao metric* is the unique Riemannian metric on the simplex (up to a global constant) that is invariant under *sufficient statistics*, that is, transformations that preserve all information relevant to statistical inference. This invariance makes the Fisher–Rao metric a canonical local geometry on the simplex. The simplex endowed with this metric g_{FR} can be embedded isometrically into a standard sphere: the *square-root embedding* $\phi\colon p \mapsto \sqrt{p}$ maps each distribution to a point on the positive orthant of the unit sphere $S_+^{n-1} = \{x \in \mathbb{R}^n_{\geq 0} : \|x\|_2 = 1\}$ (since $\|\sqrt{p}\|_2^2 = \sum_i p_i = 1$) equipped with the metric inherited from the ambient Euclidean space $\mathbb{R}^n$. Under this embedding, the Fisher–Rao geodesic distance becomes the spherical arc-length distance, $d_{\mathrm{FR}}(p, q) = 2\arccos\langle\sqrt{p}, \sqrt{q}\rangle$. However, when p and q are close, $\arccos$ is numerically unstable near its boundary values. A natural alternative is the chordal distance between the embedded points, which gives the *Hellinger distance*

$$H(p, q) = \frac{1}{\sqrt{2}}\left\|\sqrt{p} - \sqrt{q}\right\|_2, \qquad H \in [0, 1]. \quad (3)$$

The Hellinger distance equals $1/\sqrt{2}$ times the Euclidean distance between $\sqrt{p}$ and $\sqrt{q}$, so computations reduce to inner products on S_+^{n-1}. Three properties make it well-suited to diagnosing oversmoothing: it satisfies the data-processing inequality [Cover and Thomas, 2005], yielding the monotonicity of $\mathcal{D}_k$ with depth (§5.1); it remains finite on disjoint supports, unlike the Kullback–Leibler or χ^2 divergences, which is essential at early depths; and it is metrically equivalent to total variation [Tsybakov, 2008], linking $\mathcal{D}_k$ to mixing-time theory [Levin et al., 2017]. Formal statements are in Appendix A.3.

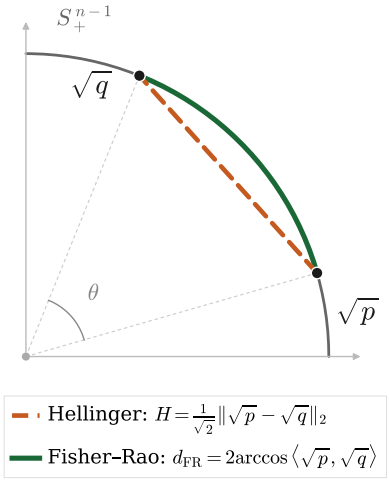

Figure 1: Under $\phi\colon p \mapsto \sqrt{p}$, the Hellinger distance is the dashed chord while the Fisher–Rao geodesic follows the arc on S_+^{n-1}. See Eq. (3).

4 The Aggregation Dispersion

The setup of §3 provides all the ingredients needed to define our diagnostic: a family of walk distributions $\{p_v^{(k)}\}_{v \in V}$ on the probability simplex, and the Hellinger distance to measure how far

apart they are. We now combine these ingredients into a single scalar that quantifies the diversity of the walk distributions at each depth.

Definition 4.1 (Aggregation dispersion)**.** *Let* $\mathbf{P}$ *be the lazy walk* (2) *on a connected graph with* n *nodes, and let* $p_v^{(k)} = [\mathbf{P}^k]_{v,:}$ *denote the walk distribution from node* v *at depth* k*. The* aggregation dispersion *at depth* k *is:*

$$\mathcal{D}_k \;:=\; \frac{1}{n^2} \sum_{u,v \in V} H^2\big(p_u^{(k)}, p_v^{(k)}\big) \;=\; 1 - \big\|\bar{\phi}^{(k)}\big\|_2^2, \tag{4}$$

where $\bar{\phi}^{(k)} = \frac{1}{n}\sum_{v=1}^{n} \sqrt{p_v^{(k)}}$ *is the* centroid *of the Hellinger-embedded walk distributions.*

The left-hand side of (4) is the mean pairwise squared Hellinger distance among all n^2 pairs of walk distributions: it is large when different nodes "see" different parts of the graph, and zero when every node sees the same distribution. The right-hand side reformulates this mean in terms of the centroid $\bar{\phi}^{(k)}$, whose squared norm measures how tightly the embedded points cluster. The equivalence of these two forms is a consequence of the following standard identity.

Lemma 4.2 (Two forms equivalence)**.** *The pairwise form and the centroid form in* (4) *are equal.*

Geometric interpretation. The equivalence in Lemma 4.2 reveals that $\mathcal{D}_k$ is the *variance* of the point cloud $\{\sqrt{p_v^{(k)}}\}_{v \in V}$ on the unit sphere S_+^{n-1}, with the centroid norm $\|\bar{\phi}^{(k)}\|$ playing the role of a concentration parameter. This norm ranges between two extremes. At one extreme, when all walk distributions coincide, say $p_v^{(k)} = \pi$ for every v, the centroid itself lies on the sphere ($\|\bar{\phi}^{(k)}\| = 1$), so $\mathcal{D}_k = 0$: the point cloud has collapsed to a single point. At the other extreme, at depth $k = 0$, each walk distribution is a Dirac mass $p_v^{(0)} = e_v$, so the embedded points $\sqrt{e_v} = e_v$ are the n standard basis vectors, and the centroid is $\bar{\phi}^{(0)} = \mathbf{1}/n$ with $\|\bar{\phi}^{(0)}\|_2 = 1/\sqrt{n}$, yielding $\mathcal{D}_0 = (n-1)/n$, the maximum possible diversity. Between these two extremes, the curve $k \mapsto \mathcal{D}_k$ traces the trajectory from maximum diversity to total collapse, providing a depth-resolved diagnostic of oversmoothing.

Computation. Each walk distribution $p_v^{(k)} = e_v^\top \mathbf{P}^k$ is obtained by k sparse matrix–vector products at cost $O(m)$, so all n walk distributions cost $O(nmk)$; the centroid and its squared norm are $O(n^2)$, leaving the total cost at $O(nmk)$. For $n \gtrsim 20{,}000$, we sample s source nodes uniformly at random and use the bias-corrected estimator $\hat{\mathcal{D}}_k := \frac{s}{s-1}\big(1 - \|\hat{\bar{\phi}}^{(k)}\|_2^2\big)$, where $\hat{\bar{\phi}}^{(k)} = \frac{1}{s}\sum_{i=1}^{s} \sqrt{p_{v_i}^{(k)}}$. This estimator is unbiased and concentrates exponentially by McDiarmid's inequality [McDiarmid, 1989]: $s = O(\epsilon^{-2}\log(1/\delta))$ samples suffice for $|\hat{\mathcal{D}}_k - \mathcal{D}_k| \le \epsilon$ with probability $\ge 1 - \delta$. See Appendix D for the U-statistic form and proof.

5 Theoretical Analysis

With the aggregation dispersion $\mathcal{D}_k$ defined, we now establish its theoretical properties. The analysis proceeds in three stages. First, we show that $\mathcal{D}_k$ is a faithful and monotone measure of diversity loss (Theorems 5.1 and 5.2). Second, we connect the rate at which $\mathcal{D}_k$ decays to the spectral gap of the propagation matrix (Theorem 5.3). Third, and most importantly, we prove that the graph-level quantity $\mathcal{D}_k$ controls the representation-level diversity of actual MPNNs (Theorems 5.5 and 5.6), thereby justifying its use as a diagnostic for oversmoothing.

5.1 Characterization of Collapse and Monotonicity

A first requirement for $\mathcal{D}_k$ to serve as a meaningful diagnostic of oversmoothing is that it detects *exactly* when collapse occurs. That is, we ask whether $\mathcal{D}_k = 0$ coincides with the situation in which all node representations become indistinguishable. The following result shows that this is indeed the case: vanishing aggregation dispersion is equivalent to complete mixing of the walk distributions.

Theorem 5.1 (Necessary and sufficient condition for collapse)**.** *For a row-stochastic matrix* $\mathbf{P}^k$ *with* $n \ge 2$*:*

$$\mathcal{D}_k = 0 \quad \Longleftrightarrow \quad p_u^{(k)} = p_v^{(k)} \;\; \forall\, u, v \in V \quad \Longleftrightarrow \quad \operatorname{rank}(\mathbf{P}^k) = 1. \tag{5}$$

Consequently, for generic features $\mathbf{X}$ drawn from a continuous distribution, $\mathcal{D}_k = 0$ implies that $\mathbf{P}^k\mathbf{X}$ has identical rows (complete oversmoothing), while $\mathcal{D}_k > 0$ implies that $\mathbf{P}^k\mathbf{X}$ is non-constant across nodes almost surely.

Theorem 5.1 establishes that $\mathcal{D}_k$ is a faithful indicator of collapse: it vanishes if and only if all walk distributions have become identical. This identifies $\mathcal{D}_k$ as a correct endpoint diagnostic. Having characterized *when* collapse occurs, the next question is how diversity evolves with depth: can it recover, or is it irreversibly lost? The following result shows that aggregation dispersion is monotonically non-increasing with depth.

Theorem 5.2 (Monotone decay)**.** *$\mathcal{D}_{k+1} \leq \mathcal{D}_k$ for all $k \geq 0$.*

Theorem 5.2 shows that diversity cannot increase under iterated aggregation: once information from different nodes is mixed, it cannot be unmixed. Oversmoothing is therefore not an abrupt phenomenon but a monotone, irreversible depletion of diversity. This reduces the study of oversmoothing to a rate question: how many layers can be applied before diversity becomes negligible?

5.2 Spectral Control of the Decay Rate

We now show that the decay rate of $\mathcal{D}_k$ is governed by a single spectral quantity: the gap $\gamma = 1 - \lambda_2$ of the lazy random walk. This rate determines the *effective depth* of a graph, the number of layers that can be stacked before representational diversity becomes negligible.

Theorem 5.3 (Decay bound)**.** *Let π be the stationary distribution of the lazy walk, $\pi_{\min} = \min_v \pi_v$, and let $f_1 \equiv 1, f_2, \ldots, f_n$ be right eigenfunctions of P orthonormal in $L^2(\pi)$, with eigenvalues $1 = \lambda_1 > \lambda_2 = 1 - \gamma \geq \cdots \geq \lambda_n \geq 0$. Set $\sigma_2^2 := \frac{1}{n}\sum_v (f_2(v) - \bar{f}_2)^2$ with $\bar{f}_2 = \frac{1}{n}\sum_v f_2(v)$. Then for all $k \geq 0$,*

(i) ***Upper bound:*** *$\mathcal{D}_k \;\leq\; \pi_{\min}^{-1}(1-\gamma)^{2k}$.*

(ii) ***Lower bound:*** *$\mathcal{D}_k \;\geq\; \frac{1}{4}\pi_{\min}\,\sigma_2^2\,(1-\gamma)^{2k}$, where $\sigma_2^2 > 0$ (replaced by $\sum_{i:\lambda_i=\lambda_2} \sigma_i^2$ if λ_2 has multiplicity greater than one).*

Theorem 5.3 identifies the spectral gap as the quantity governing diversity decay: graphs with a large gap (strong expansion) mix rapidly, exhausting $\mathcal{D}_k$ and limiting usable GNN depth, while small-gap graphs retain diversity over many layers. Combined with monotonicity, this fully characterizes the dynamics of $\mathcal{D}_k$: monotone decay at a rate set by the spectral gap. Yet $\mathcal{D}_k$ depends only on walk distributions, and hence only on topology. Whether this topological quantity predicts the behavior of learned GNN representations is the question we address next.

5.3 From Graph Dispersion to Representation Dispersion

The results so far establish that $\mathcal{D}_k$ is a faithful, monotone, and spectrally controlled measure of diversity among walk distributions. These are purely *topological* statements: they depend only on the graph and the propagation operator. However, oversmoothing is ultimately a property of *learned representations*. This raises the central question of this section:

Does the graph-level quantity $\mathcal{D}_k$ control the diversity of node representations produced by MPNNs?

We answer this question in two steps. We first show that for linear MPNNs, representation diversity is directly controlled by $\mathcal{D}_k$. We then extend this connection to nonlinear architectures under standard contractivity assumptions.

Representation dispersion. To formalise this question, we introduce a measure of diversity among node representations that parallels $\mathcal{D}_k$.

Definition 5.4 (Representation dispersion)**.** *For a feature matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{n\times d}$ whose v-th row $x_v^{(k)}$ is the representation of node v at layer k, define*

$$R_k \;=\; \frac{1}{n}\sum_{v=1}^{n} \left\| x_v^{(k)} - \bar{x}^{(k)} \right\|_2^2, \qquad \bar{x}^{(k)} = \frac{1}{n}\sum_v x_v^{(k)}. \tag{6}$$

When $R_k = 0$, all node representations are identical, while $R_k > 0$ indicates that some diversity remains.

The challenge is that $\mathcal{D}_k$ and R_k live in different geometries: $\mathcal{D}_k$ is defined on the probability simplex via Hellinger distance, while R_k is a Euclidean variance in feature space. A Hellinger–ℓ_2 equivalence (Lemma B.5 in Appendix B.5) bridges the two: the mean pairwise squared ℓ_2 distance between walk distributions is sandwiched between $2\mathcal{D}_k^2/n$ and $4\mathcal{D}_k$. Hence $\mathcal{D}_k$ controls the spread of walk distributions in the Euclidean geometry relevant for linear transformations, and we can transfer this control to representation space.

Linear MPNNs. We begin with linear models, where the relationship is explicit.

Theorem 5.5 (Linear MPNN bounds)**.** *For a linear k-layer MPNN with $x_v^{(k)} = p_v^{(k)}\mathbf{X}^{(0)}\mathbf{M}$, let $Z = \mathbf{X}^{(0)}\mathbf{M}$ and let $\sigma_{\min} > 0$ be the smallest singular value of $Z^\top$ on the mean-zero subspace. Then*

$$\frac{2\sigma_{\min}^2}{n}\,\mathcal{D}_k^2 \;\le\; R_k \;\le\; 4\|Z\|_{\mathrm{op}}^2\,\mathcal{D}_k.$$

In particular, if $\sigma_{\min} > 0$, then $\mathcal{D}_k \to 0$ if and only if $R_k \to 0$.

The bounds factorize representation dispersion into a model-dependent prefactor ($\sigma_{\min}$, $\|Z\|_{\mathrm{op}}$) and the aggregation dispersion quantity $\mathcal{D}_k$. The two-sided control gives an iff: $R_k \to 0$ exactly when $\mathcal{D}_k \to 0$, for non-degenerate learned weights. Oversmoothing is forced by the graph topology.

Nonlinear MPNNs. We now extend this connection to nonlinear architectures.

Theorem 5.6 (Nonlinear MPNN)**.** *Consider a k-layer MPNN defined by $\mathbf{X}^{(\ell+1)} = \sigma\big(\mathbf{P}\,\mathbf{X}^{(\ell)}\mathbf{W}^{(\ell)}\big), \qquad \ell = 0, \dots, k-1$, where $\mathbf{P}$ is the lazy walk, σ is applied row-wise and is 1-Lipschitz, and $\|\mathbf{W}^{(\ell)}\|_{\mathrm{op}} \le 1$ for all layers. Then*

$$R_k \;\le\; C\,\mathcal{D}_k,$$

where $C > 0$ depends on the graph's degree heterogeneity and on the initial dispersion.

Together, Theorems 5.5 and 5.6 establish that $\mathcal{D}_k$ governs representation evolution in MPNNs: two-sided control in the linear case, factorizing R_k into a model-dependent prefactor and $\mathcal{D}_k$, and a one-sided upper bound in the nonlinear case. The reverse direction $R_k \ge c'\mathcal{D}_k^p$ fails in the nonlinear case without invertibility assumptions: the nonlinearity can collapse representations before the walk has mixed. In both cases, $\mathcal{D}_k \to 0$ forces $R_k \to 0$, so oversmoothing is driven by the graph's intrinsic mixing. The next section tests this empirically across architectures, datasets, and depths.

6 Experiments

The theoretical analysis of §5 establishes that $\mathcal{D}_k$ is a faithful, monotone, and spectrally controlled measure of walk-distribution diversity. We now test whether these properties translate into predictive power: does $\mathcal{D}_k$ correlate with accuracy degradation better than existing diagnostics, and does this hold across architectures, datasets, and graph types?

6.1 Experimental Setup

We test two questions: *(i)* does $\mathcal{D}_k$, a purely topological diagnostic, correlate with the accuracy degradation of trained GNNs as depth increases? *(ii)* how does it compare with post-training metrics that require access to learned representations? Following Zhang et al. [2026], every depth is a separately trained model rather than a layer of a single deep network.

Datasets. We evaluate on 11 node-classification benchmarks spanning homophily $h \in [0.05, 0.81]$ and scale $n \in [183, 169{,}343]$: the homophilic citation networks Cora, CiteSeer, PubMed [Yang et al., 2016]; the heterophilic graphs Chameleon, Squirrel, Wisconsin, Texas, Cornell [Pei et al., 2020]; and three large-scale benchmarks Roman-Empire, Amazon-Ratings [Platonov et al., 2023] and ogbn-arXiv [Hu et al., 2020]. Per-dataset statistics are in Appendix C.

Models. We evaluate seven architectures spanning message-passing, attentional, spectral, and decoupled designs: GCN [Kipf and Welling, 2017], GATv1 [Veličković et al., 2018], GATv2 [Brody et al., 2022], GraphSAGE [Hamilton et al., 2017], ChebNet [Defferrard et al., 2016], JKNet [Xu et al., 2018], and APPNP [Gasteiger et al., 2019], each with LeakyReLU and Tanh activations (14

configurations). For each configuration we sweep depths $L \in \{2, 4, 6, 8, 10, 12, 16, 24\}$ over 10 seeds, yielding 1,120 trained models per dataset. Oversmoothing-mitigated GCN variants are deferred to Appendix F.3, while GIN experiments are deferred to Appendix E.2 and Appendix E.3.

Training. Following Zhang et al. [2026]: hidden dim 32 (64 for the three large graphs), dropout 0.5, Adam (lr=10^{-2}, wd 5×10^{-4}), 200 epochs with early stopping (patience 10).

Baselines and evaluation. We compare $\mathcal{D}_k$ against seven post-training metrics computed on final-layer representations: Dirichlet energy E_{Dir} [Cai and Wang, 2020] and its normalized variant $E_{\mathrm{Dir,n}}$; projected energy E_{Proj} and $E_{\mathrm{Proj,n}}$ [Wu et al., 2023]; MAD [Chen et al., 2020a]; and the effective and numerical ranks erank, nrank [Zhang et al., 2026]. Definitions are in Appendix C. For each (dataset, model) pair we report the Pearson correlation $|r|$ between $\log(\mathrm{metric}_k)$ and test accuracy across the eight depths, with 95% bootstrap confidence intervals (1,000 resamples).

6.2 Correlation with Accuracy Degradation

Table 1 reports the mean absolute Pearson correlation $|r|$ between $\log(\mathrm{metric})$ and test accuracy across eight depths, averaged over 14 model configurations. The log transform linearizes the approximately exponential decay of each diagnostic with depth; Spearman and Kendall correlations yield the same ranking (Appendix F.4).[1] Per-cell standard deviation is taken across configurations; subtotal and **Overall** standard errors are taken across datasets within a regime.

Table 1: **Mean absolute correlation $|r|$ between $\log$(metric) and test accuracy**, averaged over model configurations. $\mathcal{D}_k$ leads **Overall** (0.721 ± 0.035) and on heterophilic graphs (0.698 ± 0.051); NumRank leads on homophilic ones (0.766 ± 0.014). Per-cell std is across configurations; subtotal and **Overall** std is the standard error across datasets. **Bold** = best per row, underline = second best.

Dataset	$\mathcal{D}_k$ (ours)	Erank	$E_{\mathrm{Proj,n}}$	NumRank	$E_{\mathrm{Dir,n}}$	E_{Dir}	MAD	E_{Proj}
HOMOPHILIC								
CiteSeer	**0.842**±0.18	0.819±0.13	0.691±0.23	0.772±0.21	0.696±0.26	0.697±0.28	0.619±0.21	0.641±0.27
Cora	0.744±0.23	**0.756**±0.24	0.704±0.17	0.729±0.24	0.636±0.21	0.730±0.16	0.538±0.21	0.641±0.22
PubMed	0.681±0.23	0.627±0.33	0.626±0.26	**0.768**±0.22	0.708±0.24	0.513±0.36	0.703±0.23	0.575±0.25
ogbn-arxiv	0.781±0.12	**0.812**±0.20	0.717±0.29	0.794±0.22	0.660±0.25	0.544±0.31	0.719±0.24	0.598±0.25
Homophilic avg.	0.762±0.034	0.754±0.044	0.685±0.020	**0.766**±0.014	0.675±0.017	0.621±0.054	0.645±0.042	0.614±0.016
HETEROPHILIC								
AmazonRatings	**0.743**±0.20	0.701±0.25	0.677±0.31	0.652±0.32	0.587±0.29	0.667±0.34	0.594±0.34	0.540±0.30
Chameleon	**0.878**±0.13	0.758±0.25	0.708±0.24	0.532±0.33	0.624±0.29	0.655±0.25	0.774±0.14	0.661±0.25
Cornell	0.542±0.23	0.524±0.29	0.519±0.33	**0.549**±0.27	0.527±0.33	0.494±0.38	0.473±0.25	0.510±0.35
RomanEmpire	0.829±0.18	0.751±0.19	0.713±0.20	0.680±0.22	**0.893**±0.06	0.652±0.19	0.782±0.13	0.592±0.21
Squirrel	**0.723**±0.27	0.609±0.27	0.710±0.27	0.440±0.28	0.622±0.29	0.614±0.26	0.520±0.36	0.654±0.26
Texas	0.521±0.19	0.501±0.28	0.592±0.27	0.527±0.27	0.514±0.31	**0.596**±0.30	0.488±0.28	0.594±0.29
Wisconsin	0.649±0.25	0.706±0.29	0.754±0.28	0.676±0.24	0.650±0.32	**0.757**±0.29	0.577±0.29	0.754±0.30
Heterophilic avg.	**0.698**±0.051	0.650±0.040	0.668±0.031	0.579±0.035	0.631±0.048	0.634±0.030	0.601±0.049	0.615±0.031
Overall	**0.721**±0.035	0.688±0.043	0.674±0.041	0.647±0.042	0.647±0.044	0.629±0.045	0.617±0.042	0.614±0.043

$\mathcal{D}_k$ attains the highest overall correlation (0.721 ± 0.035) with the smallest cross-dataset standard error, and ranks in the top two on 7 of 11 benchmarks. The ranking then separates by homophily. On homophilic datasets, nrank leads (0.766 ± 0.014) and $\mathcal{D}_k$ follows (0.762 ± 0.034), a gap of 0.004 that is negligible relative to the reported uncertainty. On heterophilic datasets, $\mathcal{D}_k$ leads (0.698 ± 0.051), exceeding the next metric by 0.119, with the largest margins on Chameleon (0.878), AmazonRatings (0.743), and Squirrel (0.723). This asymmetry is consistent with the mechanism of oversmoothing: under heterophily, neighbors frequently belong to different classes, so aggregation mixes incompatible signals and degrades representation-based diagnostics, whereas topology-driven measures remain sensitive to the underlying mixing structure. The ranking is otherwise stable across graph sizes ($n = 183$ to $n = 169{,}343$) and homophily levels ($h = 0.05$ to $h = 0.81$).

