WHAT IS STILL MISSING IN GRAPH HOMOPHILY? DEGREE REGULARITY MATTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph homophily, the tendency of connected nodes to share similar characteristics, is widely used to understand and predict the performance of Graph Neural Networks (GNNs). Most homophily metrics, however, capture only local similarity patterns, such as label or feature consistency across edges. In this paper, we study whether a global graph property, the dispersion of the degree distribution, carries performance-relevant information that local homophily misses. We introduce *degree regularity*, which quantifies how uniform the node degrees are, relate degree dispersion to aggregation noise through elementary results, and add it to a local homophily factor to obtain *degree-aware glocal homophily*. To study degree dispersion in a controlled setting, we develop a calibrated degree-aware synthetic graph generator, which reveals that varying degree dispersion can silently vary density; at nearly fixed density, the effect of dispersion on GNN accuracy changes sign with label homophily, as a stylized aggregation model explains. On 31 real-world datasets, with baseline metrics recomputed from their definitions, degree regularity is positively associated with GNN accuracy beyond local homophily. Degree-aware glocal homophily attains the highest correlation with the accuracy of GNNs among 24 metrics. These results suggest that global degree properties complement local homophily in explaining GNN performance.

1 INTRODUCTION

Graph Neural Networks (GNNs) have achieved remarkable success in node classification (Kipf & Welling, 2017; Veličković et al., 2018), link prediction (Zhang & Chen, 2018), and graph representation learning (Hamilton et al., 2017). Their effectiveness on node-level tasks is closely related to homophily, the tendency of connected nodes to be similar (McPherson et al., 2001), and many homophily metrics have been proposed to characterize dataset difficulty and predict GNN performance (Zhu et al., 2020; Pei et al., 2020; Lim et al., 2021b). These metrics quantify local similarity patterns, such as label consistency (Zhu et al., 2020; Pei et al., 2020) or feature similarity (Jin et al., 2022; Chen et al., 2023; Yang et al., 2021) across neighboring nodes.

Existing metrics are nevertheless insufficient to fully explain GNN performance (Luan et al., 2022; Ma et al., 2021; Luan et al., 2023; Platonov et al., 2023a): GNNs can perform well on heterophilic graphs (Ma et al., 2021; Luan et al., 2022), and performance does not follow conventional homophily scores monotonically (Luan et al., 2023). The recent Tri-Hom framework (Zheng et al., 2024) disentangles homophily into label, structural, and feature aspects and derives homophily-driven factors from a contextual stochastic block model (CSBM), improving the alignment with GNN performance. However, these factors are computed from local neighborhood patterns and do not account for global graph properties.

Node degree is a natural candidate for what is still missing. GNNs are known to be biased toward high-degree nodes (Tang et al., 2020; Liu et al., 2021; Ju et al., 2023; Subramonian et al., 2024), degree-corrected block models were introduced precisely because real graphs have broad degree distributions (Karrer & Newman, 2011), large synthetic studies report that GNN performance varies with the degree distribution as well as with homophily (Palowitch et al., 2022; Yasir et al., 2023), and a regression over benchmark results identifies degree imbalance as a salient data property (Li et al., 2023b). What remains open is whether a graph-level summary of degree dispersion adds information *beyond* a performance-reflective homophily factor, how robust this is to the family structure of benchmark collections, and whether its effect has the same sign in all homophily regimes.

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

In this work, we study degree dispersion as a global complement to local homophily. We introduce *degree regularity* $r_{\deg}$, a simple dispersion statistic that equals one exactly for regular graphs, and add it to Tri-Hom as *degree-aware glocal homophily*. To examine its effect under controlled conditions, we develop DA-CSBM, a degree-aware extension of the Tri-Hom generator. We then audit 31 real-world benchmarks with baseline metrics recomputed from their definitions and with controls for significance, family structure, and our own design search.

Our key contributions are as follows:

- **Degree dispersion carries information that local homophily misses.** With baseline metrics recomputed from their definitions, significance and power analyses, family structure, and a correction for our own design search, we find that degree dispersion is associated with GNN accuracy beyond a performance-reflective homophily factor. The evidence is conditional, mostly between benchmark families, and holds for several dispersion statistics, more strongly for the Gini coefficient than for degree regularity.
- **Its effect changes sign with label homophily.** A calibrated degree-aware variant of the Tri-Hom generator, DA-CSBM (Lemma 1), exposes a density confound in degree-varying generators and shows that, at nearly fixed density, regular graphs are easier for GCN (Kipf & Welling, 2017) on homophilic graphs but harder on moderately heterophilic ones, which an elementary aggregation model explains (Proposition 2).
- **A simple way to add dispersion to homophily factors.** We relate $r_{\deg}$ to the noise of mean aggregation through elementary bounds (Proposition 1) and add it to Tri-Hom as *degree-aware glocal homophily*, a parameter-light empirical scoring rule rather than a derived model. It has the highest correlation with the accuracy of GCN, GraphSAGE (Hamilton et al., 2017), and GAT (Veličković et al., 2018) among 24 metrics (e.g., GCN: $0.666 \rightarrow 0.692$ over Tri-Hom).