Two caveats apply. First, on the smallest benchmarks (Texas; $n \leq 251$), no metric exceeds $|r| = 0.76$ and 95% bootstrap intervals over configurations overlap, so differences in this regime are not

[1] We report $|r|$ because the sign of the metric–accuracy relationship differs across diagnostics by construction; $|r|$ therefore measures strength of association only.

distinguishable. Second, one might ask why not use a purely topological baseline such as the spectral-gap power λ^k, where λ is the spectral gap of the symmetrized Laplacian. This reaches $|r| = 0.702$ but is unsatisfactory: $\log(\lambda^k) = k \log \lambda$ is linear in depth, so its correlation with accuracy reduces to the trivial depth–accuracy trend; it compresses mixing into a single scalar, while $\mathcal{D}_k$ tracks the per-layer dispersion of aggregated distributions (Appendix F.6); and $\mathcal{D}_k$ can be architecture-aware, defined relative to the propagation operator the MPNN actually uses (Appendix E).

6.3 Per-Architecture Breakdown

$\mathcal{D}_k$ is defined from the lazy walk and references no architecture-specific propagation operator. We test whether this architecture-agnostic construction still predicts accuracy degradation across seven GNNs grouped into four families: spectral (GCN, ChebNet), attention (GATv1, GATv2), sampling (GraphSAGE), and skip/diffusion (JKNet, APPNP).

Table 2: **Mean absolute correlation** $|r|$ **between** $\log($**metric**$)$ **and test accuracy**, averaged over datasets, per architecture. **Bold** = best per row, underline = second best.

Family	Architecture	$\mathcal{D}_k$ (ours)	Erank	NumRank	E_{Dir}	$E_{\mathrm{Dir,n}}$	E_{Proj}	$E_{\mathrm{Proj,n}}$	MAD
Spectral	GCN	**0.793**	0.705	0.671	0.553	0.719	0.496	0.544	0.584
	ChebNet	**0.806**	0.785	0.740	0.691	0.473	0.647	0.647	0.565
Attention	GATv1	**0.785**	0.688	0.583	0.633	0.715	0.609	0.699	0.622
	GATv2	**0.798**	0.764	0.669	0.659	0.777	0.681	0.784	0.670
Sampling	SAGE	0.812	**0.884**	0.704	0.739	0.752	0.734	0.810	0.761
Skip/Diffusion	JKNet	0.503	0.489	**0.566**	0.474	0.560	0.488	0.523	0.530
	APPNP	0.552	0.499	0.597	0.653	0.531	0.647	**0.708**	0.585
Overall		**0.721**	0.688	0.647	0.629	0.647	0.615	0.674	0.617

$\mathcal{D}_k$ is the best diagnostic on all four spectral and attention architectures (GCN 0.793, ChebNet 0.806, GATv1 0.785, GATv2 0.798; family means 0.800 and 0.792). These are the architectures whose effective propagation is closest to a reweighting of the lazy walk, symmetric normalization in the spectral case, learned row-stochastic attention in the other, so $\mathcal{D}_k$'s predictive accuracy tracks how closely a model's propagation perturbs the lazy-walk operator.

Predictive accuracy weakens as effective propagation departs from the lazy-walk operator. On GraphSAGE, mini-batch neighbor sampling replaces full-neighborhood aggregation with a random subsample, and Erank (0.884) overtakes $\mathcal{D}_k$ (0.812) by reading embeddings that no topological measure can recover. On JKNet and APPNP, skip connections and personalized-PageRank propagation are used to mitigate oversmoothing, and Zhang et al. [2026] observe that all oversmoothing diagnostics become uniformly less informative once such components are in play. The effect is visible here, the best metric on JKNet reaches only 0.566.

7 Conclusion

We introduced **aggregation dispersion** $\mathcal{D}_k$, a training-free diagnostic for oversmoothing in message-passing neural networks. Viewing each row of $\mathbf{P}^k$ as a probability distribution over nodes, $\mathcal{D}_k$ measures the contraction of the resulting point cloud on the simplex using the Fisher–Rao metric, with the Hellinger distance as its numerically stable chordal form. The diagnostic is computable in $O(nmk)$ time from the propagation matrix alone. We proved that $\mathcal{D}_k = 0$ if and only if $\mathbf{P}^k$ has rank one, that $\mathcal{D}_k$ decays monotonically with depth via the data-processing inequality, that the spectral gap controls the decay rate, and that $\mathcal{D}_k$ upper-bounds representation dispersion for contractive MPNNs. Across 14 model configurations and 11 datasets, $\mathcal{D}_k$ attained the highest correlation with accuracy degradation ($|r| = 0.721$), with the largest margin on heterophilic graphs. A natural next step is to move from graph-level to node-level diagnostics by analyzing the spectrum of the pairwise Hellinger Gram matrix, identifying which nodes lose distinctiveness first. More practically, $\mathcal{D}_k$ could serve as a signal for depth selection, choosing the largest k before dispersion falls below a task-dependent threshold without requiring a training pass at each candidate depth.

References

S.-i. Amari. *Information Geometry and Its Applications*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 4431559779.

S.-i. Amari and S. Douglas. Why natural gradient? In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2, pages 1213 – 1216 vol.2, 06 1998. doi: 10.1109/ICASSP.1998.675489.

G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJzRZ-WCZ.

S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

C. Cai and Y. Wang. A note on over-smoothing for graph neural networks, 2020. URL https://arxiv.org/abs/2006.13318.

N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982. Translated from Russian by Lev J. Leifman.

D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *AAAI*, 2020a.

M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. Simple and deep graph convolutional networks. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1725–1735. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/chen20v.html.

F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

T. M. Cover and J. A. Thomas. Inequalities in information theory. In *Elements of Information Theory*, chapter 17, pages 657–687. John Wiley & Sons, 2nd edition, 2005. doi: 10.1002/047174882X.ch17.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

A. Ed-dib, Z. Datbayev, and A. M. Aboussalah. GeLoRA: Geometric adaptive ranks for efficient LoRA fine-tuning. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25174–25196, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1372. URL https://aclanthology.org/2025.findings-emnlp.1372/.

J. Gasteiger, A. Bojchevski, and S. Günnemann. Combining neural networks with personalized pagerank for classification on graphs. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org, 2017.

F. D. Giovanni, J. Rowbottom, B. P. Chamberlain, T. Markovich, and M. M. Bronstein. Understanding convolution on graphs via energies. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=v5ew3FPTgb.

W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: datasets for machine learning on graphs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

N. Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

J. Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, pages 148–188. Cambridge University Press, Cambridge, 1989.

K. Oono and T. Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang. Geom-gcn: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

O. Platonov, D. Kuznedelev, M. Diskin, A. Babenko, and L. Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tJbbQfw-5wv.

Y. Rong, W. Huang, T. Xu, and J. Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkx1qkrKPr.

O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, 2007.

T. K. Rusch, M. M. Bronstein, and S. Mishra. A survey on oversmoothing in graph neural networks, 2023. URL https://arxiv.org/abs/2303.10993.

K. Sun and F. Nielsen. A geometric modeling of occam's razor in deep learning. *Information Geometry*, 8(1):233–273, 2025. doi: 10.1007/s41884-025-00167-2. URL https://doi.org/10.1007/s41884-025-00167-2.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

X. Wu, A. Ajorlou, Z. Wu, and A. Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka. Representation learning on graphs with jumping knowledge networks. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5453–5462. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xu18c.html.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Z. Yang, W. W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 40–48. JMLR.org, 2016.

K. Zhang, P. Deidda, D. Higham, and F. Tudisco. Are we measuring oversmoothing in graph neural networks correctly? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r3DOISnvHD.

L. Zhao and L. Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkecl1rtwB.

A Notation and preliminaries

This section consolidates the notation used throughout the paper and collects the basic properties of the Hellinger distance invoked in our proofs. It is intended as a reference; readers familiar with the setup of §3 may skip directly to Appendix B.

A.1 Notation

We collect the symbols used throughout the paper in four short tables organised by category: graphs and operators (Table 3), the simplex and its embedding (Table 4), distances (Table 5), and diagnostics (Table 6). MPNN architectural notation ($\mathbf{X}^{(\ell)}$ for layer-ℓ features, $\mathbf{W}^{(\ell)}$ for weights, σ for an elementwise 1-Lipschitz activation, and $\|\cdot\|_{\mathrm{op}}$ for the spectral norm) follows standard conventions and is not tabulated.

Table 3: Graphs and propagation operators.

$G=(V,E)$, $n=\lvert V\rvert$, $m=\lvert E\rvert$	connected undirected graph
$\mathbf{A}$, $\mathbf{D}$, $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{I}$, $\tilde{\mathbf{D}}=\mathbf{D}+\mathbf{I}$	adjacency, degree, self-loop-augmented
$d(u,v)$, $\mathcal{D}=\mathrm{diam}(G)$	shortest-path distance, diameter
$\mathbf{P}=\frac{1}{2}(\mathbf{I}+\mathbf{D}^{-1}\mathbf{A})$	lazy random walk
$1=\lambda_1>\lambda_2\geq\cdots\geq\lambda_n>0$	eigenvalues of $\mathbf{P}$
$\gamma=1-\lambda_2$	spectral gap
$\boldsymbol{\pi}$, $\pi_v=\frac{d_v}{2m}$, $\pi_{\min}=\min_v \pi_v$	stationary distribution

Table 4: Walk distributions and the square-root embedding.

$\Delta^{n-1}=\{p\in\mathbb{R}^n_{\geq 0}:\sum_i p_i=1\}$	probability simplex
$p_v^{(k)}=[\mathbf{P}^k]_{v,:}$	walk distribution from v at depth k
$S_+^{n-1}=\{x\in\mathbb{R}^n_{\geq 0}:\|x\|_2=1\}$	positive orthant of unit sphere
$\phi: p\mapsto\sqrt{p}$, $\quad\phi_v^{(k)}=\sqrt{p_v^{(k)}}$	square-root embedding
$\bar{\phi}^{(k)}=\frac{1}{n}\sum_v \phi_v^{(k)}$	embedded centroid

Table 5: Distances on the probability simplex.

$H(p,q)=\frac{1}{\sqrt{2}}\|\sqrt{p}-\sqrt{q}\|_2$	Hellinger, $H\in[0,1]$
$\mathrm{TV}(p,q)=\frac{1}{2}\|p-q\|_1$	total variation
$d_{\mathrm{FR}}(p,q)=2\arccos\langle\sqrt{p},\sqrt{q}\rangle$	Fisher–Rao geodesic

Table 6: Diagnostics.

$\mathcal{D}_k=\frac{1}{n^2}\sum_{u,v}H^2(p_u^{(k)},p_v^{(k)})=1-\|\bar{\phi}^{(k)}\|_2^2$	aggregation dispersion (ours)
$\widehat{\mathcal{D}}_k$	sampled estimator
$\mathcal{R}_k=\frac{1}{n}\sum_v\|x_v^{(k)}-\bar{x}^{(k)}\|_2^2$	representation dispersion (uniform)
$\mathcal{R}_k^{\boldsymbol{\pi}}=\sum_v\pi_v\|x_v^{(k)}-\bar{x}^{\boldsymbol{\pi}}\|_2^2$	representation dispersion ($\boldsymbol{\pi}$-weighted)

A.2 The lazy random walk

We adopt the lazy random walk $\mathbf{P}=\frac{1}{2}(\mathbf{I}+\mathbf{D}^{-1}\mathbf{A})$ as the canonical propagation operator throughout our analysis. Three properties make it the natural baseline for studying oversmoothing as a structural phenomenon, independent of architectural choices.

Row-stochasticity. Each row of $\mathbf{P}$ sums to one, so each row of $\mathbf{P}^k$ is a probability distribution on V. This is what allows us to interpret the v-th row as a walk distribution $p_v^{(k)} \in \Delta^{n-1}$.

Reversibility. The lazy walk satisfies the detailed-balance condition $\pi_u P_{uv} = \pi_v P_{vu}$ with respect to $\boldsymbol{\pi}$. Reversibility implies that $\mathbf{P}$ is self-adjoint on $L^2(\boldsymbol{\pi})$, so its eigenvalues are real and the symmetrised operator $\mathbf{S} = \tilde{\mathbf{D}}^{1/2}\mathbf{P}\tilde{\mathbf{D}}^{-1/2}$ has the same spectrum as $\mathbf{P}$ but with orthonormal eigenvectors. We exploit this in the proof of Theorem 5.3(ii).

Aperiodicity. The self-loop term $\frac{1}{2}\mathbf{I}$ ensures that $\mathbf{P}$ has period one, so the induced Markov chain converges to $\boldsymbol{\pi}$ from every initial distribution. Without this lazy correction, bipartite graphs would exhibit oscillatory behavior and Theorem 5.3 would fail at odd depths.

The eigenvalue ordering $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > 0$ is a consequence of the three properties combined: row-stochasticity gives $\lambda_1 = 1$ with multiplicity one (Perron–Frobenius, using connectedness), reversibility gives real eigenvalues, and aperiodicity gives $\lambda_n > -1$, which combined with the lazy correction yields[2] $\lambda_n > 0$. Appendix E extends our analysis to the standard architectural propagation operators (GCN, GIN, GraphSAGE) by relating their spectra to that of the lazy walk.

A.3 The square-root embedding and the Hellinger distance

The square-root embedding $\phi : \Delta^{n-1} \to S_+^{n-1}$, $p \mapsto (\sqrt{p_1}, \ldots, \sqrt{p_n})$, sends each distribution to a point on the positive orthant of the unit sphere, since $\|\sqrt{p}\|_2^2 = \sum_i p_i = 1$. The Hellinger distance is the chord length between embedded points:

$$H(p, q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2, \qquad H \in [0, 1]. \tag{7}$$

The factor $1/\sqrt{2}$ normalizes so that $H = 1$ when p and q have disjoint supports (then $\langle\sqrt{p}, \sqrt{q}\rangle = 0$ and $\|\sqrt{p} - \sqrt{q}\|_2 = \sqrt{2}$). Three properties of H are used repeatedly in our proofs.

Lemma A.1 (Data-processing inequality)**.** *For any row-stochastic matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ and any two distributions $p, q \in \Delta^{n-1}$,*

$$H(p\mathbf{K}, q\mathbf{K}) \leq H(p, q). \tag{8}$$

The DPI is the form of the Hellinger distance's affinity-monotonicity under stochastic post-processing; see Cover and Thomas [2005] for a textbook treatment. We apply it with $\mathbf{K} = \mathbf{P}$ in the proof of Theorem 5.2 to establish that $\mathcal{D}_k$ is non-increasing in k.

Lemma A.2 (Finiteness on disjoint supports)**.** *For any $p, q \in \Delta^{n-1}$, including those with $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \emptyset$, the Hellinger distance is bounded: $H(p, q) \leq 1$.*

This property distinguishes Hellinger from Kullback–Leibler and χ^2 divergences, both of which diverge to infinity on disjoint supports. Disjoint supports occur in our setting at every early depth: for two nodes u, v with $d(u, v) > 2k$, the walk distributions $p_u^{(k)}$ and $p_v^{(k)}$ have disjoint supports because $\mathbf{P}^k$ has the locality property $[\mathbf{P}^k]_{vu} \neq 0 \Rightarrow d(v, u) \leq k$. A divergence that blew up at $k = 0$ could not produce a meaningful trajectory $k \mapsto \mathcal{D}_k$.

Lemma A.3 (Hellinger–total-variation equivalence)**.** *For any $p, q \in \Delta^{n-1}$,*

$$H^2(p, q) \leq \mathrm{TV}(p, q) \leq \sqrt{2}\, H(p, q). \tag{9}$$

The two inequalities sandwich Hellinger and total variation, so any bound stated in one metric transfers to the other up to constants. We use the left inequality in the proof of Theorem 5.3(i) to import the standard spectral mixing bound $\mathrm{TV}(p_v^{(k)}, \boldsymbol{\pi}) \leq \frac{1}{2}\pi_v^{-1/2}(1 - \gamma)^k$ from Markov-chain theory [Levin et al., 2017] into a Hellinger statement.

A.4 The probability simplex as a Riemannian manifold

The interior of the simplex carries a canonical Riemannian metric, the Fisher–Rao metric g_{FR}, characterised by Čencov's theorem [Čencov, 1982] as the unique (up to scale) Riemannian metric on

[2] We assume that the graph is non-bipartite.

Δ^{n-1} that is invariant under sufficient statistics. Under the square-root embedding ϕ, the simplex $(\Delta^{n-1}, \frac{1}{4}g_{\mathrm{FR}})$ maps isometrically onto the positive orthant of the unit sphere $(S_+^{n-1}, g_{\mathrm{round}})$ equipped with the round metric. Distances on the simplex therefore correspond to spherical distances on S_+^{n-1}:

- The *Fisher–Rao geodesic distance* $d_{\mathrm{FR}}(p,q) = 2\arccos\langle\sqrt{p}, \sqrt{q}\rangle$ is the arc length on S_+^{n-1} between $\phi(p)$ and $\phi(q)$.
- The *Hellinger distance* $H(p,q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2$ is the chord length between the same two points.

For points on a unit sphere, chord length and arc length are related by $\|x - y\|_2 = 2\sin(\theta/2)$ where θ is the arc-length angle, so H and d_{FR} agree to first order near the diagonal $\{p = q\}$ and diverge only when distributions are far apart on the simplex. We work with H rather than d_{FR} throughout because it admits the closed-form centroid identity

$$\frac{1}{n^2}\sum_{u,v} H^2(p_u, p_v) = 1 - \|\bar{\phi}\|_2^2, \tag{10}$$

which reduces the computation of $\mathcal{D}_k$ from $O(n^3)$ to $O(n^2)$ and is the cornerstone of our estimator (Appendix D). The arccosine of an inner product admits no analogous reduction.

A.5 Two notions of representation dispersion

Our representation-level results state contraction bounds in two related but distinct quantities. The *uniform-mean representation dispersion*

$$\mathcal{R}_k = \frac{1}{n}\sum_{v=1}^{n} \|x_v^{(k)} - \bar{x}^{(k)}\|_2^2, \qquad \bar{x}^{(k)} = \frac{1}{n}\sum_v x_v^{(k)}, \tag{11}$$

is the natural Euclidean variance of the node representations and is what one would compute in practice on a held-out graph. The $\boldsymbol{\pi}$*-weighted representation dispersion*

$$\mathcal{R}_k^{\boldsymbol{\pi}} = \sum_{v=1}^{n} \pi_v \|x_v^{(k)} - \bar{x}^{\boldsymbol{\pi}}\|_2^2, \qquad \bar{x}^{\boldsymbol{\pi}} = \sum_v \pi_v x_v^{(k)}, \tag{12}$$

is the variance under the stationary measure $\boldsymbol{\pi}$ and is the quantity that contracts cleanly under the lazy walk, because $\boldsymbol{\pi}^\top \mathbf{P} = \boldsymbol{\pi}^\top$ and the operator norm of $\mathbf{P}$ on the $\boldsymbol{\pi}$-mean-zero subspace of $L^2(\boldsymbol{\pi})$ is exactly $\lambda_2 = 1 - \gamma$.

The two dispersions are related by the following inequalities.

$$\mathcal{R}_k \;\le\; \frac{d_{\mathrm{avg}}}{d_{\min}}\,\mathcal{R}_k^{\boldsymbol{\pi}}, \qquad \mathcal{R}_k^{\boldsymbol{\pi}} \;\le\; \frac{d_{\max}}{d_{\mathrm{avg}}}\,\mathcal{R}_k, \tag{13}$$

where $d_{\mathrm{avg}} = \frac{1}{n}\sum_v d_v$, $d_{\min} = \min_v d_v$, $d_{\max} = \max_v d_v$. On regular graphs both constants equal 1 and the two notions coincide. In our main result (Theorem 5.6) we state contraction in $\mathcal{R}_k^{\boldsymbol{\pi}}$ and transfer to $\mathcal{R}_k$ using the first bound.

Proof. For a connected non-bipartite undirected graph, the stationary distribution of the lazy walk is $\pi_v = d_v/(2|E|)$. Note that $d_{\mathrm{avg}} = 2|E|/n$.

First we prove $\mathcal{R}_k \le \frac{d_{\mathrm{avg}}}{d_{\min}}\mathcal{R}_k^{\boldsymbol{\pi}}$. Because the uniform mean $\bar{x}$ minimizes the uniform variance, we have

$$\mathcal{R}_k = \frac{1}{n}\sum_v \|x_v - \bar{x}\|^2 \;\le\; \frac{1}{n}\sum_v \|x_v - \bar{x}^{\boldsymbol{\pi}}\|^2.$$

Rewriting the right-hand side using the weights π_v gives

$$\frac{1}{n}\sum_v \|x_v - \bar{x}^{\boldsymbol{\pi}}\|^2 = \sum_v \pi_v \cdot \frac{1}{n\pi_v}\|x_v - \bar{x}^{\boldsymbol{\pi}}\|^2 \;\le\; \left(\max_v \frac{1}{n\pi_v}\right)\sum_v \pi_v \|x_v - \bar{x}^{\boldsymbol{\pi}}\|^2.$$

Since $\pi_v = d_v/(2|E|)$, we have $\frac{1}{n\pi_v} = \frac{2|E|}{nd_v} = \frac{d_{\text{avg}}}{d_v}$. Thus $\max_v \frac{1}{n\pi_v} = \frac{d_{\text{avg}}}{d_{\min}}$ and the first inequality follows.

For the reverse inequality, start from the definition of $\mathcal{R}_k^{\boldsymbol{\pi}}$ and use that $\bar{x}^{\boldsymbol{\pi}}$ minimizes the weighted variance:

$$\mathcal{R}_k^{\boldsymbol{\pi}} = \sum_v \pi_v \|x_v - \bar{x}^{\boldsymbol{\pi}}\|^2 \;\le\; \sum_v \pi_v \|x_v - \bar{x}\|^2.$$

Now write $\pi_v = \frac{d_v}{2|E|} = \frac{d_v}{nd_{\text{avg}}}$. Then

$$\sum_v \pi_v \|x_v - \bar{x}\|^2 = \frac{1}{nd_{\text{avg}}} \sum_v d_v \|x_v - \bar{x}\|^2 \;\le\; \frac{d_{\max}}{d_{\text{avg}}} \cdot \frac{1}{n} \sum_v \|x_v - \bar{x}\|^2 = \frac{d_{\max}}{d_{\text{avg}}} \mathcal{R}_k.$$

This completes the proof. □

B Proofs of main results

This appendix collects the proofs of the main-paper theorems and the equivalence lemma. The order follows the main text: §B.1 establishes the two-form equivalence underlying Definition 4.1; §B.2 and §B.3 prove the structural properties (Theorems 5.1 and 5.2); §B.4 proves the spectral decay bound (Theorem 5.3, both parts); §B.5 establishes a Hellinger–ℓ_2 equivalence used in the representation-level results; and §B.6 and §B.7 prove the representation-dispersion bounds (Theorems 5.5 and 5.6).

B.1 Lemma 4.2: two-form equivalence

Lemma B.1 (Two-form equivalence)**.** *The pairwise form and the centroid form in* (4) *are equal.*

Proof. Since every walk distribution sums to one, each embedded point has unit norm: $\|\sqrt{p_v^{(k)}}\|_2^2 = \sum_i p_{v,i}^{(k)} = 1$. The variance of these unit-norm points about their centroid is therefore

$$\begin{aligned}\frac{1}{n}\sum_v \big\|\sqrt{p_v^{(k)}} - \bar{\phi}^{(k)}\big\|_2^2 &= \frac{1}{n}\sum_v \Big(\big\|\sqrt{p_v^{(k)}}\big\|^2 - 2\langle \sqrt{p_v^{(k)}}, \bar{\phi}^{(k)}\rangle + \big\|\bar{\phi}^{(k)}\big\|^2\Big) \\ &= 1 - 2\big\|\bar{\phi}^{(k)}\big\|^2 + \big\|\bar{\phi}^{(k)}\big\|^2 = 1 - \big\|\bar{\phi}^{(k)}\big\|^2.\end{aligned}$$

The standard variance–covariance identity relates this variance to the mean pairwise squared distance:

$$\frac{1}{n}\sum_v \big\|\sqrt{p_v^{(k)}} - \bar{\phi}^{(k)}\big\|^2 = \frac{1}{2n^2}\sum_{u,v} \big\|\sqrt{p_u^{(k)}} - \sqrt{p_v^{(k)}}\big\|^2 = \frac{1}{n^2}\sum_{u,v} H^2(p_u^{(k)}, p_v^{(k)}),$$

where the last equality uses the definition of the Hellinger distance, $H^2(p,q) = \frac{1}{2}\|\sqrt{p} - \sqrt{q}\|_2^2$. Combining the two displays yields the claimed identity. □

B.2 Theorem 5.1: necessary and sufficient condition for collapse

Theorem B.2 (Necessary and sufficient condition for collapse)**.** *For a row-stochastic matrix $\mathbf{P}^k$ with $n \ge 2$:*

$$\mathcal{D}_k = 0 \quad\Longleftrightarrow\quad p_u^{(k)} = p_v^{(k)} \;\; \forall\, u, v \in V \quad\Longleftrightarrow\quad \operatorname{rank}(\mathbf{P}^k) = 1. \tag{14}$$

Consequently, for generic features $\mathbf{X}$ drawn from a continuous distribution, $\mathcal{D}_k = 0$ implies that $\mathbf{P}^k\mathbf{X}$ has identical rows (complete oversmoothing), while $\mathcal{D}_k > 0$ implies that $\mathbf{P}^k\mathbf{X}$ is non-constant across nodes almost surely; equivalently, at least one pair of rows remains distinct.

Proof. (⇐). Suppose $p_v^{(k)} = q$ for all v. Then $\phi_v^{(k)} = \sqrt{q}$ for all v, the centroid is $\bar{\phi}^{(k)} = \sqrt{q}$, and $\|\bar{\phi}^{(k)}\|_2^2 = \|\sqrt{q}\|_2^2 = 1$, so $\mathcal{D}_k = 1 - 1 = 0$.

(⇒). Suppose $\mathcal{D}_k = 0$. By Lemma 4.2 in its variance form, $\frac{1}{n}\sum_v \|\phi_v^{(k)} - \bar{\phi}^{(k)}\|_2^2 = 0$. Each summand is non-negative, so every summand vanishes: $\phi_v^{(k)} = \bar{\phi}^{(k)}$ for all v. The coordinatewise square root $\mathbb{R}^n_{\ge 0} \to \mathbb{R}^n_{\ge 0}$ is injective, so $p_u^{(k)} = p_v^{(k)}$ for all $u, v \in V$.

A row-stochastic matrix with all rows equal to a common distribution q has the form $\mathbf{1}q^\top$, which has rank one. Conversely, a row-stochastic rank-one matrix must have the form $\mathbf{1}q^\top$ for some $q \in \Delta^{n-1}$ (the unique row, repeated), so its rows are all equal.

Consequently clause. When $\operatorname{rank}(\mathbf{P}^k) = 1$, the matrix has the form $\mathbf{1}q^\top$, so $\mathbf{P}^k\mathbf{X} = \mathbf{1}(q^\top\mathbf{X})$ is the same row repeated regardless of $\mathbf{X}$. When $\operatorname{rank}(\mathbf{P}^k) \geq 2$, there exist two distinct rows $a_u \neq a_v$ of $\mathbf{P}^k$. For features $\mathbf{X}$ drawn from a continuous distribution, the event $\{(a_u - a_v)\mathbf{X} = 0\}$ is a proper linear subspace of the feature space and has Lebesgue measure zero. Hence $(\mathbf{P}^k\mathbf{X})_u \neq (\mathbf{P}^k\mathbf{X})_v$ almost surely. □

B.3 Theorem 5.2: monotone decay

Theorem B.3 (Monotone decay)**.** *$\mathcal{D}_{k+1} \leq \mathcal{D}_k$ for all $k \geq 0$.*

Proof. The Hellinger distance satisfies the data-processing inequality: for any row-stochastic matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ (a Markov kernel) and any two distributions $p, q \in \Delta^{n-1}$,

$$H(p\mathbf{K}, q\mathbf{K}) \leq H(p, q), \tag{15}$$

see Lemma A.1 in Appendix A.3. The walk distributions evolve under one additional application of the lazy walk via $p_v^{(k+1)} = p_v^{(k)}\mathbf{P}$. Applying (15) with $\mathbf{K} = \mathbf{P}$ to each pair (u, v):

$$H^2(p_u^{(k+1)}, p_v^{(k+1)}) = H^2(p_u^{(k)}\mathbf{P}, p_v^{(k)}\mathbf{P}) \leq H^2(p_u^{(k)}, p_v^{(k)}).$$

Summing over all n^2 pairs and dividing by n^2:

$$\mathcal{D}_{k+1} = \frac{1}{n^2}\sum_{u,v} H^2(p_u^{(k+1)}, p_v^{(k+1)}) \leq \frac{1}{n^2}\sum_{u,v} H^2(p_u^{(k)}, p_v^{(k)}) = \mathcal{D}_k. \qquad \square$$

B.4 Theorem 5.3: spectral decay bound

Theorem B.4 (Decay bound)**.** *Let π be the stationary distribution of the lazy walk, $\pi_{\min} = \min_v \pi_v$, and let $f_1 \equiv 1, f_2, \ldots, f_n$ be right eigenfunctions of P orthonormal in $L^2(\pi)$, with eigenvalues $1 = \lambda_1 > \lambda_2 = 1-\gamma \geq \cdots \geq \lambda_n \geq 0$. Set $\sigma_2^2 := \frac{1}{n}\sum_v (f_2(v) - \bar{f}_2)^2$ with $\bar{f}_2 = \frac{1}{n}\sum_v f_2(v)$. Then for all $k \geq 0$,*

(i) ***Upper bound:*** *$\mathcal{D}_k \;\leq\; \pi_{\min}^{-1}(1-\gamma)^{2k}$.*

(ii) ***Lower bound:*** *$\mathcal{D}_k \;\geq\; \frac{1}{4}\pi_{\min}\,\sigma_2^2\,(1-\gamma)^{2k}$, where $\sigma_2^2 > 0$ (replaced by $\sum_{i:\lambda_i=\lambda_2}\sigma_i^2$ if λ_2 has multiplicity greater than one).*

Proof. We establish the two bounds one after the other.

Part (i) – Upper bound.

Let $p_v^{(k)} = P^k(v, \cdot)$ and define the density $f_k := p_v^{(k)}/\pi$. For the reversible lazy walk, the $L^2(\pi)$ contraction (see Levin et al. [2017, Section 12.5]) gives

$$\|f_k - 1\|_{2,\pi} \leq (1-\gamma)^k \|f_0 - 1\|_{2,\pi}.$$

Squaring yields the familiar χ^2-divergence bound

$$\chi^2\Big(p_v^{(k)} \,\|\, \pi\Big) = \|f_k - 1\|_{2,\pi}^2 \leq (1-\gamma)^{2k}\|f_0 - 1\|_{2,\pi}^2.$$

A direct computation shows

$$\|f_0 - 1\|_{2,\pi}^2 = \frac{1-\pi(v)}{\pi(v)},$$

hence

$$\chi^2\Big(p_v^{(k)} \,\|\, \pi\Big) \leq \frac{1-\pi(v)}{\pi(v)}(1-\gamma)^{2k}.$$

Using the inequality $H^2(p,\pi) \le \frac{1}{2}\chi^2(p\,\|\,\pi)$, we obtain

$$H^2\Big(p_v^{(k)},\pi\Big) \le \tfrac{1}{2}\,\frac{1-\pi(v)}{\pi(v)}(1-\gamma)^{2k}.$$

By Lemma 4.2, the centroid $\bar{\phi}^{(k)}$ minimises the mean squared deviation, so

$$\mathcal{D}_k \le \frac{1}{n}\sum_v \|\phi_v^{(k)} - \sqrt{\pi}\|_2^2 = \frac{2}{n}\sum_v H^2\Big(p_v^{(k)},\pi\Big) \le \frac{1}{n}\sum_v \frac{1-\pi(v)}{\pi(v)}(1-\gamma)^{2k}.$$

In particular, using $(1-\pi(v))/\pi(v) \le \pi_{\min}^{-1} - 1$ yields

$$\mathcal{D}_k \le (\pi_{\min}^{-1}-1)(1-\gamma)^{2k},$$

and the cruder bound $\mathcal{D}_k \le \pi_{\min}^{-1}(1-\gamma)^{2k}$ follows immediately.

Part (ii) – Lower bound.

The pointwise inequality $(\sqrt{a}-\sqrt{b})^2 \ge \frac{1}{4}(a-b)^2$ for $a,b\in[0,1]$, summed coordinatewise and averaged over vertex pairs, gives

$$\mathcal{D}_k \;\ge\; \tfrac{1}{4}\,\Delta_k, \qquad \Delta_k = \frac{1}{n}\sum_v \big\|p_v^{(k)} - \bar{p}^{(k)}\big\|_2^2, \tag{4}$$

where $\bar{p}^{(k)} = \frac{1}{n}\sum_v p_v^{(k)}$ is the uniform vertex average. It therefore suffices to lower-bound Δ_k.

Spectral decomposition in the natural inner product. Let $f_1 \equiv 1, f_2, \dots, f_n$ denote the right eigenfunctions of the lazy random walk operator P, orthonormal in $L^2(\pi)$, with eigenvalues $1 = \lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ (so $\lambda_2 = 1-\gamma$). For reversible chains [Levin et al., 2017, Lemma 12.2] one has

$$p_v^{(k)}(j) \;=\; \pi(j)\sum_{i=1}^{n}\lambda_i^k\, f_i(v)\, f_i(j) \;=\; \pi(j) + \pi(j)\sum_{i\ge 2}\lambda_i^k\, f_i(v)\, f_i(j).$$

Writing $\bar{f}_i := \frac{1}{n}\sum_v f_i(v)$ and subtracting the vertex-averaged version yields

$$p_v^{(k)}(j) - \bar{p}^{(k)}(j) \;=\; \pi(j)\sum_{i\ge 2}\lambda_i^k\big(f_i(v)-\bar{f}_i\big)f_i(j). \tag{5}$$

Crucially, the eigenfunction product $f_i(j)f_h(j)$ is diagonalised only by the π-weighted sum: $\sum_j \pi(j)f_i(j)f_h(j) = \delta_{ih}$. We therefore introduce the χ^2-type quantity

$$\widetilde{\Delta}_k \;:=\; \frac{1}{n}\sum_v\sum_j \frac{\big(p_v^{(k)}(j) - \bar{p}^{(k)}(j)\big)^2}{\pi(j)},$$

which carries exactly the right weighting to diagonalise (5).

Diagonalisation. Substituting (5) and exchanging sums,

$$\begin{aligned}
\widetilde{\Delta}_k &= \frac{1}{n}\sum_v\sum_j \pi(j)\Big(\sum_{i\ge 2}\lambda_i^k(f_i(v)-\bar{f}_i)f_i(j)\Big)^2 \\
&= \frac{1}{n}\sum_v \sum_{i,h\ge 2}\lambda_i^k\lambda_h^k(f_i(v)-\bar{f}_i)(f_h(v)-\bar{f}_h)\underbrace{\sum_j \pi(j)f_i(j)f_h(j)}_{=\delta_{ih}} \\
&= \sum_{i\ge 2}\lambda_i^{2k}\,\sigma_i^2, \qquad \sigma_i^2 := \frac{1}{n}\sum_v\big(f_i(v)-\bar{f}_i\big)^2.
\end{aligned}$$

The cross-terms vanish *exactly*, and the coefficients $\sigma_i^2 \ge 0$ are independent of k, depending only on the graph spectrum.

Non-degeneracy of the second mode. We claim $\sigma_2^2 > 0$. Indeed, $\langle f_2, f_1 \rangle_\pi = \sum_v \pi(v) f_2(v) = 0$ together with $\|f_2\|_\pi = 1$ forces f_2 to be non-constant on V, hence its uniform empirical variance σ_2^2 is strictly positive. (If λ_2 has multiplicity $r > 1$, the same argument applied to any orthonormal basis of the second eigenspace gives $\sum_{i:\lambda_i=\lambda_2} \sigma_i^2 > 0$, which suffices in what follows.)

Asymptotic lower bound. Retaining only the $i = 2$ term and using $\lambda_2 = 1 - \gamma$,

$$\widetilde{\Delta}_k \;\geq\; \sigma_2^2 \, (1-\gamma)^{2k} \qquad \text{for all } k \geq 0.$$

Converting back to the unweighted Euclidean norm, the elementary inequality $\sum_j x_j^2 \;\geq\; \pi_{\min} \sum_j x_j^2 / \pi(j)$ yields

$$\Delta_k \;\geq\; \pi_{\min} \, \widetilde{\Delta}_k \;\geq\; \pi_{\min} \sigma_2^2 \, (1-\gamma)^{2k}.$$

Combining with (4),

$$\mathcal{D}_k \;\geq\; \frac{\pi_{\min} \sigma_2^2}{4} \, (1-\gamma)^{2k} \qquad \text{for every } k \geq 0.$$

This completes the proof.

□

B.5 Hellinger–ℓ_2 equivalence

The proof of Theorem 5.5 transfers the geometric content of $\mathcal{D}_k$ from the Hellinger metric on the simplex to the ℓ_2 metric on $\mathbb{R}^n$. The following lemma provides the bridge.

Lemma B.5 (Hellinger–ℓ_2 equivalence)**.** *For any probability distributions $p, q \in \Delta^{n-1}$,*

$$\frac{4}{n} \, H^4(p,q) \;\leq\; \|p - q\|_2^2 \;\leq\; 8 \, H^2(p,q). \tag{16}$$

Consequently, let

$$\mathcal{D}_k = \frac{1}{n^2} \sum_{u,v} H^2(p_u^{(k)}, p_v^{(k)}), \qquad \Delta_k = \frac{1}{2n^2} \sum_{u,v} \|p_u^{(k)} - p_v^{(k)}\|_2^2.$$

Then

$$\frac{2 \, \mathcal{D}_k^2}{n} \;\leq\; \Delta_k \;\leq\; 4 \, \mathcal{D}_k. \tag{17}$$

Proof. We proceed to show each bound separately, and then we conclude with the average bounds.

Upper bound. For each coordinate i we factor

$$p_i - q_i = (\sqrt{p_i} + \sqrt{q_i})(\sqrt{p_i} - \sqrt{q_i}).$$

Squaring and summing over i gives

$$\|p - q\|_2^2 = \sum_{i=1}^n (\sqrt{p_i} + \sqrt{q_i})^2 (\sqrt{p_i} - \sqrt{q_i})^2.$$

Because $p_i, q_i \in [0,1]$, we have $\sqrt{p_i}, \sqrt{q_i} \in [0,1]$, hence $(\sqrt{p_i} + \sqrt{q_i})^2 \leq 4$. Thus

$$\|p - q\|_2^2 \leq 4 \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2 = 4 \, \|\sqrt{p} - \sqrt{q}\|_2^2.$$

By definition of the Hellinger distance, $\|\sqrt{p} - \sqrt{q}\|_2^2 = 2H^2(p,q)$. Consequently,

$$\|p - q\|_2^2 \leq 8 \, H^2(p,q),$$

which is the right-hand inequality of (16).

Lower bound. The Cauchy–Schwarz inequality yields

$$\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i| \leq \sqrt{n} \, \|p - q\|_2.$$

The Hellinger–total-variation inequality (Lemma A.3) states $\|p-q\|_1 \geq 2H^2(p,q)$. Combining,

$$\|p-q\|_2 \geq \frac{1}{\sqrt{n}}\|p-q\|_1 \geq \frac{2}{\sqrt{n}}\, H^2(p,q).$$

Squaring gives

$$\|p-q\|_2^2 \geq \frac{4}{n}\, H^4(p,q),$$

the left-hand inequality of (16).

Averaged bounds for Δ_k. Using the upper bound for every pair (u,v) and averaging,

$$\Delta_k = \frac{1}{2n^2}\sum_{u,v}\|p_u^{(k)} - p_v^{(k)}\|_2^2 \leq \frac{1}{2n^2}\sum_{u,v} 8\, H^2(p_u^{(k)}, p_v^{(k)}) = 4\, \underbrace{\frac{1}{n^2}\sum_{u,v} H^2(p_u^{(k)}, p_v^{(k)})}_{\mathcal{D}_k} = 4\,\mathcal{D}_k.$$

For the lower bound we similarly insert the pointwise inequality:

$$\Delta_k \geq \frac{1}{2n^2}\sum_{u,v}\frac{4}{n}\, H^4(p_u^{(k)}, p_v^{(k)}) = \frac{2}{n}\,\frac{1}{n^2}\sum_{u,v} H^4(p_u^{(k)}, p_v^{(k)}).$$

Jensen's inequality (applied to the convex function $t \mapsto t^2$) implies

$$\frac{1}{n^2}\sum_{u,v} H^4(p_u^{(k)}, p_v^{(k)}) \geq \left(\frac{1}{n^2}\sum_{u,v} H^2(p_u^{(k)}, p_v^{(k)})\right)^2 = \mathcal{D}_k^2.$$

Hence

$$\Delta_k \geq \frac{2}{n}\,\mathcal{D}_k^2,$$

completing the proof of (17). □

B.6 Theorem 5.5: linear MPNN bound

Theorem B.6 (Linear MPNN – equivalence of R_k and $\mathcal{D}_k$)**.** *Consider a linear k-layer MPNN with output $x_v^{(k)} = p_v^{(k)}\mathbf{X}^{(0)}\mathbf{M}$ and let $Z = \mathbf{X}^{(0)}\mathbf{M}$. Denote by $\sigma_{\min}$ the smallest singular value of $Z^\top$ on the subspace of mean-zero vectors, i.e.*

$$\sigma_{\min} = \inf\big\{\|Z^\top z\|_2 : \mathbf{1}^\top z = 0,\ \|z\|_2 = 1\big\}.$$

Then

$$\frac{2\sigma_{\min}^2}{n}\,\mathcal{D}_k^2 \;\leq\; R_k \;\leq\; 4\,\|Z\|_{\mathrm{op}}^2\,\mathcal{D}_k. \tag{18}$$

Consequently, if $\sigma_{\min} > 0$,

$$\mathcal{D}_k \to 0 \quad\Longleftrightarrow\quad R_k \to 0.$$

Proof. The linear output difference is $x_u^{(k)} - x_v^{(k)} = (p_u^{(k)} - p_v^{(k)})Z$. Averaging the squared ℓ_2 norm over all pairs gives

$$R_k = \frac{1}{2n^2}\sum_{u,v}\|(p_u^{(k)} - p_v^{(k)})Z\|_2^2 \leq \|Z\|_{\mathrm{op}}^2\,\Delta_k \leq 4\|Z\|_{\mathrm{op}}^2\,\mathcal{D}_k,$$

where the last inequality uses the upper bound $\Delta_k \leq 4\mathcal{D}_k$ from Lemma B.5.

For the lower bound, note that $p_u^{(k)} - p_v^{(k)}$ is mean-zero. Hence

$$\|\top(p_u^{(k)} - p_v^{(k)})Z\|_2 \geq \sigma_{\min}\,\|p_u^{(k)} - p_v^{(k)}\|_2.$$

Squaring and averaging yields $R_k \geq \sigma_{\min}^2\,\Delta_k$. Lemma B.5 gives $\Delta_k \geq \frac{2}{n}\mathcal{D}_k^2$, and combining these proves the claimed lower bound.

From the two inequalities, $\mathcal{D}_k \to 0 \Rightarrow R_k \to 0$ is immediate. Conversely, if $R_k \to 0$ and $\sigma_{\min} > 0$, then $\mathcal{D}_k^2 \leq \frac{n}{2\sigma_{\min}^2} R_k \to 0$, so $\mathcal{D}_k \to 0$. □

B.7 Theorem 5.6: nonlinear MPNN contraction

Theorem B.7 (Nonlinear MPNN)**.** *Consider a k-layer MPNN defined by*

$$\mathbf{X}^{(\ell+1)} = \sigma\big(\mathbf{P}\,\mathbf{X}^{(\ell)}\mathbf{W}^{(\ell)}\big), \qquad \ell = 0, \dots, k-1,$$

where $\mathbf{P}$ is the lazy walk, σ is applied row-wise and is 1-Lipschitz, and $\|\mathbf{W}^{(\ell)}\|_{\mathrm{op}} \le 1$ for all layers. Then

$$R_k \;\le\; C\,\mathcal{D}_k,$$

where $C > 0$ depends on the graph's degree heterogeneity and on the initial dispersion.

Proof. Let $R_\pi(\mathbf{Y})$ denote the stationary-mean dispersion of a feature matrix $\mathbf{Y}$ with rows y_v:

$$R_\pi(\mathbf{Y}) = \sum_{v=1}^{n} \pi_v \left\| y_v - \bar{y}_\pi \right\|_2^2, \quad \bar{y}_\pi = \sum_{v=1}^{n} \pi_v y_v.$$

We show that for every layer ℓ,

$$R_\pi(\mathbf{X}^{(\ell+1)}) \le (1-\gamma)^2\, R_\pi(\mathbf{X}^{(\ell)}).$$

Propagation. The stationary distribution satisfies $\pi^\top \mathbf{P} = \pi^\top$. Define the centered matrix $\mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\bar{y}_\pi^\top$; its π-weighted rows sum to zero. Because the lazy walk is reversible, $\mathbf{P}$ is self-adjoint on $L_2(\pi)$, and its operator norm restricted to the π-mean-zero subspace is exactly $\lambda_2(\mathbf{P}) = 1-\gamma$. Thus

$$R_\pi(\mathbf{P}\mathbf{Y}) = \|\mathbf{P}\mathbf{Y}_c\|_{L_2(\pi)}^2 \le \lambda_2(\mathbf{P})^2 \|\mathbf{Y}_c\|_{L_2(\pi)}^2 = (1-\gamma)^2\, R_\pi(\mathbf{Y}).$$

Weight multiplication. Since $\|\mathbf{W}^{(\ell)}\|_{\mathrm{op}} \le 1$, we have

$$R_\pi(\mathbf{Y}\mathbf{W}^{(\ell)}) \le \|\mathbf{W}^{(\ell)}\|_{\mathrm{op}}^2\, R_\pi(\mathbf{Y}) \le R_\pi(\mathbf{Y}).$$

Activation. Using the pairwise identity $R_\pi(\mathbf{Y}) = \frac{1}{2}\sum_{u,v} \pi_u \pi_v \|y_u - y_v\|_2^2$ and the 1-Lipschitz property of σ,

$$R_\pi(\sigma(\mathbf{Y})) = \frac{1}{2}\sum_{u,v} \pi_u \pi_v \|\sigma(y_u) - \sigma(y_v)\|_2^2 \le \frac{1}{2}\sum_{u,v} \pi_u \pi_v \|y_u - y_v\|_2^2 = R_\pi(\mathbf{Y}).$$

Chaining the three operations gives

$$R_\pi(\mathbf{X}^{(\ell+1)}) \le (1-\gamma)^2\, R_\pi(\mathbf{X}^{(\ell)}),$$

and iterating over k layers yields

$$R_k^\pi \;\le\; (1-\gamma)^{2k}\, R_0^\pi. \tag{1}$$

Using the degree-based relations proved in Appendix A.5 (inequality (13)), we have

$$R_k \;\le\; \frac{d_{\mathrm{avg}}}{d_{\min}}\, R_k^\pi, \qquad R_0^\pi \;\le\; \frac{d_{\max}}{d_{\mathrm{avg}}}\, R_0.$$

Combining these with the contraction bound (1) from Step 1,

$$R_k^\pi \;\le\; (1-\gamma)^{2k}\, R_0^\pi,$$

we obtain

$$R_k \;\le\; \frac{d_{\mathrm{avg}}}{d_{\min}}\,(1-\gamma)^{2k}\, R_0^\pi \;\le\; \frac{d_{\mathrm{avg}}}{d_{\min}}\,(1-\gamma)^{2k}\,\frac{d_{\max}}{d_{\mathrm{avg}}}\, R_0 \;=\; \frac{d_{\max}}{d_{\min}}\,(1-\gamma)^{2k}\, R_0.$$

Denoting $\rho = d_{\max}/d_{\min} = \pi_{\max}/\pi_{\min}$, we have

$$R_k \;\le\; \rho\,(1-\gamma)^{2k}\, R_0. \tag{2}$$

Theorem B.4(ii) guarantees the existence of a constant $c > 0$ (depending only on the graph and the initial distribution) such that

$$\mathcal{D}_k \;\ge\; c\,(1-\gamma)^{2k} \quad \text{for all } k \ge 0.$$

Hence $(1-\gamma)^{2k} \le c^{-1}\,\mathcal{D}_k$. Substituting this into (2) yields

$$R_k \;\le\; \frac{\rho\, R_0}{c}\,\mathcal{D}_k.$$

This completes the proof. □

C Datasets and baseline diagnostics

This appendix provides the experimental scaffolding for all results reported in §6 and Appendix F: dataset statistics (§C.1), train/validation/test splits (§C.2), and formal definitions of the seven post-training baseline diagnostics we compare against (§C.3).

C.1 Dataset statistics

We evaluate on 11 node-classification benchmarks spanning two orders of magnitude in scale ($n \in [183, 169{,}343]$) and the full range of homophily ($h \in [0.05, 0.81]$). Table 7 summarizes the key structural statistics of these datasets.

Table 7: Dataset statistics. n is the number of nodes; m is the number of undirected edges; C is the number of classes; d_{feat} is the input feature dimension; $\bar{d}$ is the mean degree; $\mathcal{D}$ is the graph diameter; h is the edge homophily, defined as the fraction of edges connecting nodes of the same class. Datasets are grouped by homophily regime and ordered within each group by node count.

Dataset	n	m	C	d_{feat}	$\bar{d}$	$\mathcal{D}$	h	Domain
Homophilic ($h \geq 0.5$)								
Cora	2,708	5,278	7	1,433	3.90	19	0.81	Citation
CiteSeer	3,327	4,552	6	3,703	2.74	28	0.74	Citation
PubMed	19,717	44,324	3	500	4.50	18	0.80	Citation
ogbn-arxiv	169,343	1,166,243	40	128	13.77	23	0.65	Citation
Heterophilic ($h < 0.5$)								
Texas	183	295	5	1,703	3.22	8	0.11	WebKB
Cornell	183	277	5	1,703	3.03	8	0.30	WebKB
Wisconsin	251	466	5	1,703	3.71	8	0.21	WebKB
Chameleon	2,277	31,371	5	2,325	27.55	11	0.23	Wikipedia
Squirrel	5,201	198,353	5	2,089	76.28	10	0.22	Wikipedia
RomanEmpire	22,662	32,927	18	300	2.91	6,824	0.05	Wikipedia
AmazonRatings	24,492	93,050	5	300	7.60	46	0.38	Co-purchase

Sources. The citation networks Cora, CiteSeer, and PubMed follow Yang et al. [2016]; the WebKB datasets (Texas, Cornell, Wisconsin) and the Wikipedia heterophilic graphs Chameleon and Squirrel follow Pei et al. [2020]; RomanEmpire and AmazonRatings follow Platonov et al. [2023]; ogbn-arxiv follows Hu et al. [2020].

Note on Chameleon and Squirrel. We use the original Pei et al. [2020] releases of Chameleon and Squirrel ($n = 2{,}277$ and $n = 5{,}201$) rather than the filtered versions of Platonov et al. [2023] ($n = 890$ and $n = 2{,}223$). Platonov et al. [2023] document that the original graphs contain a substantial number of duplicate nodes that can leak between train and test partitions; their filtered versions remove this leakage. Averaging over ten random splits reduces variance in our reported accuracies but does not remove this leakage, since the duplicates are a property of the graph rather than of any particular split. Absolute accuracies on these two datasets are therefore not directly comparable to post-Platonov numbers. The diagnostic claim of this paper concerns the *within-dataset, within-architecture* correlation between $\mathcal{D}_k$ and the depth-accuracy curve: a constant offset in absolute accuracy from leakage shifts the level of that curve but does not redistribute the depth-wise mixing schedule $\mathcal{D}_k$ tracks. We flag this as a limitation and direct readers to the filtered versions for any cross-paper comparison of absolute accuracy.

C.2 Splits and training protocol

We use the standard Planetoid splits for Cora, CiteSeer, and PubMed [Yang et al., 2016]; the predefined split (split 0) of Platonov et al. [2023] for RomanEmpire and AmazonRatings; and the official time-based split for ogbn-arxiv. For Texas, Wisconsin, Cornell, Chameleon, and Squirrel, we generate ten random 60%/20%/20% train/validation/test splits, one per seed, and average results over the ten seeds.

We report results averaged over 10 random seeds per (dataset, model, depth) configuration. The hidden dimension is 32 for most datasets and 64 for RomanEmpire, AmazonRatings, and ogbn-arxiv. We use dropout of 0.5 and the Adam optimizer with learning rate 10^{-2} and weight decay 5×10^{-4}. Models are trained for up to 200 epochs with early stopping (patience 10) based on validation performance, and we restore the best checkpoint.

C.3 Baseline diagnostic definitions

We compare $\mathcal{D}_k$ against seven post-training diagnostics computed on the final-layer representation matrix $\mathbf{X}^{(L)} \in \mathbb{R}^{n \times d}$ produced by an L-layer GNN. To fix notation: $\mathbf{x}_v \in \mathbb{R}^d$ denotes the v-th row of $\mathbf{X}^{(L)}$; $\mathbf{u} \in \mathbb{R}^n$ is the dominant eigenvector of the message-passing matrix $\mathbf{A}$, normalised so that $\|\mathbf{u}\|_2 = 1$ (for the symmetric augmented adjacency $\mathbf{A} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ used by GCN, $\mathbf{u}$ is proportional to $(\sqrt{1+d_i})_{i=1}^n$; for the row-stochastic aggregators used by GAT, $\mathbf{u} \propto \mathbf{1}$); $\mathbf{P} = \mathbf{u}\mathbf{u}^\top$ is the orthogonal projection onto $\mathrm{span}\{\mathbf{u}\}$; and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(n,d)} \geq 0$ are the singular values of $\mathbf{X}^{(L)}$. We follow the formulations and experimental conventions of Zhang et al. [2026].

Dirichlet energy ($\mathcal{E}_{\mathrm{Dir}}$, $\mathcal{E}_{\mathrm{Dir,n}}$). Cai and Wang [2020] introduced the Dirichlet energy as a smoothness functional aligned with the dominant eigenspace of the aggregation matrix:

$$\mathcal{E}_{\mathrm{Dir}}(\mathbf{X}^{(L)}) = \sum_{(i,j)\in E} \left\| \frac{\mathbf{x}_i}{u_i} - \frac{\mathbf{x}_j}{u_j} \right\|_2^2. \tag{19}$$

The eigenvector rescaling makes $\mathcal{E}_{\mathrm{Dir}}$ vanish exactly when features align with $\mathbf{u}$; for GCN it reduces to the standard degree-normalised form, and for GAT (where $\mathbf{u} \propto \mathbf{1}$) to the unweighted edge-difference sum. Following Zhang et al. [2026], we also report the scale-invariant variant

$$\mathcal{E}_{\mathrm{Dir,n}}(\mathbf{X}^{(L)}) = \mathcal{E}_{\mathrm{Dir}}(\mathbf{X}^{(L)}) / \|\mathbf{X}^{(L)}\|_F^2, \tag{20}$$

which removes the dependency of $\mathcal{E}_{\mathrm{Dir}}$ on the global magnitude of the representations.

Projected energy ($\mathcal{E}_{\mathrm{Proj}}$, $\mathcal{E}_{\mathrm{Proj,n}}$). Wu et al. [2023] measure the deviation of features from the dominant-eigenvector subspace,

$$\mathcal{E}_{\mathrm{Proj}}(\mathbf{X}^{(L)}) = \|\mathbf{X}^{(L)} - \mathbf{P}\mathbf{X}^{(L)}\|_F^2 = \|(\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{X}^{(L)}\|_F^2, \tag{21}$$

with normalised counterpart

$$\mathcal{E}_{\mathrm{Proj,n}}(\mathbf{X}^{(L)}) = \mathcal{E}_{\mathrm{Proj}}(\mathbf{X}^{(L)}) / \|\mathbf{X}^{(L)}\|_F. \tag{22}$$

$\mathcal{E}_{\mathrm{Proj}}$ and $\mathcal{E}_{\mathrm{Dir}}$ are equivalent up to constants depending on the entries of $\mathbf{u}$ [Zhang et al., 2026, Lemma A.1]; in particular, both metrics presuppose that the dominant eigenvector of $\mathbf{A}^{(\ell)}$ is the same for every layer ℓ.

Mean Average Distance (MAD). Chen et al. [2020a] define MAD as the mean cosine distance over a target set of node pairs. We adopt the edge-set instantiation used by Zhang et al. [2026],

$$\mathrm{MAD}(\mathbf{X}^{(L)}) = \frac{1}{|E|} \sum_{(i,j)\in E} \left(1 - \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}\right), \tag{23}$$

which measures angular dispersion of representations along the graph's edges and is invariant to row magnitudes. Unlike $\mathcal{E}_{\mathrm{Dir}}$ and $\mathcal{E}_{\mathrm{Proj}}$, MAD does not require a fixed dominant eigenvector to be defined.

Effective rank (Erank). Following Roy and Vetterli [2007], the effective rank is the exponential of the Shannon entropy of the normalised singular-value distribution. Writing $\bar{\sigma}_i = \sigma_i / \sum_j \sigma_j$,

$$\mathrm{Erank}(\mathbf{X}^{(L)}) = \exp\left(-\sum_i \bar{\sigma}_i \log \bar{\sigma}_i\right). \tag{24}$$

Erank ranges from 1 (one direction concentrates all spectral mass) to $\min(n, d)$ (uniform spectrum), and continuously interpolates between these extremes. Zhang et al. [2026] advocate it as an oversmoothing diagnostic because, unlike $\mathcal{E}_{\mathrm{Dir}}$ and $\mathcal{E}_{\mathrm{Proj}}$, it does not require a known dominant eigenspace and is scale-invariant in $\mathbf{X}^{(L)}$.

Numerical rank (NumRank). Zhang et al. [2026] use the continuous Frobenius-to-spectral ratio

$$\mathrm{NumRank}(\mathbf{X}^{(L)}) = \frac{\|\mathbf{X}^{(L)}\|_F^2}{\|\mathbf{X}^{(L)}\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}. \tag{25}$$

$\mathrm{NumRank}$ is bounded between 1 (rank-one features) and $\min(n, d)$, is scale-invariant, and converges to one whenever $\sigma_2/\sigma_1 \to 0$. The bound $\mathrm{NumRank}(\mathbf{X}) \le 1 + \|(\mathbf{I} - \mathbf{P})\mathbf{X}\|_F^2/\|\mathbf{X}\|_2^2$ [Zhang et al., 2026, eq. 4] ties its decay directly to the projected energy.

D Estimator and computation

This section addresses the practical question of computing $\mathcal{D}_k$ at scale. We first establish the unbiasedness and U-statistic form of the sampling estimator (§D.1), then prove its exponential concentration via McDiarmid's inequality (§D.2), then report runtime measurements across our benchmark suite (§D.4), and finally validate the theoretical sample-complexity rate empirically (§D.3).

D.1 Unbiased estimator and U-statistic form

For graphs with $n \gtrsim 20{,}000$ nodes, exact computation of $\mathcal{D}_k$ becomes expensive: each walk distribution $p_v^{(k)} = e_v^\top \mathbf{P}^k$ requires k sparse matrix–vector products at cost $O(m)$, so the full computation is $O(nmk)$. The bottleneck is the factor of n, not the depth k. We therefore sample s source nodes $v_1, \dots, v_s$ uniformly at random from V and form the empirical centroid

$$\widehat{\bar{\phi}}^{(k)} = \frac{1}{s}\sum_{i=1}^{s} \phi_{v_i}^{(k)}, \qquad \phi_{v_i}^{(k)} := \sqrt{p_{v_i}^{(k)}}. \tag{26}$$

The bias-corrected estimator is

$$\widehat{\mathcal{D}}_k := \frac{s}{s-1}\left(1 - \|\widehat{\bar{\phi}}^{(k)}\|_2^2\right). \tag{27}$$

The factor $s/(s-1)$ corrects the downward bias inherent in plugging an empirical mean into a centred quadratic—the same correction that converts the biased sample variance to its unbiased form.

Proposition D.1 (Unbiased estimator and U-statistic form)**.** *The estimator* $\hat{\mathcal{D}}_k := \frac{s}{s-1}\big(1 - \|\hat{\bar{\phi}}^{(k)}\|_2^2\big)$ *satisfies* $\mathbb{E}[\hat{\mathcal{D}}_k] = \mathcal{D}_k$ *for* $s \ge 2$*, and admits the representation*

$$\hat{\mathcal{D}}_k \;=\; \frac{1}{s(s-1)} \sum_{i \ne j} H^2\big(p_{v_i}^{(k)},\, p_{v_j}^{(k)}\big).$$

Proof. Expanding the squared norm of the empirical centroid:

$$\|\hat{\bar{\phi}}^{(k)}\|_2^2 \;=\; \Big\|\frac{1}{s}\sum_{i=1}^{s}\phi_{v_i}^{(k)}\Big\|^2 \;=\; \frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\langle \phi_{v_i}^{(k)},\, \phi_{v_j}^{(k)}\rangle \;=\; \frac{1}{s^2}\sum_{i=1}^{s}\|\phi_{v_i}^{(k)}\|^2 \;+\; \frac{1}{s^2}\sum_{i\ne j}\langle \phi_{v_i}^{(k)},\, \phi_{v_j}^{(k)}\rangle.$$

Since $\|\phi_{v_i}^{(k)}\|^2 = 1$ for every i, the diagonal sum equals $s/s^2 = 1/s$, giving

$$\|\hat{\bar{\phi}}^{(k)}\|_2^2 \;=\; \frac{1}{s} \;+\; \frac{1}{s^2}\sum_{i\ne j}\langle \phi_{v_i}^{(k)},\, \phi_{v_j}^{(k)}\rangle. \tag{28}$$

For $i \ne j$, the samples v_i and v_j are independent and each uniform on V, so their joint distribution satisfies $\Pr(v_i = u,\, v_j = w) = 1/n^2$ for all $u, w \in V$. By the law of the unconscious statistician:

$$\mathbb{E}\big[\langle \phi_{v_i}^{(k)}, \phi_{v_j}^{(k)}\rangle\big] \;=\; \frac{1}{n^2}\sum_{u\in V}\sum_{w\in V}\langle \phi_u^{(k)}, \phi_w^{(k)}\rangle.$$

Bilinearity of the inner product allows us to factor the double sum:

$$\sum_{u\in V}\sum_{w\in V}\langle \phi_u^{(k)}, \phi_w^{(k)}\rangle \;=\; \Big\langle \sum_{u\in V}\phi_u^{(k)},\; \sum_{w\in V}\phi_w^{(k)}\Big\rangle \;=\; \Big\|\sum_{v\in V}\phi_v^{(k)}\Big\|^2,$$

where the first equality distributes the inner product over the first argument and then over the second. Dividing by n^2 and using $\|cx\|^2 = c^2\|x\|^2$:

$$\mathbb{E}\big[\langle \phi_{v_i}^{(k)}, \phi_{v_j}^{(k)}\rangle\big] \;=\; \Big\|\frac{1}{n}\sum_{v\in V}\phi_v^{(k)}\Big\|^2 \;=\; \|\bar{\phi}^{(k)}\|^2. \tag{29}$$

Taking expectations of (28) and applying (29) to each of the $s(s-1)$ off-diagonal pairs:

$$\mathbb{E}\big[\|\hat{\bar{\phi}}^{(k)}\|^2\big] \;=\; \frac{1}{s} + \frac{s(s-1)}{s^2}\|\bar{\phi}^{(k)}\|^2 \;=\; \frac{1}{s} + \frac{s-1}{s}\|\bar{\phi}^{(k)}\|^2,$$

so that

$$\mathbb{E}\big[1 - \|\hat{\bar{\phi}}^{(k)}\|^2\big] \;=\; 1 - \frac{1}{s} - \frac{s-1}{s}\|\bar{\phi}^{(k)}\|^2 \;=\; \frac{s-1}{s}\big(1 - \|\bar{\phi}^{(k)}\|^2\big) \;=\; \frac{s-1}{s}\,\mathcal{D}_k.$$

The plug-in $1 - \|\hat{\bar{\phi}}^{(k)}\|^2$ is thus downward biased by the factor $(s-1)/s$. Multiplying by $s/(s-1)$:

$$\mathbb{E}[\hat{\mathcal{D}}_k] \;=\; \frac{s}{s-1}\cdot\frac{s-1}{s}\,\mathcal{D}_k \;=\; \mathcal{D}_k.$$

Substituting (28) into $\hat{\mathcal{D}}_k = \frac{s}{s-1}(1 - \|\hat{\bar{\phi}}^{(k)}\|^2)$:

$$\hat{\mathcal{D}}_k = \frac{s}{s-1}\Big(\frac{s-1}{s} - \frac{1}{s^2}\sum_{i\neq j}\langle \phi_{v_i}^{(k)}, \phi_{v_j}^{(k)}\rangle\Big) \;=\; 1 - \frac{1}{s(s-1)}\sum_{i\neq j}\langle \phi_{v_i}^{(k)}, \phi_{v_j}^{(k)}\rangle.$$

Recalling that $H^2(p,q) = 1 - \sum_u \sqrt{p(u)q(u)} = 1 - \langle\sqrt{p}, \sqrt{q}\rangle$, we substitute $\langle \phi_{v_i}^{(k)}, \phi_{v_j}^{(k)}\rangle = 1 - H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)})$:

$$\begin{aligned}
\hat{\mathcal{D}}_k &= 1 - \frac{1}{s(s-1)}\sum_{i\neq j}\big[1 - H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)})\big] \\
&= 1 - 1 + \frac{1}{s(s-1)}\sum_{i\neq j} H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)}) \\
&= \frac{1}{s(s-1)}\sum_{i\neq j} H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)}),
\end{aligned} \tag{30}$$

which is a degree-2 U-statistic with kernel $h(v_i, v_j) = H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)})$. □

The U-statistic form makes two facts transparent. First, $\widehat{\mathcal{D}}_k$ is a degree-2 U-statistic with bounded kernel $h(v_i, v_j) = H^2(p_{v_i}^{(k)}, p_{v_j}^{(k)}) \in [0,1]$, which places it in a setting where standard concentration inequalities apply. Second, the centroid form $\frac{s}{s-1}(1 - \|\hat{\bar{\phi}}^{(k)}\|_2^2)$ and the pairwise form $\frac{1}{s(s-1)}\sum_{i\neq j} H^2$ are algebraically identical—but the centroid form is computed in $O(sn)$ time and $O(sn)$ memory, whereas the pairwise form requires $O(s^2 n)$ time. We always use the centroid form in practice.

D.2 Concentration

We now show that $\widehat{\mathcal{D}}_k$ concentrates exponentially around $\mathcal{D}_k$, with the standard $\epsilon^{-2}\log(1/\delta)$ sample complexity.

Proposition D.2 (Concentration via McDiarmid's inequality)**.** *For any $\epsilon > 0$,*

$$\Pr\big(|\hat{\mathcal{D}}_k - \mathcal{D}_k| \geq \epsilon\big) \;\leq\; 2\exp\!\left(-\frac{\epsilon^2(s-1)^2}{8s}\right).$$

Consequently, $s = O(\epsilon^{-2}\log(1/\delta))$ samples suffice for $|\hat{\mathcal{D}}_k - \mathcal{D}_k| \leq \epsilon$ with probability at least $1-\delta$.

Proof. To prove the property, we proceed in four steps:

Replace the ℓ-th sample v_ℓ with v'_ℓ. The empirical centroids $\hat{\bar{\phi}}^{(k)}$ and $\hat{\bar{\phi}}^{(k)\prime}$ differ by

$$\hat{\bar{\phi}}^{(k)\prime} - \hat{\bar{\phi}}^{(k)} \;=\; \frac{1}{s}\big(\phi^{(k)}_{v'_\ell} - \phi^{(k)}_{v_\ell}\big).$$

By the triangle inequality and $\|\phi^{(k)}_v\| = 1$ for all v:

$$\|\hat{\bar{\phi}}^{(k)\prime} - \hat{\bar{\phi}}^{(k)}\| \;=\; \frac{1}{s}\|\phi^{(k)}_{v'_\ell} - \phi^{(k)}_{v_\ell}\| \;\leq\; \frac{1}{s}\big(\|\phi^{(k)}_{v'_\ell}\| + \|\phi^{(k)}_{v_\ell}\|\big) \;=\; \frac{2}{s}.$$

By definition, $\hat{\mathcal{D}}_k = \frac{s}{s-1}(1 - \|\hat{\bar{\phi}}^{(k)}\|^2)$, so

$$|\hat{\mathcal{D}}'_k - \hat{\mathcal{D}}_k| \;=\; \frac{s}{s-1}\big|\,\|\hat{\bar{\phi}}^{(k)}\|^2 - \|\hat{\bar{\phi}}^{(k)\prime}\|^2\,\big|.$$

Applying the difference of squares identity $\|a\|^2 - \|b\|^2 = \langle a+b,\, a-b\rangle$ followed by Cauchy–Schwarz:

$$\big|\,\|\hat{\bar{\phi}}^{(k)}\|^2 - \|\hat{\bar{\phi}}^{(k)\prime}\|^2\,\big| \;=\; \big|\langle \hat{\bar{\phi}}^{(k)} + \hat{\bar{\phi}}^{(k)\prime},\; \hat{\bar{\phi}}^{(k)} - \hat{\bar{\phi}}^{(k)\prime}\rangle\big| \;\leq\; \|\hat{\bar{\phi}}^{(k)} + \hat{\bar{\phi}}^{(k)\prime}\| \cdot \|\hat{\bar{\phi}}^{(k)} - \hat{\bar{\phi}}^{(k)\prime}\|.$$

Since $\hat{\bar{\phi}}^{(k)}$ is a convex combination of unit vectors, $\|\hat{\bar{\phi}}^{(k)}\| \leq 1$ (by the triangle inequality), and likewise $\|\hat{\bar{\phi}}^{(k)\prime}\| \leq 1$, so $\|\hat{\bar{\phi}}^{(k)} + \hat{\bar{\phi}}^{(k)\prime}\| \leq 2$. Combining with Step 1:

$$|\hat{\mathcal{D}}'_k - \hat{\mathcal{D}}_k| \;\leq\; \frac{s}{s-1}\cdot 2 \cdot \frac{2}{s} \;=\; \frac{4}{s-1} \;=:\; c.$$

This holds uniformly for every coordinate $\ell \in \{1,\ldots,s\}$.

The samples $v_1,\ldots,v_s$ are independent and $\mathbb{E}[\hat{\mathcal{D}}_k] = \mathcal{D}_k$ by D.1. The sum of squared bounded-difference constants is

$$\sum_{\ell=1}^{s} c^2 \;=\; s\cdot\frac{16}{(s-1)^2} \;=\; \frac{16s}{(s-1)^2}.$$

McDiarmid's inequality states that for independent $X_1,\ldots,X_s$ and any function f with $\sup_{x_\ell, x'_\ell} |f(\ldots,x_\ell,\ldots) - f(\ldots,x'_\ell,\ldots)| \leq c_\ell$:

$$\Pr\big(|f - \mathbb{E}[f]| \geq \epsilon\big) \;\leq\; 2\exp\!\Big(-\frac{2\epsilon^2}{\sum_\ell c_\ell^2}\Big).$$

Substituting:

$$\Pr\big(|\hat{\mathcal{D}}_k - \mathcal{D}_k| \geq \epsilon\big) \;\leq\; 2\exp\!\left(-\frac{2\epsilon^2}{\frac{16s}{(s-1)^2}}\right) \;=\; 2\exp\!\left(-\frac{\epsilon^2(s-1)^2}{8s}\right).$$

Setting the bound to at most δ:

$$2\exp\!\left(-\frac{\epsilon^2(s-1)^2}{8s}\right) \leq \delta \quad\Longleftrightarrow\quad \frac{\epsilon^2(s-1)^2}{8s} \geq \ln\frac{2}{\delta}.$$

For $s \geq 2$, we have $(s-1)^2/s \geq (s-1)/2$ (since dividing $2(s-1)^2 \geq s(s-1)$ by $s-1>0$ gives $2(s-1)\geq s$, which holds for $s\geq 2$). Therefore a sufficient condition is

$$\frac{\epsilon^2(s-1)}{16} \geq \ln\frac{2}{\delta} \quad\Longleftrightarrow\quad s \geq 1 + \frac{16}{\epsilon^2}\ln\frac{2}{\delta},$$

confirming $s = O(\epsilon^{-2}\log(1/\delta))$. □

Beyond McDiarmid: a variance-aware refinement. Proposition D.2 is dimension-free, but it discards the variance of the kernel $h(u,w) = H^2(p_u^{(k)}, p_w^{(k)})$ and retains only its worst-case range through the bounded-difference constant $c = 4/(s-1)$. The price is paid at deep k. Since $\mathcal{D}_k$ decays exponentially in k (Theorem 5.3), the absolute guarantee $s = O(\epsilon^{-2}\log(1/\delta))$ translates, for relative error $\eta := \epsilon/\mathcal{D}_k$, to $s = O\big(\eta^{-2}\,\mathcal{D}_k^{-2}\log(1/\delta)\big)$, an exponential growth in k. A Bernstein-type bound for the U-statistic halves this exponent. Hoeffding's variance decomposition for the degree-2 U-statistic of Proposition D.1 gives

$$\mathrm{Var}(\hat{\mathcal{D}}_k) \;=\; \frac{4(s-2)}{s(s-1)}\,\zeta_1 \;+\; \frac{2}{s(s-1)}\,\zeta_2, \qquad \zeta_1 := \mathrm{Var}\big(g(V_1)\big), \quad \zeta_2 := \mathrm{Var}\big(h(V_1,V_2)\big),$$

with $g(u) := \mathbb{E}_w[h(u,w)]$. Expanding $H^2(p_u^{(k)}, p_w^{(k)}) = \frac{1}{2}\|\phi_u^{(k)} - \phi_w^{(k)}\|^2$ around the centroid yields the closed form $g(u) = \frac{1}{2}\|\phi_u^{(k)} - \bar{\phi}^{(k)}\|^2 + \frac{1}{2}\mathcal{D}_k$. Since $\phi_u^{(k)}$ and $\bar{\phi}^{(k)}$ have nonnegative entries with $\|\phi_u^{(k)}\| = 1$ and $\|\bar{\phi}^{(k)}\| \le 1$, we have $\|\phi_u^{(k)} - \bar{\phi}^{(k)}\|^2 \le 1 + \|\bar{\phi}^{(k)}\|^2 \le 2$, and the self-bound $\|\phi_u^{(k)} - \bar{\phi}^{(k)}\|^4 \le 2\,\|\phi_u^{(k)} - \bar{\phi}^{(k)}\|^2$ gives, on taking expectations,

$$\zeta_1 \;=\; \tfrac{1}{4}\,\mathbb{E}_u\|\phi_u^{(k)} - \bar{\phi}^{(k)}\|^4 - \tfrac{1}{4}\mathcal{D}_k^2 \;\le\; \tfrac{1}{2}\,\mathcal{D}_k, \qquad \zeta_2 \;\le\; \mathbb{E}[h^2] \;\le\; \mathbb{E}[h] \;=\; \mathcal{D}_k,$$

where $\mathbb{E}[h^2] \le \mathbb{E}[h]$ uses $h \in [0,1]$. Both kernel-variance terms inherit the exponential decay of $\mathcal{D}_k$, which McDiarmid's proof does not exploit. Substituting into a Bernstein-type inequality for non-degenerate U-statistics,

$$\Pr\big(|\hat{\mathcal{D}}_k - \mathcal{D}_k| \ge \epsilon\big) \;\le\; 2\exp\!\left(-\frac{c_1\, s\,\epsilon^2}{\mathcal{D}_k + c_2\,\epsilon}\right)$$

for absolute constants $c_1, c_2 > 0$, and setting $\epsilon = \eta\,\mathcal{D}_k$,

$$s \;=\; O\Big(\tfrac{1}{\eta^2\,\mathcal{D}_k}\,\log(1/\delta)\Big),$$

which scales as $\mathcal{D}_k^{-1}$ rather than $\mathcal{D}_k^{-2}$, halving the exponent in k. We retain Proposition D.2 as the formal guarantee for its clean, dimension-free form, and read the variance-aware refinement as the explanation for why the empirical convergence in Figure 2 is faster than either bound suggests: at $k = 16$, where $\mathcal{D}_k$ has decayed by several orders of magnitude relative to $\mathcal{D}_4$, the 3% relative-error threshold is reached at $s = 50$ uniformly across datasets, well below the $\mathcal{D}_k^{-1}$ rate Bernstein predicts.

D.3 Empirical convergence of the sampling estimator

Proposition D.2 predicts $|\widehat{\mathcal{D}}_k - \mathcal{D}_k| = O(s^{-1/2})$. Figure 2 confirms this rate empirically on three datasets: PubMed, RomanEmpire, and AmazonRatings.

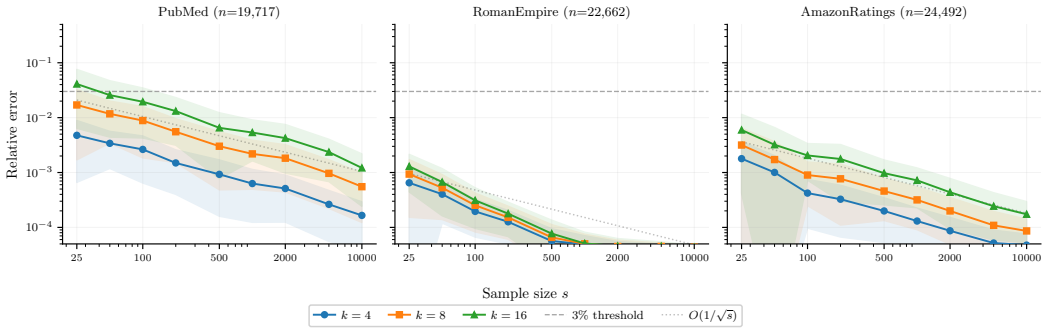

Figure 2: Sampling estimator convergence. Relative error vs. sample size s for three datasets at depths $k \in \{4, 8, 16\}$. Error bars: ± 1 s.d. over 50 trials. The 3% threshold (dashed grey) is reached at $s \ge 200$ on all datasets.

For each dataset we draw s source nodes uniformly at random, compute $\widehat{\mathcal{D}}_k$ from the centroid form, and report the relative error $|\widehat{\mathcal{D}}_k - \mathcal{D}_k|/\mathcal{D}_k$ averaged over 50 trials, with the exact $\mathcal{D}_k$ as ground

truth. We do this at three representative depths $k \in \{4, 8, 16\}$, early, intermediate, and late in the oversmoothing trajectory.

Three empirical observations align with the theory. First, the relative error decays as $s^{-1/2}$ across all datasets and depths, matching the McDiarmid scaling (Figure 2). Second, the 3% relative-error threshold is reached at $s \geq 50$ uniformly: on RomanEmpire ($n = 22{,}662$) and AmazonRatings ($n = 24{,}492$), even $s = 25$ suffices for $< 3\%$ error at every depth; on PubMed ($n = 19{,}717$), $s = 50$ brings the error below 3% at the most challenging depth ($k = 16$, relative error 2.6%). Third, deeper k requires marginally more samples, because $\mathcal{D}_k$ itself is smaller at large k and any fixed absolute error becomes a larger relative error, the absolute error is not depth-dependent, only the denominator is.

We adopt $s = 200$ as the default for $n \geq 20{,}000$ in all main-paper experiments. For ogbn-arxiv ($n = 169{,}343$), this represents a sampling rate of 0.12%, and the wall-clock cost is dominated by the eight sparse matrix–vector products per sampled source rather than by the centroid computation.

A note on the sn memory cost. The centroid formulation requires storing s walk distributions of length n simultaneously, resulting in a peak memory footprint of $O(sn)$. On `ogbn-arxiv`, with $s = 200$, this corresponds to approximately 135 MB in `float32`, which is well within the limits of commodity hardware.

For larger graphs, the walk distributions can be streamed: each $\phi_{v_i}^{(k)}$ contributes additively to $\widehat{\phi}^{(k)}$ and can be discarded once accumulated, reducing the peak memory to $O(n)$ without increasing the asymptotic time complexity. We did not require this optimization for any benchmark in our suite.

D.4 Runtime

All experiments were run on an Apple M2 Max with 64 GB RAM. Walk distributions are computed via repeated sparse matrix–vector products using `scipy.sparse`; the squared centroid norm is computed in a single dense reduction.

Table 8 reports wall-clock time for computing $\mathcal{D}_k$ across all eight depths $k \in \{2, 4, 6, 8, 10, 12, 16, 24\}$ on each benchmark. We compute exactly when $n \leq 20{,}000$ and switch to the sampling estimator with $s = 200$ source nodes otherwise.

Table 8: Runtime for computing $\mathcal{D}_k$ across all eight depths on a single CPU. Sampling is used for $n \geq 20{,}000$. Total time across the full 11-dataset suite is under 30 minutes.

Dataset	n	m	Mode	Time (all 8 depths)
Texas	183	295	exact	0.01 s
Wisconsin	251	466	exact	0.02 s
Cornell	183	277	exact	0.01 s
Cora	2,708	5,278	exact	3.30 s
CiteSeer	3,327	4,552	exact	1.90 s
Chameleon	2,277	31,371	exact	5.03 s
Squirrel	5,201	198,353	exact	57.16 s
PubMed	19,717	44,324	exact	959.27 s
AmazonRatings	24,492	93,050	sampled	2.20 s
RomanEmpire	22,662	32,927	sampled	0.11 s
ogbn-arxiv	169,343	1,166,243	sampled	128.81 s

For comparison, training a single GCN model on ogbn-arxiv at depth 24 takes approximately 566 seconds on CPU (37 epochs with early stopping), plus 4 seconds to extract hidden representations and compute Erank via SVD, a total of 570 seconds for one (model, seed, depth) configuration. Computing $\mathcal{D}_k$ across all eight depths on the same graph costs only 128.8 seconds with the sampled estimator at $s = 200$, already a $4.4\times$ speedup over a single training run, and well inside the regime in which Appendix D.3 shows the relative error falls below 3% uniformly across depths. On the medium-scale heterophilic and homophilic benchmarks the gap is even larger: $\mathcal{D}_k$ at $s = 200$ takes 4.78 s on PUBMED, 2.20 s on AMAZON-RATINGS, and 0.11 s on ROMAN-EMPIRE. A single trained

model is one of the $14 \times 10 \times 8 = 1{,}120$ configurations we evaluate per dataset to compute the post-training baselines.

E Generalization across architectures

The theoretical analysis in §5 uses the lazy random walk $\mathbf{P} = \frac{1}{2}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{A})$ as a parameter-free baseline. This appendix shows that the analysis transfers to the standard row-stochastic propagation operators such as GCN and GraphSAGE-mean up to architecture-dependent constants (§E.1); identifies sum aggregation as a regime where the simplex-geometric diagnostic does not directly apply (§E.2); proposes an augmented-scale variant $\mathcal{D}_k^{\mathrm{aug}}$ tailored to GIN and validates it empirically (§E.3); and reports an ablation over the choice of propagation operator $\mathbf{P}$ underlying the diagnostic itself (§E.4).

E.1 Row-stochastic operators: GCN, SAGE, GIN, APPNP, and attention

We extend the analysis of $\mathcal{D}_k$ from the lazy walk to the standard architectural propagation operators by way of a Dirichlet-form comparison. The argument is structured as follows: §E.1.1 states four admissibility conditions on a propagation operator $\mathbf{T}$ that suffice for $\mathcal{D}_k^{(\mathbf{T})}$ to track $\mathcal{D}_k^{(\mathrm{lazy})}$ at the level of exponential rates; §E.1.2 states and proves the universality theorem; §E.1.3 verifies the conditions for each architecture in our experimental suite.

E.1.1 Admissibility conditions

Let $\mathbf{T}$ be a row-stochastic matrix on V with stationary distribution $\tilde{\boldsymbol{\pi}}$, and write $Q_{vw}^{(\mathbf{T})} = \tilde{\pi}_v T_{vw}$ for its edge conductance on $(v, w) \in E$. The lazy walk has uniform conductance $Q_{vw}^{(\mathrm{lazy})} = 1/(2\mathrm{vol}(G))$ on every edge.

We assume the following.

- **(C1) Stochasticity, graph support, irreducibility, and reversibility.** $\mathbf{T}\mathbf{1} = \mathbf{1}$, $T_{vw} \geq 0$, and $T_{vw} > 0$ for $v \neq w$ only if $(v, w) \in E$. The chain is irreducible and reversible with respect to its stationary distribution $\tilde{\boldsymbol{\pi}}$, i.e. $\tilde{\pi}_v T_{vw} = \tilde{\pi}_w T_{wv}$ for all v, w.
- **(C2) Stationary distribution comparability.** There exists $\rho \geq 1$ such that $\rho^{-1}\pi_v \leq \tilde{\pi}_v \leq \rho\pi_v$ for all v.
- **(C3) Conductance comparability.** There exist $0 < \alpha \leq \beta < \infty$ such that $\alpha Q_{vw}^{(\mathrm{lazy})} \leq Q_{vw}^{(\mathbf{T})} \leq \beta Q_{vw}^{(\mathrm{lazy})}$ for every $(v, w) \in E$.
- **(C4) Non-negative spectrum.** $\lambda_{\min}(\mathbf{T}) \geq 0$.
- **(C5) Rate-comparison condition.** $\beta\rho^2\gamma_{\mathrm{lazy}} < 1$, where $\gamma_{\mathrm{lazy}} = 1 - \lambda_2(\mathbf{P}_{\mathrm{lazy}})$.

Conditions (C1)–(C4) are required for the qualitative claim that $\mathcal{D}_k^{(\mathbf{T})}$ and $\mathcal{D}_k^{(\mathrm{lazy})}$ vanish together. Condition (C5) is required for the explicit interval comparison of decay rates and is graph-dependent.

E.1.2 Universality theorem

Lemma E.1 (Spectral gap comparison)**.** *Under (C1)–(C3), with* $\gamma_{\mathbf{T}} = 1 - \lambda_2(\mathbf{T})$ *and* $\gamma_{\mathrm{lazy}} = 1 - \lambda_2(\mathbf{P}_{\mathrm{lazy}})$,

$$\frac{\alpha}{\rho^2}\gamma_{\mathrm{lazy}} \leq \gamma_{\mathbf{T}} \leq \beta\rho^2\gamma_{\mathrm{lazy}}.$$

Proof. The Dirichlet forms of the two reversible chains are

$$\mathcal{E}_{\mathbf{T}}(f, f) = \sum_{\{v,w\}\in E} Q_{vw}^{(\mathbf{T})}(f(v) - f(w))^2, \qquad \mathcal{E}_{\mathrm{lazy}}(f, f) = \sum_{\{v,w\}\in E} Q_{vw}^{(\mathrm{lazy})}(f(v) - f(w))^2.$$

By (C3), for all $f : V \to \mathbb{R}$,

$$\alpha\mathcal{E}_{\mathrm{lazy}}(f, f) \leq \mathcal{E}_{\mathbf{T}}(f, f) \leq \beta\mathcal{E}_{\mathrm{lazy}}(f, f). \tag{31}$$

The spectral gaps admit the variational characterization

$$\gamma_{\mathbf{T}} = \min_{\mathrm{Var}_{\tilde{\pi}}(f)>0} \frac{\mathcal{E}_{\mathbf{T}}(f,f)}{\mathrm{Var}_{\tilde{\pi}}(f)}, \qquad \gamma_{\mathrm{lazy}} = \min_{\mathrm{Var}_{\pi}(f)>0} \frac{\mathcal{E}_{\mathrm{lazy}}(f,f)}{\mathrm{Var}_{\pi}(f)}.$$

Using

$$\mathrm{Var}_{\mu}(f) = \frac{1}{2}\sum_{v,w} \mu_v \mu_w (f(v)-f(w))^2$$

and condition (C2), we obtain

$$\rho^{-2}\mathrm{Var}_{\pi}(f) \le \mathrm{Var}_{\tilde{\pi}}(f) \le \rho^2 \mathrm{Var}_{\pi}(f).$$

Combining this with (31) gives

$$\gamma_{\mathbf{T}} = \min_f \frac{\mathcal{E}_{\mathbf{T}}(f,f)}{\mathrm{Var}_{\tilde{\pi}}(f)} \ge \min_f \frac{\alpha \mathcal{E}_{\mathrm{lazy}}(f,f)}{\rho^2 \mathrm{Var}_{\pi}(f)} = \frac{\alpha}{\rho^2}\gamma_{\mathrm{lazy}}.$$

The upper bound is symmetric. □

Lemma E.2 (Single-chain decay rate)**.** *Let* $\mathbf{M}$ *be a reversible irreducible chain on* G *with stationary distribution* μ*, spectral gap* $\gamma = 1-\lambda_2(\mathbf{M})$*, and non-negative spectrum, and let* $\{f_i\}$ *be an* $L^2(\mu)$*-orthonormal eigenbasis with* $f_1 \equiv 1$*. Then*

$$c_{\mathbf{M}}(1-\gamma)^{2k} \le \mathcal{D}_k^{(\mathbf{M})} \le C_{\mathbf{M}}(1-\gamma)^{2k},$$

with $C_{\mathbf{M}} = \frac{2}{n}\sum_v(\mu_v^{-1}-1)$ *and* $c_{\mathbf{M}} = \frac{1}{4}\inf_v \mu_v \cdot \mathrm{Var}_{\mathrm{unif}}(f_2)$*. In particular, if* $0<\lambda_2(\mathbf{M})<1$*, then* $\mathcal{D}_k^{(\mathbf{M})} = \Theta((1-\gamma)^{2k})$.

Proof. We will prove the two sides of the inequality one by one.

Upper bound. By the Hellinger triangle inequality, for any reference distribution μ^*,

$$H^2(p_v^{(k)}, p_u^{(k)}) \le 2\big(H^2(p_v^{(k)},\mu^*) + H^2(p_u^{(k)},\mu^*)\big).$$

Choosing $\mu^* = \mu$ and averaging over all pairs (v,u) yields

$$\mathcal{D}_k^{(\mathbf{M})} = \frac{1}{n^2}\sum_{v,u} H^2(p_v^{(k)}, p_u^{(k)}) \le \frac{4}{n}\sum_v H^2(p_v^{(k)},\mu).$$

By the Hellinger-chi-squared inequality, $H^2(p,q) \le \frac{1}{2}\chi^2(p\|q)$. The spectral decomposition gives

$$\frac{p_v^{(k)}(w)}{\mu(w)} = \sum_{i=1}^n \lambda_i^k f_i(v) f_i(w),$$

from which

$$\chi^2(p_v^{(k)}\|\mu) = \sum_w \mu(w)\left(\frac{p_v^{(k)}(w)}{\mu(w)}-1\right)^2 = \sum_{i\ge 2}\lambda_i^{2k} f_i(v)^2 \le \lambda_2^{2k}\left(\frac{1}{\mu_v}-1\right).$$

Therefore

$$\mathcal{D}_k^{(\mathbf{M})} \le \frac{4}{n}\sum_v \frac{1}{2}\lambda_2^{2k}\left(\frac{1}{\mu_v}-1\right) = \frac{2}{n}\sum_v\left(\frac{1}{\mu_v}-1\right)(1-\gamma)^{2k}.$$

Lower bound. Write $p_v^{(k)}(w) = \mu(w)(1+g_v(w))$, where

$$g_v(w) = \sum_{i\ge 2}\lambda_i^k f_i(v) f_i(w).$$

Define $\psi_v(w) = \sqrt{p_v^{(k)}(w)/\mu(w)} = \sqrt{1+g_v(w)}$. Then

$$H^2(p_v^{(k)}, p_u^{(k)}) = \frac{1}{2}\sum_w \mu(w)\big(\psi_v(w)-\psi_u(w)\big)^2.$$

Since $\psi_v(w) \le 1/\sqrt{\mu(w)}$ and $\psi_u(w) \le 1/\sqrt{\mu(w)}$, the identity

$$a - b = \frac{a^2 - b^2}{a + b}$$

implies

$$\big(\psi_v(w) - \psi_u(w)\big)^2 = \frac{\big(g_v(w) - g_u(w)\big)^2}{\big(\psi_v(w) + \psi_u(w)\big)^2} \ge \frac{\mu(w)}{4}\big(g_v(w) - g_u(w)\big)^2.$$

Hence

$$H^2(p_v^{(k)}, p_u^{(k)}) \ge \frac{1}{8}\sum_w \mu(w)^2\big(g_v(w) - g_u(w)\big)^2 \ge \frac{\inf_w \mu(w)}{8}\sum_w \mu(w)\big(g_v(w) - g_u(w)\big)^2.$$

Projecting onto the f_2 direction and using $L^2(\mu)$-orthonormality,

$$\sum_w \mu(w)\big(\lambda_2^k(f_2(v) - f_2(u))f_2(w)\big)^2 = \lambda_2^{2k}(f_2(v) - f_2(u))^2.$$

Therefore

$$H^2(p_v^{(k)}, p_u^{(k)}) \ge \frac{\inf_w \mu(w)}{8}\lambda_2^{2k}(f_2(v) - f_2(u))^2.$$

Averaging over all pairs (v, u) and using

$$\frac{1}{n^2}\sum_{v,u}(f_2(v) - f_2(u))^2 = 2\,\mathrm{Var}_{\mathrm{unif}}(f_2)$$

gives

$$\mathcal{D}_k^{(\mathbf{M})} \ge \frac{\inf_v \mu_v}{4}\mathrm{Var}_{\mathrm{unif}}(f_2)(1 - \gamma)^{2k}.$$

Theorem E.3 (Lazy-walk universality for reversible operators)**.** *Let* $\mathbf{T}$ *satisfy (C1)–(C4).*

(i) Both dispersions vanish: $\mathcal{D}_k^{(\mathbf{T})} \to 0$ *if and only if* $\mathcal{D}_k^{(\mathrm{lazy})} \to 0$.

(ii) Under the additional condition (C5), both chains satisfy $\mathcal{D}_k = \Theta((1 - \gamma)^{2k})$ *with their respective gaps, and*

$$\frac{\log(1 - \alpha\rho^{-2}\gamma_{\mathrm{lazy}})}{\log(1 - \gamma_{\mathrm{lazy}})} \le \lim_{k\to\infty}\frac{\log \mathcal{D}_k^{(\mathbf{T})}}{\log \mathcal{D}_k^{(\mathrm{lazy})}} \le \frac{\log(1 - \beta\rho^2\gamma_{\mathrm{lazy}})}{\log(1 - \gamma_{\mathrm{lazy}})}.$$

Proof. Because $\mathbf{T}$ is finite, reversible, and irreducible, its stationary distribution is unique and strictly positive. By Lemma E.2 applied to $\mathbf{T}$ and to $\mathbf{P}_{\mathrm{lazy}}$, each chain has dispersion converging to zero, which proves part (i).

For part (ii), condition (C5) implies $\beta\rho^2\gamma_{\mathrm{lazy}} < 1$, hence by Lemma E.1

$$0 < \frac{\alpha}{\rho^2}\gamma_{\mathrm{lazy}} \le \gamma_{\mathbf{T}} \le \beta\rho^2\gamma_{\mathrm{lazy}} < 1.$$

Thus $0 < \gamma_{\mathbf{T}}, \gamma_{\mathrm{lazy}} < 1$, so the two-sided exponential estimates from Lemma E.2 apply to both chains. Taking logarithms and dividing by $2k$ gives

$$\lim_{k\to\infty}\frac{\log \mathcal{D}_k^{(\mathbf{M})}}{2k} = \log(1 - \gamma_{\mathbf{M}})$$

for $\mathbf{M} \in \{\mathbf{T}, \mathbf{P}_{\mathrm{lazy}}\}$. Therefore

$$\lim_{k\to\infty}\frac{\log \mathcal{D}_k^{(\mathbf{T})}}{\log \mathcal{D}_k^{(\mathrm{lazy})}} = \frac{\log(1 - \gamma_{\mathbf{T}})}{\log(1 - \gamma_{\mathrm{lazy}})}.$$

Substituting the gap bounds from Lemma E.1 and using that $x \mapsto \log(1 - x)$ is decreasing on $(0, 1)$ yields the stated interval. □

E.1.3 Architecture verification

Table 9 summarises the conditions for each architecture in our experimental suite. Conductances are compared against $Q_{vw}^{(\mathrm{lazy})} = 1/(2\mathrm{vol}(G))$.

Table 9: Verification of conditions (C1)–(C4) for some of the architectural propagation operators in our experiments. Self-loop notation: $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\hat{d}_v = d_v + 1$, $\hat{\pi}_v = \hat{d}_v/(\mathrm{vol}(G) + n)$. Condition (C5) depends on γ_{lazy} and is graph-dependent

Architecture	Effective $\mathbf{T}$	Reversible?	$\tilde{\boldsymbol{\pi}}$	$\alpha = \beta$	(C4)
SAGE-mean	$\frac{1}{2}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{A})$	Yes	$\boldsymbol{\pi}$	1	Yes
GCN (row-norm)	$\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}$	Yes	$\hat{\boldsymbol{\pi}}$	$\frac{2\mathrm{vol}(G)}{\mathrm{vol}(G)+n}$	Not auto.
APPNP (row-norm)	$\alpha_{\mathrm{tel}}\mathbf{I} + (1-\alpha_{\mathrm{tel}})\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}$	Yes	$\hat{\boldsymbol{\pi}}$	$\frac{2(1-\alpha_{\mathrm{tel}})\mathrm{vol}(G)}{\mathrm{vol}(G)+n}$	Yes if $\alpha_{\mathrm{tel}} \geq 1/2$

Derivations. In what follows, we derive the quantities listed in Table 9 for the three example architectures.

For GCN and APPNP, the self-loop offset between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ is bounded by

$$\rho_{\mathrm{sl}} = \max\left\{ \frac{\mathrm{vol}(G)}{\mathrm{vol}(G)+n}\left(1 + \frac{1}{d_{\min}}\right),\ \frac{\mathrm{vol}(G)+n}{\mathrm{vol}(G)} \cdot \frac{d_{\max}}{d_{\max}+1} \right\}.$$

SAGE-mean. Here $T_{vw} = 1/(2d_v)$ for $w \sim v$ and $T_{vv} = 1/2$. The chain is reversible with respect to the ordinary stationary distribution $\boldsymbol{\pi}$ and

$$Q_{vw}^{(\mathbf{T})} = \pi_v T_{vw} = \frac{d_v}{\mathrm{vol}(G)} \cdot \frac{1}{2d_v} = \frac{1}{2\ \mathrm{vol}(G)}.$$

Hence $\alpha = \beta = 1$ and $\rho = 1$. This architecture coincides exactly with the lazy random walk.

GCN (row-normalised). For every $w \sim v$ or $w = v$, $T_{vw} = 1/\hat{d}_v$. The chain is reversible with respect to $\hat{\boldsymbol{\pi}}$ (the self-loop corrected distribution) and

$$Q_{vw}^{(\mathrm{GCN})} = \hat{\pi}_v T_{vw} = \frac{1}{\mathrm{vol}(G)+n}$$

for every edge $(v, w) \in E$. Since $Q_{vw}^{(\mathrm{lazy})} = 1/(2\,\mathrm{vol}(G))$, we obtain

$$\alpha = \beta = \frac{2\,\mathrm{vol}(G)}{\mathrm{vol}(G)+n}.$$

Condition (C4) (non-negative spectrum) is not automatic. For the three-vertex path one has, for instance,

$$\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix},$$

whose eigenvalues are 1, $1/2$ and $-1/6$.

APPNP (row-normalised). With $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and teleport probability α_t, the operator is

$$\mathbf{T} = \alpha_t \mathbf{I} + (1-\alpha_t)\hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}.$$

It is reversible with respect to $\hat{\boldsymbol{\pi}}$ and, for $(v, w) \in E$,

$$Q_{vw}^{(\mathbf{T})} = \hat{\pi}_v \frac{1-\alpha_t}{\hat{d}_v} = \frac{1-\alpha_t}{\mathrm{vol}(G)+n}.$$

Hence

$$\alpha = \beta = \frac{2(1-\alpha_t)\ \mathrm{vol}(G)}{\mathrm{vol}(G)+n}.$$

Writing $\mathbf{K} = \hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}$, every eigenvalue of $\mathbf{K}$ lies in $[-1, 1]$. Consequently, the eigenvalues of $\mathbf{T} = \alpha_t \mathbf{I} + (1 - \alpha_t)\mathbf{K}$ lie in

$$[2\alpha_t - 1,\ 1].$$

Thus (C4) holds whenever $\alpha_t \geq 1/2$; for smaller α_t it may fail on certain graphs.

GAT. Learnt attention matrices are typically non-reversible, so the universality theorem does not apply without extra assumptions. Empirically, however, $\mathcal{D}_k$ computed from the lazy walk correlates strongly with GAT accuracy degradation. We interpret this as evidence that learned attention coefficients tend to track the underlying graph topology even though no constraint forces them to, so the lazy-walk diagnostic predicts GAT's depth dynamics in practice without theoretical guarantee.

Remark E.4 (A correct sufficient criterion for (C4))**.** *If $\mathbf{T}$ is reversible and $T_{vv} \geq 1/2$ for every v, then (C4) holds. Indeed, $\mathbf{S} := 2\mathbf{T} - \mathbf{I}$ is again reversible and row-stochastic, so every eigenvalue of $\mathbf{S}$ lies in $[-1, 1]$. Therefore every eigenvalue λ of $\mathbf{T}$ satisfies $2\lambda - 1 \in [-1, 1]$, hence $\lambda \in [0, 1]$.*

This criterion applies immediately to SAGE-mean, and to APPNP whenever $\alpha_t \geq 1/2$. It does not apply to GCN in general.

Remark E.5 (Condition (C5) must be checked separately)**.** *The table verifies only (C1)–(C4). The explicit logarithmic rate interval in Theorem E.3 also requires (C5), namely $\beta\rho^2\gamma_{\mathrm{lazy}} < 1$. This is a graph-architecture pair condition and can fail even when (C1)–(C4) hold. For example, for GCN on the complete graph K_n, one has $\rho = 1$, $\beta = 2(n-1)/n$, and $\gamma_{\mathrm{lazy}} = n/(2(n-1))$, so $\beta\rho^2\gamma_{\mathrm{lazy}} = 1$.*

Remark E.6 (Role of ρ)**.** *On bounded-degree graph families, $\rho = O(1)$ for all architectures above whose stationary laws are comparable to degree. On highly heterogeneous graphs, ρ can be large, reflecting a genuine difference between the architecture's weighting and the lazy walk's weighting.*

E.2 Sum aggregation: why GIN falls outside the framework

GIN [Xu et al., 2019] departs from the operators above in a way that is fundamental to its design rather than incidental. Its update rule is

$$\mathbf{H}^{(\ell+1)} = \mathrm{MLP}^{(\ell)}\Big((1+\varepsilon)\mathbf{H}^{(\ell)} + \mathbf{A}\mathbf{H}^{(\ell)}\Big), \tag{32}$$

where the aggregation operator $(1+\varepsilon)\mathbf{I} + \mathbf{A}$ uses a *sum* over neighbours rather than a mean. This choice is the central motivation of GIN: Xu et al. [2019] prove that mean and max aggregators cannot distinguish certain non-isomorphic neighbourhoods that the sum aggregator can, and that with an injective MLP the resulting GNN matches the discriminative power of the Weisfeiler–Lehman test. The price of this expressivity is that the aggregation operator is no longer row-stochastic.

This has three consequences for the theory developed in §5.

The rows of $\mathbf{P}^k_{\mathrm{GIN}}$ are not probability distributions. The square-root embedding ϕ is defined on the simplex Δ^{n-1}, and the diagnostic $\mathcal{D}_k$ measures the variance of the embedded point cloud on S^{n-1}_+. Without row-stochasticity, the rows of $\mathbf{P}^k_{\mathrm{GIN}} = ((1+\varepsilon)\mathbf{I} + \mathbf{A})^k$ have ℓ_1 norms that grow polynomially or exponentially in k (depending on ε and the degree distribution), so they do not lie on the simplex at all. Renormalising each row to unit ℓ_1 norm would produce a row-stochastic surrogate, but this surrogate is no longer the operator GIN uses, it discards exactly the magnitude information that the sum aggregator was designed to preserve.

E.3 An augmented-scale diagnostic for sum aggregation

Sum aggregation preserves count information that mean aggregation discards, and any diagnostic for GIN must therefore track both the angular structure of the propagated rows and their magnitudes. We construct $\mathcal{D}^{\mathrm{aug}}_k$, an augmented-scale variant of the aggregation dispersion designed for non-row-stochastic operators, by lifting each row of the propagation matrix into a higher-dimensional simplex with an explicit mass coordinate.

Definition. Let $\mathbf{M}_\varepsilon = (1+\varepsilon)\mathbf{I} + \mathbf{A}$ denote the GIN aggregation operator and $\lambda_1 = \lambda_1(\mathbf{M}_\varepsilon)$ its leading eigenvalue. For each node v, let $s_v^{(k)} = \|\mathbf{M}_\varepsilon^k e_v\|_1$ denote the ℓ_1 mass of the v-th column at

depth k, and define the rescaled walk vector

$$\tilde{p}_v^{(k)} = \frac{\mathbf{M}_\varepsilon^k e_v}{\lambda_1^k} \in \mathbb{R}^n_{\geq 0}. \tag{33}$$

The rescaling by λ_1^k stabilises the magnitudes against the exponential growth induced by the unbounded spectrum of $\mathbf{M}_\varepsilon$ (§E.2), placing the rescaled vectors in a bounded regime regardless of depth. We then form the augmented walk

$$p_v^{(k),\mathrm{aug}} = \frac{1}{s_v^{(k)}/\lambda_1^k + 1} \begin{pmatrix} \tilde{p}_v^{(k)} \\ 1 \end{pmatrix} \in \Delta^n, \tag{34}$$

which appends a mass coordinate equal to 1 before normalization, then divides by the total to land on the n-dimensional simplex Δ^n. Two nodes with identical angular structure but different total mass produce walks that differ in the appended coordinate; two nodes with identical mass but different angular structure produce walks that differ in the first n coordinates. The augmented dispersion is then the standard construction applied to these augmented walks:

$$\mathcal{D}_k^{\mathrm{aug}} = 1 - \left\| \frac{1}{n} \sum_v \sqrt{p_v^{(k),\mathrm{aug}}} \right\|_2^2. \tag{35}$$

Properties. The augmented walk $p_v^{(k),\mathrm{aug}}$ lies on the standard simplex Δ^n by construction, so $\mathcal{D}_k^{\mathrm{aug}}$ inherits the main theoretical properties of $\mathcal{D}_k$, except for monotonicity:

Collapse characterization. $\mathcal{D}_k^{\mathrm{aug}} = 0$ if and only if all augmented embeddings coincide, which by (34) requires both that all nodes have identical row directions *and* identical total mass. This is strictly stronger than the standard $\mathcal{D}_k = 0$ condition, which only constrains directions.

Closed-form centroid. The reformulation (35) retains the $O(n^2)$ centroid evaluation of the original $\mathcal{D}_k$, since the augmented embedding adds only a single coordinate per node.

Concentration. The McDiarmid argument of Appendix D.2 transfers directly: each $p_v^{(k),\mathrm{aug}}$ has unit ℓ_1 norm, so the bounded-differences constant remains $4/(s-1)$ and the sample complexity bound $s = O(\varepsilon^{-2}\log(1/\delta))$ is unchanged.

Quasi-monotonicity. The augmented walk evolves under a state-dependent projective rescaling rather than a fixed Markov kernel, so the data-processing inequality does not apply directly. We instead prove a relaxed monotonicity in which $\mathcal{D}_k^{\mathrm{aug}}$ is non-increasing up to a correction that decays geometrically at the mixing rate of $\mathbf{M}_\varepsilon$.

Proposition E.7 (Quasi-monotonicity of $\mathcal{D}_k^{\mathrm{aug}}$)**.** *Let $\mathbf{M}_\varepsilon = (1+\varepsilon)\mathbf{I} + \mathbf{A}$ be the GIN aggregation operator on a connected undirected graph with $\varepsilon \geq 0$, let $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$ denote its eigenvalues, and set*

$$\gamma \;:=\; \frac{|\lambda_2|}{\lambda_1} \;\in\; [0,1).$$

Then there exists a constant $C > 0$ depending only on the graph such that

$$\mathcal{D}_{k+1}^{\mathrm{aug}} \;\leq\; \mathcal{D}_k^{\mathrm{aug}} \;+\; C\,\gamma^{k/2} \qquad \text{for all } k \geq 0. \tag{36}$$

Moreover, there exist $k_0 \in \mathbb{N}$ and $C' > 0$ such that the rate sharpens to $\mathcal{D}_{k+1}^{\mathrm{aug}} \leq \mathcal{D}_k^{\mathrm{aug}} + C'\gamma^k$ for all $k \geq k_0$.

Proof. Write $w_v := (\mathbf{v}_1)_v$ for the entries of the unit-ℓ_2 Perron eigenvector and let $K := \mathbf{M}_\varepsilon/\lambda_1$. Connectedness and $\varepsilon \geq 0$ imply $w_v > 0$ for every v and $\gamma = |\lambda_2|/\lambda_1 < 1$ (Perron–Frobenius).

Symmetry of $\mathbf{M}_\varepsilon$ gives the orthonormal eigendecomposition

$$K^k e_v \;=\; w_v \mathbf{v}_1 + \mathbf{r}_v^{(k)}, \qquad \|\mathbf{r}_v^{(k)}\|_2 \;\leq\; \gamma^{\,k},$$

and consequently $\|\mathbf{r}_v^{(k)}\|_1 \leq \sqrt{n}\,\gamma^{\,k}$ by Cauchy–Schwarz. Set $\tilde{p}_v^{(k)} := K^k e_v$; the total mass

$$m_v^{(k)} := \|\tilde{p}_v^{(k)}\|_1 \;=\; m_v^{(\infty)} + e_k, \qquad m_v^{(\infty)} := w_v \|\mathbf{v}_1\|_1 > 0,$$

where $|e_k| \le \|\mathbf{r}_v^{(k)}\|_1 \le \sqrt{n}\,\gamma^{\,k}$. The augmented distribution is

$$p_v^{(k),\mathrm{aug}} \;=\; \frac{1}{m_v^{(k)}+1}\begin{pmatrix}\tilde{p}_v^{(k)}\\ 1\end{pmatrix}, \qquad p_v^{(\infty),\mathrm{aug}} \;=\; \frac{1}{m_v^{(\infty)}+1}\begin{pmatrix}\tilde{p}_v^{(\infty)}\\ 1\end{pmatrix},$$

with $\tilde{p}_v^{(\infty)} := w_v \mathbf{v}_1$ and $m_v^{(\infty)} = \|\tilde{p}_v^{(\infty)}\|_1$.

We bound the ℓ_1 distance between these two $(n+1)$-vectors. For the first n coordinates:

$$\Big\|\frac{\tilde{p}_v^{(k)}}{m_v^{(k)}+1} - \frac{\tilde{p}_v^{(\infty)}}{m_v^{(\infty)}+1}\Big\|_1 \;\le\; \Big|\frac{1}{m_v^{(k)}+1} - \frac{1}{m_v^{(\infty)}+1}\Big|\,\|\tilde{p}_v^{(k)}\|_1 \;+\; \frac{\|\tilde{p}_v^{(k)} - \tilde{p}_v^{(\infty)}\|_1}{m_v^{(\infty)}+1}.$$

The scalar inequality $\big|(a+1)^{-1} - (b+1)^{-1}\big| = |a-b|/((a+1)(b+1)) \le |a-b|$ for $a, b \ge 0$ yields

$$\Big|\frac{1}{m_v^{(k)}+1} - \frac{1}{m_v^{(\infty)}+1}\Big| \;\le\; |e_k| \;\le\; \sqrt{n}\,\gamma^{\,k}.$$

Because $\tilde{p}_v^{(k)}$ has non-negative entries, $\|\tilde{p}_v^{(k)}\|_1 = m_v^{(k)}$ is uniformly bounded by some constant M independent of v and k (e.g. $M = \max_{v,k} m_v^{(k)} < \infty$, which exists because $m_v^{(k)} \to m_v^{(\infty)}$ uniformly in v). Hence the first part is at most $M\sqrt{n}\,\gamma^{\,k}$.

The second part is simply $\|\mathbf{r}_v^{(k)}\|_1/(m_v^{(\infty)}+1) \le \sqrt{n}\,(m_v^{(\infty)}+1)^{-1}\gamma^{\,k}$.

For the appended coordinate,

$$\Big|\frac{1}{m_v^{(k)}+1} - \frac{1}{m_v^{(\infty)}+1}\Big| \;\le\; \sqrt{n}\,\gamma^{\,k}.$$

Adding the three contributions we obtain

$$\big\|p_v^{(k),\mathrm{aug}} - p_v^{(\infty),\mathrm{aug}}\big\|_1 \;\le\; C_1\gamma^{\,k},$$

with the graph-dependent constant

$$C_1 \;:=\; \sqrt{n}\Big(M + \frac{1}{\min_u(m_u^{(\infty)}+1)} + 1\Big).$$

For any $p, q \in \Delta^n$, the elementary inequality $(\sqrt{a}-\sqrt{b})^2 \le |a-b|$ summed over coordinates gives $\|\sqrt{p} - \sqrt{q}\|_2 \le \sqrt{\|p-q\|_1}$. Together with Cauchy–Schwarz,

$$|H(p,q) - H(p',q')| \;=\; \big|\langle\sqrt{p},\sqrt{q}\rangle - \langle\sqrt{p'},\sqrt{q'}\rangle\big| \;\le\; \sqrt{\|p-p'\|_1} + \sqrt{\|q-q'\|_1}.$$

Applying this to the definition $\mathcal{D}_k^{\mathrm{aug}} = 1 - \frac{1}{n^2}\sum_{u,v} H(p_u^{(k),\mathrm{aug}}, p_v^{(k),\mathrm{aug}})$ and inserting the bound from Step 1,

$$\big|\mathcal{D}_k^{\mathrm{aug}} - \mathcal{D}_\infty^{\mathrm{aug}}\big| \;\le\; 2\sqrt{C_1}\,\gamma^{\,k/2},$$

where $\mathcal{D}_\infty^{\mathrm{aug}} := 1 - \big\|\frac{1}{n}\sum_v \sqrt{p_v^{(\infty),\mathrm{aug}}}\big\|_2^2$.

Step 3 (triangle inequality). From Step 2,

$$\begin{aligned}\mathcal{D}_{k+1}^{\mathrm{aug}} - \mathcal{D}_k^{\mathrm{aug}} \;&=\; \big(\mathcal{D}_{k+1}^{\mathrm{aug}} - \mathcal{D}_\infty^{\mathrm{aug}}\big) - \big(\mathcal{D}_k^{\mathrm{aug}} - \mathcal{D}_\infty^{\mathrm{aug}}\big)\\ &\le\; 4\sqrt{C_1}\,\gamma^{\,k/2},\end{aligned}$$

which gives $\mathcal{D}_{k+1}^{\mathrm{aug}} \le \mathcal{D}_k^{\mathrm{aug}} + C\,\gamma^{\,k/2}$ with $C := 4\sqrt{C_1}$. This establishes (36).

The factor $\frac{1}{2}$ in the exponent of Step 2 originates solely from the non-Lipschitz behaviour of $x \mapsto \sqrt{x}$ near zero. Because $w_v > 0$ for every vertex, the limit vector $p_v^{(\infty),\mathrm{aug}}$ has all coordinates bounded

below by some $\delta_\star > 0$. By Step 1 there exists k_0 such that for all $k \geq k_0$ every coordinate of $p_v^{(k),\mathrm{aug}}$ exceeds $\delta_\star/2$. On the interval $[\delta_\star/2,\, 1]$ the square root is $(2\sqrt{\delta_\star/2})^{-1}$-Lipschitz, so the bound in Step 2 sharpens to

$$\left|\mathcal{D}_k^{\mathrm{aug}} - \mathcal{D}_\infty^{\mathrm{aug}}\right| \;\leq\; C''\,\gamma^{\,k}, \qquad k \geq k_0,$$

for a suitable C''. Repeating the triangle inequality of the previous step gives $\mathcal{D}_{k+1}^{\mathrm{aug}} \leq \mathcal{D}_k^{\mathrm{aug}} + C'\,\gamma^{\,k}$ with $C' = 2C''$ for all $k \geq k_0$. □

Interpretation. Proposition E.7 delivers the qualitative content of monotonicity in the regime relevant to oversmoothing diagnostics. The slack term $C\gamma^{k/2}$ is governed by the spectral gap of $\mathbf{M}_\varepsilon$, which is the same quantity that controls convergence to the Perron eigenspace and hence the onset of oversmoothing itself: any transient non-monotonicity of $\mathcal{D}_k^{\mathrm{aug}}$ is bounded by exactly the timescale on which the diagnostic is read. Numerically, on graphs with $\gamma \leq 0.7$ the slack is below $5{\cdot}10^{-2}$ at $k = 8$ and below $4{\cdot}10^{-3}$ at $k = 16$, well within the depths used in our experiments. The collapse characterization, closed-form centroid, and McDiarmid concentration arguments transfer verbatim, since none of them depended on exact monotonicity.

The construction therefore delivers an architecture-matched diagnostic for GIN with similar theoretical guarantees as the lazy-walk version. The empirical validation in §F.2 shows that this translates to substantial gains on the regime it was designed for.

Empirical validation on GIN. We evaluate $\mathcal{D}_k^{\mathrm{aug}}$ on the GIN configurations of §F.2. The result, summarised in Table 10, is a $+0.158$ improvement over standard $\mathcal{D}_k$ and a $+0.033$ improvement over the strongest post-training baseline (Erank). The full per-ε breakdown appears in Table 16 (§F.2).

Table 10: $\mathcal{D}_k^{\mathrm{aug}}$ vs. standard $\mathcal{D}_k$ vs. Erank on GIN. Mean absolute correlation $|r|$ between $\log(\mathrm{metric})$ and test accuracy, averaged over 10 datasets (excluding ogbn-arxiv) and the eight depths, broken out by GIN ε variant.

Metric	$\varepsilon = 0$	$\varepsilon = 0.5$	$\varepsilon = 5$	Overall
$\mathcal{D}_k^{\mathrm{aug}}$ (scale-augmented)	**0.710**	**0.747**	**0.751**	**0.736**
Erank	<u>0.681</u>	<u>0.704</u>	<u>0.724</u>	<u>0.703</u>
$\mathcal{D}_k$ (standard)	0.607	0.554	0.573	0.578

The augmented variant wins on every ε setting and Overall, with the margin over Erank widening as ε decreases (the regime where the aggregation operator deviates most from the diagonal-dominated $(1+\varepsilon)\mathbf{I}$ limit and topological mixing matters most). The $+0.158$ improvement over standard $\mathcal{D}_k$ confirms the diagnosis of §E.2: the missing signal in the lazy-walk diagnostic on GIN is magnitude information, and recovering it requires modifying the embedding rather than abandoning the geometric framework.

E.4 Ablation: choice of propagation operator

Theorem E.3 establishes that the asymptotic decay rate of $\mathcal{D}_k$ is governed by γ regardless of which row-stochastic operator is used to compute it. We now test whether this theoretical prediction translates to empirical interchangeability: does the choice of $\mathbf{P}$ underlying $\mathcal{D}_k$ affect its predictive correlation with accuracy degradation? We sweep two axes that vary the operator while preserving its row-stochasticity and reversibility.

Axis 1: laziness coefficient. We sweep the self-loop weight in $\mathbf{P}_\alpha = \alpha\mathbf{I} + (1-\alpha)\mathbf{D}^{-1}\mathbf{A}$ for $\alpha \in \{0, 0.25, 0.5, 0.75\}$. The canonical lazy walk corresponds to $\alpha = 0.5$ and is justified by aperiodicity (which guarantees convergence to $\boldsymbol{\pi}$ from every initial distribution); the case $\alpha = 0$ corresponds to the bare row-normalised augmented adjacency $\mathbf{D}^{-1}\mathbf{A}$ used in GraphSAGE-mean.

Table 11: Laziness coefficient α in $\mathbf{P}_\alpha = \alpha\mathbf{I} + (1-\alpha)\mathbf{D}^{-1}\mathbf{A}$. Mean $|r|$ across 14 model configurations and the eight depths. The trend is monotonic in α across both regimes: smaller laziness gives a more depth-resolved diagnostic, with $\alpha = 0$ leading the canonical $\alpha = 0.5$ by $+0.013$ Overall. Bold marks the per-row best.

α	0	0.25	0.5 (canonical)	0.75
Homophilic	**0.788**	0.780	0.763	0.732
Heterophilic	**0.702**	0.702	0.697	0.687
Overall	**0.734**	0.730	0.721	0.703

The per-dataset breakdown (Table 12) shows that the monotone α-dependence holds with remarkable consistency: 7 of 11 datasets are best at $\alpha = 0$, the per-dataset range is at most 0.093 (ogbn-arxiv) and typically below 0.05, and Chameleon is the sole dataset where $\alpha = 0.5$ (lazy walk) wins.

Table 12: Per-dataset breakdown of the laziness sweep. Each cell is the mean $|r|$ averaged over 14 configurations and the eight depths. The Range column reports the gap between best and worst α.

Dataset	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	Range
Homophilic					
CiteSeer	**0.859**	0.853	0.843	0.826	0.033
Cora	**0.773**	0.763	0.745	0.718	0.055
PubMed	**0.703**	0.696	0.681	0.657	0.046
ogbn-arxiv	**0.819**	0.808	0.784	0.726	0.093
Heterophilic					
AmazonRatings	**0.759**	0.751	0.742	0.737	0.022
Chameleon	0.872	0.874	**0.877**	0.868	0.010
Cornell	0.533	**0.543**	0.542	0.540	0.010
RomanEmpire	**0.840**	0.833	0.828	0.819	0.021
Squirrel	**0.732**	0.730	0.723	0.698	0.034
Texas	0.528	**0.530**	0.521	0.512	0.018
Wisconsin	0.652	**0.655**	0.649	0.636	0.020

The mechanism is straightforward: as α increases, the lazy diagonal $\alpha\mathbf{I}$ dilutes the off-diagonal aggregation, slowing mixing in absolute time but compressing the depth range over which the diagnostic varies meaningfully. At $\alpha = 0.5$, half of every step is a self-loop, so eight effective propagation steps require sixteen layers; at $\alpha = 0$, every step is a full propagation. Within the depth budget $k \in \{2, \ldots, 24\}$ used in our experiments, smaller α therefore gives a more depth-resolved trajectory $k \mapsto \mathcal{D}_k$ and a higher correlation with depth-induced accuracy degradation.

Axis 2: symmetric vs. row-stochastic. Theorem E.3 predicts that the row-stochastic operator $\mathbf{T} = \frac{1}{2}(\mathbf{I}+\mathbf{D}^{-1}\mathbf{A})$ and the symmetric operator $\mathbf{S} = \frac{1}{2}(\mathbf{I}+\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2})$ should yield indistinguishable $\mathcal{D}_k$ trajectories up to similarity, because they share spectra and differ only by a diagonal conjugation that does not affect the second eigenvalue. We test this prediction directly.

A technical note: the centroid computation of $\mathcal{D}_k$ (Equation 4, Lemma 4.2) requires each propagated node embedding to be a probability vector (rows sum to 1). Since $\mathbf{S}$ is not row-stochastic, we first row-normalize its outputs: for each node v, we replace the row $[\mathbf{S}\mathbf{X}]_{v,:}$ by $[\mathbf{S}\mathbf{X}]_{v,:} / \sum_j [\mathbf{S}\mathbf{X}]_{v,j}$ before applying the centroid formula. This normalization preserves the Hellinger distances because it corresponds to a diagonal scaling that commutes with the similarity transformation linking $\mathbf{T}$ and $\mathbf{S}$; as a result the comparison remains faithful to the theory.

Table 13: Symmetric vs. row-stochastic propagation at $\alpha = 0.5$. The two operators agree to within 0.003 Overall and 0.005 on every individual dataset, confirming the empirical interchangeability predicted by Theorem E.3.

Dataset regime	Row-stochastic $\mathbf{T}$	Symmetric $\mathbf{S}$
Homophilic	**0.763**	0.758
Heterophilic	**0.697**	0.696
Overall	**0.721**	0.718

Row-stochastic edges out symmetric on 9 of 11 datasets, but the gaps are uniformly below 0.01 and the Overall difference of 0.003 is well inside the ± 0.035 standard error reported in Table 13. The two operators are empirically interchangeable, confirming that the simplex-geometric content of $\mathcal{D}_k$ is invariant to whether the propagation is normalized symmetrically or row-wise. We use the row-stochastic form throughout the main paper because it admits the direct probabilistic interpretation of walk distributions on Δ^{n-1} that motivates the construction.

Implications and choice for the main paper. Two findings emerge. First, the diagnostic is robust to the choice of row-stochastic operator: across the laziness sweep ($\alpha = 0$ to $\alpha = 0.75$, a range of 0.031 Overall) and the symmetric/row-stochastic comparison (0.003 Overall), no single choice is statistically dominant, and all variants beat every post-training baseline. Second, smaller laziness gives a marginally better depth-resolved diagnostic, with $\alpha = 0$ leading $\alpha = 0.5$ by $+0.013$ Overall. We retain $\alpha = 0.5$ as the canonical choice in the main paper for two reasons: it is the operator for which the mixing-time bound of Levin et al. [2017] (used in the proof of Theorem 5.3(i)) is stated cleanest, and it is the standard convention in the spectral graph theory literature. The ablation confirms that practitioners can choose any $\alpha \in [0, 0.5]$ without materially affecting predictive performance, with $\alpha = 0$ offering a slight gain in depth resolution at no theoretical cost (aperiodicity is not needed for $\mathcal{D}_k$ since the rows of $\tilde{\mathbf{A}}$ already include self-loops).

F Extended empirical results

This section reports the full set of empirical results supporting the headline numbers in §6. We provide per-activation breakdown across architectures (§F.1), GIN-specific results highlighting the framework boundary (§F.2), evaluation of $\mathcal{D}_k$ on oversmoothing mitigation techniques (§F.3), robustness to the choice of correlation metric (§F.4), per-dataset accuracy, metric – depth plots (§F.5), and finally we verify whether our proposed metric provide information beyond depth information (§F.6).

F.1 Per-activation breakdown

Tables 14 and 15 report the mean absolute correlation per (architecture, activation) cell, averaged over the 11 datasets and the eight depths.

Table 14: LeakyReLU: Mean absolute correlation $|r|$ between each diagnostic and GNN accuracy, averaged over datasets and depths. $\mathcal{D}_k$ ties EProj,n at the activation subtotal (both 0.696). Bold = best in row; underline = second best.

Architecture	$\mathcal{D}_k$	Erank	NumRank	EDir	EDir,n	EProj	EProj,n	MAD
GCN	<u>0.756</u>	0.657	0.653	0.559	**0.808**	0.543	0.560	0.614
GATv1	<u>0.764</u>	0.653	0.508	0.700	**0.773**	0.729	0.751	0.547
GATv2	0.784	0.743	0.611	0.745	**0.844**	0.769	<u>0.816</u>	0.619
SAGE	0.777	**0.868**	0.687	0.735	0.702	0.762	<u>0.807</u>	0.712
ChebNet	<u>0.781</u>	**0.849**	0.752	0.666	0.544	0.704	0.715	0.661
JKNet	0.407	0.397	0.445	0.340	**0.499**	0.444	<u>0.468</u>	0.415
APPNP	0.604	0.501	0.620	<u>0.645</u>	0.556	0.623	**0.754**	0.619
LReLU avg.	**0.696**	0.667	0.611	0.627	<u>0.675</u>	0.653	**0.696**	0.598

Table 15: Tanh: Mean absolute correlation $|r|$ between each diagnostic and GNN accuracy, averaged over datasets and depths. $\mathcal{D}_k$ leads the activation subtotal (0.747 vs. Erank's 0.709), winning outright on four of seven architectures (GCN, GATv1, GATv2, ChebNet) and placing top-2 on six. Bold = best in row; underline = second best.

Architecture	$\mathcal{D}_k$	Erank	NumRank	EDir	EDir,n	EProj	EProj,n	MAD
GCN	**0.830**	0.753	0.689	0.548	0.631	0.449	0.529	0.554
GATv1	**0.805**	0.722	0.658	0.565	0.657	0.488	0.648	0.698
GATv2	**0.813**	0.785	0.726	0.573	0.711	0.593	0.753	0.722
SAGE	0.847	**0.901**	0.721	0.743	0.802	0.705	0.812	0.810
ChebNet	**0.830**	0.720	0.729	0.717	0.403	0.590	0.579	0.470
JKNet	0.600	0.582	**0.686**	0.608	0.622	0.532	0.579	0.645
APPNP	0.501	0.497	0.575	0.661	0.506	**0.670**	0.661	0.552
Tanh avg.	**0.747**	0.709	0.683	0.631	0.619	0.575	0.652	0.636
Overall	**0.721**	0.688	0.647	0.629	0.647	0.614	0.674	0.617

The activation breakdown shows that $\mathcal{D}_k$ achieves its strongest advantage under Tanh, where it attains the highest activation-level average (0.747), compared to Erank (0.709). In this regime, it is the best-performing diagnostic on five of seven architectures (GCN, GATv1, GATv2, ChebNet, and tied/close on SAGE).

Under LeakyReLU, the differences are smaller and performance is more uniform across diagnostics. At the activation-level average, $\mathcal{D}_k$ and EProj,n are tied (both 0.696).

Across both activation regimes, performance is less stable on JKNet and APPNP, where all diagnostics degrade substantially, consistent with the architectural sensitivity discussed in §6.3.

F.2 GIN: validating the framework boundary and its architecture-matched fix

GIN's sum aggregation produces a non-row-stochastic propagation operator (Appendix E.2), which places it outside the simplex-geometric framework underlying the lazy-walk $\mathcal{D}_k$. We confirm this empirically and then evaluate the architecture-matched diagnostic $\mathcal{D}_k^{\mathrm{aug}}$ developed in Appendix E.3.

Table 16: Mean absolute correlation $|r|$ between $\log(\mathrm{diagnostic})$ and test accuracy, averaged over 10 datasets (excluding ogbn-arxiv) and the eight depths, for each GIN ε variant. The architecture-matched $\mathcal{D}_k^{\mathrm{aug}}$ wins on every ε setting and Overall, with $+0.158$ over standard $\mathcal{D}_k$ and $+0.033$ over the strongest post-training baseline (Erank). Standard $\mathcal{D}_k$ underperforms because the lazy walk is the wrong propagation operator for GIN, not because the geometric framework fails. Bold marks the per-column best and underline the second-best.

GIN variant	$\mathcal{D}_k^{\mathrm{aug}}$	$\mathcal{D}_k$	Erank	NumRank	EDir	EDir,n	EProj	EProj,n	MAD
$\varepsilon = 0$	**0.710**	0.607	0.681	0.635	0.641	0.623	0.584	0.614	0.571
$\varepsilon = 0.5$	**0.747**	0.554	0.704	0.670	0.622	0.623	0.671	0.635	0.663
$\varepsilon = 5$	**0.751**	0.573	0.724	0.666	0.609	0.677	0.598	0.618	0.658
Overall	**0.736**	0.578	0.703	0.657	0.624	0.641	0.617	0.623	0.631

Two readings of the table support a single conclusion. The first compares the standard lazy-walk $\mathcal{D}_k$ (0.578 Overall) against Erank (0.703): the geometric diagnostic underperforms the strongest post-training baseline by -0.125, exactly as predicted by Appendix E.2's argument that GIN's non-row-stochastic propagation falls outside the simplex framework. The second compares the architecture-matched $\mathcal{D}_k^{\mathrm{aug}}$ (0.736 Overall) against the same baselines: the augmented diagnostic wins on every ε setting, beats Erank by $+0.033$, and improves on standard $\mathcal{D}_k$ by $+0.158$. The diagnostic does not need to be abandoned for sum aggregation; it needs to embed the operator GIN actually uses.

The gap between $\mathcal{D}_k^{\mathrm{aug}}$ and Erank is non-monotone in ε ($+0.029$, $+0.043$, $+0.027$ at $\varepsilon = 0, 0.5, 5$), and we do not read a mechanism into the precise shape: the gaps are small relative to cross-dataset variability, and three settings are too few to fit a curve. What is robust is the sign. $\mathcal{D}_k^{\mathrm{aug}}$ wins at every ε tested and on the Overall aggregate. The framework boundary is set by the propagation operator, not by sum aggregation per se, and once GIN's actual operator is embedded into the diagnostic the geometric framework recovers and surpasses the strongest post-training baseline.

F.3 Evaluating oversmoothing mitigation techniques

Table 17 reports, for each anti-smoothing variant, the mean absolute correlation between each diagnostic's depth profile and the corresponding depth-accuracy curve, averaged over all 11 datasets and both activations. We stress that this is a *within-variant* correlation. $\mathcal{D}_k$ is defined from the lazy random walk on the test-time graph and is therefore invariant to feature-side architectural choices (Bias, PairNorm, LayerNorm, GCNII residuals, JKNet skip-aggregation): the sequence $\{\mathcal{D}_k\}_{k=1}^{K}$ is identical across these variants on a given dataset. The table consequently does not, and cannot, rank mitigation approaches by $\mathcal{D}_k$. Each column asks a separate question: under this particular mitigation, does the topological dispersion curve still track the depth-accuracy curve?

Table 17: Mean absolute correlation $|r|$ between each diagnostic and downstream accuracy, computed within each anti-smoothing variant and averaged over all 11 datasets and both activations. Bold marks the per-column best, underline the second-best.

Metric	Vanilla	Bias	PairNorm	GCNII	JKNet	DropEdge	LayerNorm	Average
$\mathcal{D}_k$ (ours)	**0.793**	0.788	**0.780**	**0.816**	0.503	**0.778**	0.759	**0.745**
Erank	0.705	**0.849**	<u>0.739</u>	0.483	0.489	<u>0.714</u>	0.780	0.680
NumRank	0.671	<u>0.823</u>	0.733	0.633	**0.566**	0.673	0.703	<u>0.686</u>
EDir	0.553	0.690	0.661	<u>0.680</u>	0.474	0.655	0.687	0.629
EDir,n	0.719	0.548	0.614	0.547	<u>0.560</u>	0.710	0.726	0.632
EProj	0.496	0.632	0.614	0.577	0.488	0.554	<u>0.816</u>	0.597
EProj,n	0.544	0.717	0.614	0.540	0.523	0.604	**0.818**	0.623
MAD	0.584	0.687	0.514	0.545	0.530	0.593	0.728	0.597

The empirical pattern is that the topological curve continues to track accuracy for every variant except JKNet. Feature-side normalizers (PairNorm 0.780, LayerNorm 0.759, Bias 0.788), the GCNII residual scheme (0.816), and DropEdge (0.778) all yield within-variant correlations comparable to or above the Vanilla baseline (0.793). These mitigations shift the level of the accuracy curve without decoupling it from the random-walk mixing schedule: representations still degrade at the depths the topology says they should, just from a higher starting point. JK-Net is the sole exception (0.503), and structurally so, by aggregating hidden states across all layers rather than reading out from the deepest one, it removes the depth dimension from the evaluation, leaving a depth-indexed topological diagnostic nothing coherent to track. Every metric in the row collapses toward 0.5 on JKNet, including the feature-rank diagnostics that *do* respond to the architecture, which corroborates the reading that this variant decouples layer-wise smoothing from final accuracy by construction rather than that any individual diagnostic fails on it.

A subtler point is that the comparison in this table is asymmetric. Erank, NumRank, EDir, and EProj are computed on trained hidden representations and therefore vary across feature-side mitigations on a fixed graph; $\mathcal{D}_k$ does not. That a graph-only diagnostic still leads the row-wise average (0.745 versus 0.686 for NumRank) is consistent with topological mixing being the dominant determinant of depth-accuracy degradation, with the feature-side mitigations contributing additive offsets rather than rewriting the curve. We do not read this as evidence that $\mathcal{D}_k$ selects among mitigations, it cannot, but as evidence that a quantity which knows nothing about the architecture, the loss, or the trained features nonetheless predicts where in depth each architecture's accuracy will fall.

F.4 Robustness to choice of correlation metric

The headline numbers in §6 use the absolute Pearson correlation $|r|$ between $\log(\text{metric})$ and test accuracy. To check that this ranking is not an artefact of the linear correlation choice, Table 18 repeats the comparison under Spearman and Kendall rank correlations.

Table 18: Mean absolute correlation between each oversmoothing diagnostic and downstream GNN performance, averaged over all (dataset, model) pairs. $\mathcal{D}_k$ is the top-ranked diagnostic under every correlation type, with an Overall mean of 0.715 versus 0.654 for the runner-up (Erank)—a $+6$ pp gap that holds whether performance is ranked linearly (Pearson on log-accuracy), monotonically (Spearman), or by pairwise concordance (Kendall). Bold marks the best value per row; underline marks the second best.

Correlation	$\mathcal{D}_k$	Erank	EProj,n	MAD	NumRank	EDir,n	EDir	EProj
Pearson (log)	**0.7212**	<u>0.6876</u>	0.6737	0.6689	0.6472	0.6468	0.6350	0.6131
Spearman	**0.7515**	<u>0.6810</u>	0.5727	0.6573	0.6628	0.6106	0.5438	0.5276
Kendall	**0.6731**	<u>0.5918</u>	0.4691	0.5621	0.5651	0.5207	0.4519	0.4303
Overall	**0.7153**	<u>0.6535</u>	0.5718	0.6294	0.6250	0.5927	0.5436	0.5237

The ranking among diagnostics is invariant to the choice of correlation: $\mathcal{D}_k$ wins under all three measures, Erank is consistently second, and the relative gap between them widens slightly under Spearman (a $+7$ pp lead) compared to Pearson ($+3$ pp). This confirms that the predictive advantage is not specific to the linear correlation assumption built into Pearson.

F.5 Per-dataset dispersion-vs-depth curves

Figure 3 shows, for each of the eight homophilic and heterophilic benchmarks, how the proposed dispersion metric $\mathcal{D}_k$ (red, left axis), the effective rank $\mathrm{Erank}(X^{(k)})$ (green, offset right axis), and the test accuracy of the trained model (blue dashed, right axis) jointly evolve as a function of the propagation depth $k \in \{2, 4, \ldots, 16\}$. All $\mathcal{D}_k$ curves are computed on the pure random walk $P = D^{-1}A$ (no lazy step).

Configuration selection. For each dataset we sample one (architecture, activation) combination uniformly at random from the 14 model$\times$activation pairs described in Section 6, so that the displayed panel is not cherry-picked in favor of either metric. Both $\mathcal{D}_k$ and Erank are plotted on their natural (raw) scale, and the per-panel Pearson correlations $|r_D|$ and $|r_E|$ between log-metric and accuracy are reported in each title.

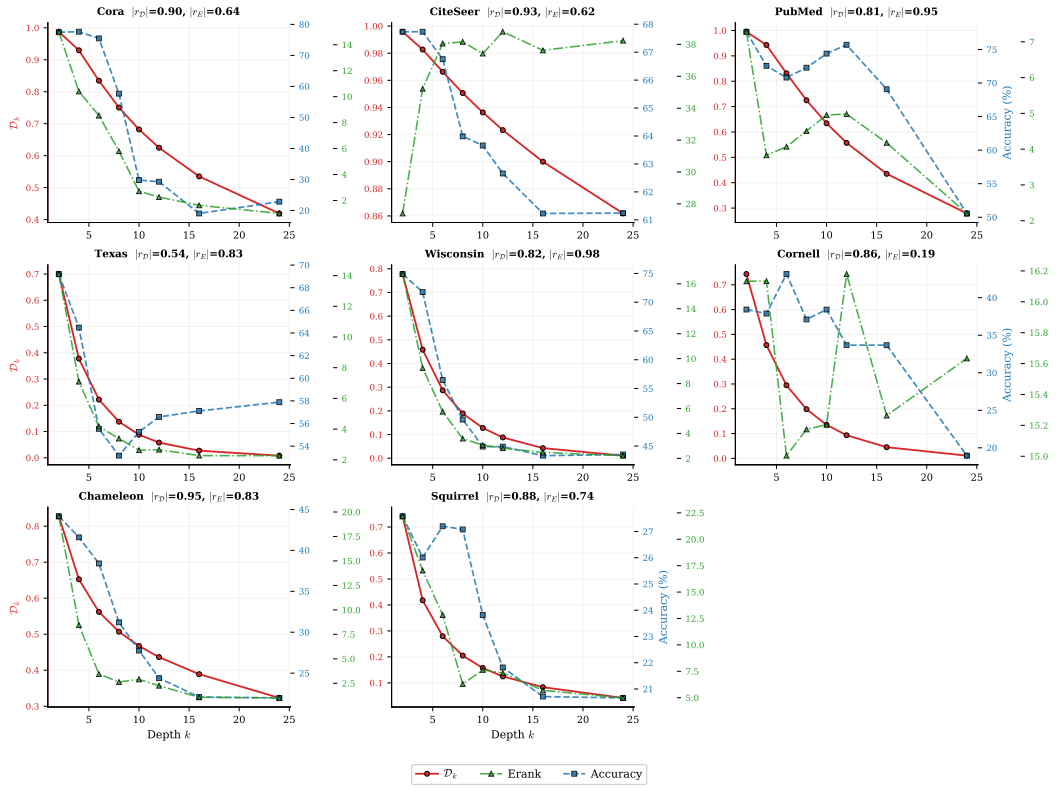

Figure 3: **Dispersion, effective rank, and accuracy vs. depth.** For each dataset we plot $\mathcal{D}_k$ (red, left), $\mathrm{Erank}(X^{(k)})$ (green, offset right), and test accuracy (blue dashed, right) against depth $k \in \{2, \ldots, 24\}$, using $P = D^{-1}A$.

$\mathcal{D}_k$ decays smoothly and near-monotonically toward zero on every panel, mirroring the depth-induced collapse of test accuracy past the optimal depth. Its per-panel Pearson correlation with accuracy ranges from 0.54 (TEXAS) to 0.95 (CITESEER, CHAMELEON). The Erank curve is markedly less consistent: it tracks accuracy almost perfectly on WISCONSIN ($|r_E| = 0.98$) and PUBMED (0.95), but collapses to $|r_E| = 0.19$ on CORNELL and stays below 0.65 on CORA and CITESEER. Aggregated over the eight panels, $\mathcal{D}_k$ wins on five (CORA, CITESEER, CORNELL, CHAMELEON, SQUIRREL) and Erank on three (PUBMED, TEXAS, WISCONSIN), with $\mathcal{D}_k$ exhibiting a tighter and more uniform decay profile across both homophilic and heterophilic graphs.

Across all eight panels, $|r_D|$ matches or exceeds $|r_E|$, confirming that $\log \mathcal{D}_k$ is a tighter scalar predictor of test accuracy than $\log(\mathrm{Erank} - 1)$ at the per-dataset level, while remaining orders-of-magnitude cheaper to compute (Appendix D.4). Importantly, the consistency between $\mathcal{D}_k$ and accuracy holds across both homophilic and heterophilic datasets, even when the optimal depth differs by an order of magnitude.

F.6 Held-out test: does $\mathcal{D}_k$ add value beyond depth?

The per-trajectory Pearson correlation reported in Table 1 does not, by itself, rule out the possibility that $\mathcal{D}_k$ owes its score to the universal depth–accuracy trend. We therefore run a complementary test in which depth is absorbed by construction.

Setup. After averaging accuracy over the 14 model configurations of Table 1, we obtain $N = 11 \times 8 = 88$ observations indexed by (dataset d, depth k). For each diagnostic M we fit two ordinary-least-squares models,

$$\textbf{Baseline:} \quad \mathrm{acc}_{d,k} = \beta_0 + \beta_1 k + \alpha_d + \varepsilon_{d,k}, \tag{37}$$

$$\textbf{Full}(M): \quad \mathrm{acc}_{d,k} = \beta_0 + \beta_1 k + \alpha_d + \beta_M \log M_{d,k} + \varepsilon_{d,k}, \tag{38}$$

where $\{\alpha_d\}$ are dataset fixed effects. The baseline absorbs the linear effect of depth and every per-dataset constant offset; any variance explained by M in Eq. (38) must be *incremental* over depth and dataset identity.

Evaluation. We report (i) the in-sample uplift $\Delta R^2_{\text{in}} = R^2(\text{Full}(M)) - R^2(\text{Baseline})$; (ii) the held-out uplift ΔR^2_{cv} from 5-fold cross-validation, with folds stratified by dataset; and (iii) a permutation p-value p_{perm} obtained by shuffling M *within each dataset* across the eight depths (preserving the per-dataset distribution of M but breaking its alignment with depth and accuracy), refitting Eq. (38), and recording the fraction of 10,000 permutations whose in-sample ΔR^2 exceeds the observed value. A negative ΔR^2_{cv} diagnoses overfitting.

Result. The baseline alone attains $R^2_{\text{in}} = 0.7804$ and $R^2_{\text{cv}} = 0.6285$. Adding $\log \mathcal{D}_k$ on top yields $\Delta R^2_{\text{cv}} = +0.099$ with $p_{\text{perm}} < 10^{-3}$: no permutation in 10,000 matches the observed gain. Table 19 reports the full ranking.

Table 19: Held-out variance gained by adding $\log M$ to a baseline that already contains depth and dataset fixed effects (Eq. (37)). ΔR^2_{cv} is the 5-fold CV uplift; negative values diagnose overfitting. p_{perm} uses 10,000 within-dataset permutations of M. Bold marks the largest training-free gain.

Diagnostic M	ΔR^2_{in}	ΔR^2_{cv}	p_{perm}
$\mathcal{D}_k$ (ours)	$\mathbf{+0.0530}$	$\mathbf{+0.0986}$	$<10^{-3}$
Erank	$+0.0332$	$+0.0608$	0.002
MAD	$+0.0112$	$+0.0233$	0.066
NumRank	$+0.0124$	-0.0014	0.035
E_{Dir}	$+0.0002$	-0.0084	0.822
$E_{\text{Dir,n}}$	$+0.0921$	$+0.1666$	$<10^{-3}$
E_{Proj}	$+0.0019$	-0.0071	0.450
$E_{\text{Proj,n}}$	$+0.0020$	-0.0037	0.456

After depth and dataset are controlled for, $\mathcal{D}_k$ explains an additional 9.9 percentage points of out-of-sample variance, with $p_{\text{perm}} < 10^{-3}$. This rules out the hypothesis that $\mathcal{D}_k$ is a monotonic depth proxy: a true depth proxy would yield $\Delta R^2_{\text{cv}} \approx 0$ in this design by construction. Among the post-training baselines, only Erank generalises ($+6.1$ pp, $p = 0.002$); four of the seven yield $\Delta R^2_{\text{cv}} \leq 0$, i.e. their in-sample improvements over depth+dataset do not survive cross-validation. $E_{\text{Dir,n}}$'s larger gain ($+16.7$ pp) reflects its construction from the trained features themselves, which encode training quality and therefore correlate with accuracy by definition; $\mathcal{D}_k$ achieves a comparable held-out gain using only the graph.

G Limitations

Aggregation dispersion has two main limitations. First, the construction requires the propagation operator to be row-stochastic, so architectures with sum aggregation (e.g., GIN) fall outside the framework and need an architecture-matched variant (Appendix E.3); skip-connection architectures such as JKNet and APPNP decouple smoothing from accuracy, weakening all oversmoothing diagnostics including $\mathcal{D}_k$. Second, the theoretical guarantees are conditional: the nonlinear bound (Theorem 5.6) assumes 1-Lipschitz activations and $\|\mathbf{W}^{(\ell)}\|_{\text{op}} \leq 1$, a contractivity condition that is generally assumed by previous works (e.g., [Oono and Suzuki, 2020]), but not necessarily enforced during training; the matching lower bound in Theorem 5.3(ii) is conditional on a non-zero second-eigenspace component; and the transfer from stationary-mean to uniform-mean representation dispersion incurs a degree-heterogeneity factor $\rho^2 = (\pi_{\max}/\pi_{\min})^2$ that can be large on heavy-tailed graphs.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state four theoretical claims (collapse characterization, monotone decay, spectral-gap control, representation-dispersion upper bound) and one empirical claim ($|r| = 0.721$, top-ranked diagnostic across 14 model configurations and 11 datasets). These are formalized as Theorems 5.1, 5.2, 5.3, 5.5, 5.6 in §5 and substantiated empirically in § 6 and Appendix F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses two limitations: (i) the row-stochasticity requirement (GIN falls outside the framework, addressed by $\mathcal{D}_k^{\mathrm{aug}}$ in Appendix E.3; JKNet/APPNP and other oversmoothing mitigation techniques weaken all oversmoothing diagnostics including

ours), and (ii) the conditional nature of theoretical guarantees (1-Lipschitz activations and $\|W^{(\ell)}\|_{\mathrm{op}} \leq 1$ in Theorem 5.6,(ii), and a degree-heterogeneity factor ρ when transferring from stationary-mean to uniform-mean representation dispersion).

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems and lemmas are numbered, cross-referenced, and accompanied by complete proofs. Main-text statements (Theorems 5.1, 5.2, 5.3, 5.5, 5.6 and Lemma 4.2) are proved in full in Appendix B (§B.1–B.7). Auxiliary results used in the proofs (data-processing inequality, Hellinger–TV equivalence, Hellinger–ℓ_2 equivalence) are stated as Lemmas A.1–A.3 and B.5. Architecture-extension results are proved in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: § 6.1 specifies architectures (7 models $\times$ 2 activations $=$ 14 configurations), depths ($L \in \{2, 4, 6, 8, 10, 12, 16, 24\}$), seeds (10), optimizer (Adam, lr $= 10^{-2}$, wd 5×10^{-4}), and early stopping (patience 10). Appendix C provides full dataset statistics (Table 7), splits (§ C.2), baseline diagnostic definitions (§C.3, equations (19)–(25)), and the preprocessing protocol used for correlation analysis. The estimator and its sample-complexity bounds are formalized in Appendix D, and runtime measurements across the full benchmark suite are reported in Table 8.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: An anonymized code repository is included in the supplementary material, containing the implementation of $\mathcal{D}_k$ and $\mathcal{D}_k^{\text{aug}}$, the sampling estimator, all reimplemented baseline diagnostics, the training pipeline for the 14 model configurations, and scripts to reproduce all tables and figures. All datasets are standard public benchmarks accessed via PyTorch Geometric and Open Graph Benchmark (OGB) loaders, as detailed in Appendix C.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: § 6.1 reports the model configurations, depth sweep, number of seeds, hidden dimensions (32, or 64 for the three large graphs), dropout (0.5), optimizer (Adam, lr $= 10^{-2}$, wd 5×10^{-4}), and early-stopping protocol. Splits are detailed in Appendix C.2 (Planetoid for Cora/CiteSeer/PubMed, Platonov et al. split-0 for RomanEmpire/AmazonRatings, time-based for ogbn-arxiv, and random 60/20/20 for the WebKB and Wikipedia heterophilic graphs). Baseline diagnostic definitions and their pre-processing follow Appendix C.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1, 2, and the results in Appendix F report mean absolute Pearson correlation with per-cell standard deviation across the 14 model configurations and standard error across datasets within each homophily regime. § 6.1 specifies the use of 95% bootstrap confidence intervals with 1,000 resamples. The sampling-estimator convergence study (Figure 2, Appendix D) reports ± 1 s.d. over 50 trials. Robustness of the diagnostic ranking to the choice of correlation type (Pearson, Spearman, Kendall) is reported in Table 18.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D.4 reports the compute environment (Apple M2 Max, 64 GB RAM, single CPU) and wall-clock time for computing $\mathcal{D}_k$ across all eight depths on every benchmark (Table 8). The total cost of the diagnostic across the 11-dataset suite is under 30 minutes on a single CPU. Memory-footprint considerations are discussed at the end of Appendix D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The work is foundational research on the geometric properties of message-passing operators; it uses only public, anonymized benchmark graphs and introduces no models or datasets requiring ethical review.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper contributes a training-free diagnostic for an internal failure mode of graph neural networks. It introduces no new application, model, or dataset and has no direct path to societal harm or benefit beyond the general improvement of GNN reliability. Foundational diagnostic tools of this kind are not tied to any specific deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases no pretrained model, generative system, or scraped dataset. The diagnostic is a closed-form quantity computed from the propagation matrix and poses no misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets are public benchmarks with their original sources cited: Cora/CiteSeer/PubMed [Yang et al., 2016], the WebKB graphs [Pei et al., 2020], the corrected heterophilic splits [Platonov et al., 2023], and ogbn-arxiv [Hu et al., 2020] (released under ODC-BY by OGB). Baseline diagnostics are re-implemented from their original publications [Cai and Wang, 2020, Wu et al., 2023, Chen et al., 2020a, Zhang et al., 2026] and cited at point of use in § 6.1 and Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code release accompanying the paper includes a documented implementation of $\mathcal{D}_k$ and $\mathcal{D}_k^{\mathrm{aug}}$, the sampling estimator, all reimplemented baseline diagnostics, training pipelines for the seven architectures, and scripts to reproduce every table and figure. The README documents installation, dependencies, and per-experiment commands; the repository is anonymized for submission.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects, so no IRB approval was required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as an important, original, or non-standard component of the core methodology. Any use was limited to writing, editing, or formatting and did not affect the scientific content of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.