2 Related Work

Homophily metrics. Label-based metrics measure same-class connections at the edge, node, or class level (Zhu et al., 2020; Pei et al., 2020; Lim et al., 2021b). Their values depend on the number and balance of classes, which motivated adjusted variants (Platonov et al., 2023a; Li et al., 2023a) and the unbiased measure of Mironov & Prokhorenkova (2025); since our datasets have 2 to 70 classes, we include it as a baseline and control for the number of classes. Other metrics use multi-hop or neighborhood-distribution information (Cavallo et al., 2022; Gong et al., 2023; Ma et al., 2021; Luan et al., 2022), feature similarity (Jin et al., 2022; Chen et al., 2023; Yang et al., 2021; Lee et al., 2024), or classifier-based separability (Luan et al., 2023); Zhao et al. (2026) relate the distinguishability of neighborhood label distributions to node classification accuracy and propose a corresponding homophily metric, which we include as a baseline. Tri-Hom (Zheng et al., 2024) is the only one that combines label, structural, and feature aspects in a closed-form, performance-reflective factor derived from a generative model. These metrics are built from local similarity patterns. Some use graph-level quantities, such as the spectral radius in h_F and Tri-Hom, but none uses the dispersion of the degree distribution as an explicit graph-level covariate. **Degree and GNNs.** Degree-related bias is well documented at the node level: GCNs are more accurate on high-degree nodes (Tang et al., 2020), tail nodes benefit from dedicated designs (Liu et al., 2021; Ju et al., 2023), and Subramonian et al. (2024) trace the bias to the aggregation operator. In the heterophilous SBM analysis of Wang et al. (2024), the separability gain of graph convolution depends on the distance between neighborhood distributions and on the expected node degree. In network science, degree heterogeneity is a basic descriptor of a graph (Newman, 2003), summarized by moments, the Gini coefficient (Kunegis & Preusse, 2012), or spectral indices (Estrada, 2010). These GNN works explain accuracy differences *within* a graph; we ask whether a graph-level summary of the degree distribution explains accuracy differences *across* graphs. At the graph level, GraphWorld (Palowitch et al., 2022) benchmarks GNNs on degree-corrected SBM graphs with power-law degrees, and Yasir et al. (2023) compare SBM, LFR, and CABAM generators on 300,000 synthetic graphs, finding that performance depends on homophily, degree distribution, and feature signal. Across real benchmarks, Li et al. (2023b) regress GNN performance on 15 graph properties, including clustering and homophily, identify the Gini coefficient of the degree distribution as one of the most salient, and confirm in homophilic SBM graphs that accuracy declines with degree imbalance; Liu et al. (2022) profile benchmarks by their sensitivity to structural perturbations. These works already establish that degree distributions can affect GNN performance and motivate degree dispersion as the first global descriptor to examine,

so our novelty is not that degree matters. It is whether a degree-dispersion statistic adds predictive information beyond a performance-reflective homophily factor, tested with significance, family, and design-search controls, and whether its effect has the same sign in all homophily regimes. **Degree-heterogeneous random graphs.** Our DA-CSBM (Section 5) extends the CSBM of Tri-Hom in the manner of the degree-corrected SBM (Karrer & Newman, 2011), with log-normal weights, which fit many empirical degree distributions (Broido & Clauset, 2019). GraphUniverse (Van Langendonck et al., 2026) also controls homophily and degree distributions, but its goal differs: it generates families of graphs with shared communities to test inductive generalization under distribution shift. DA-CSBM instead generates single graphs for factor analysis, parameterized by the three Tri-Hom factors and a target degree regularity, and we report how closely each target is realized.

3 PRELIMINARIES

Notation. We consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes, adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$, and degree vector $\mathbf{d} = \mathbf{A}\mathbf{1}$. Each node v_i has a label $y_i \in \{1,\dots,C\}$ and a feature vector $\mathbf{x}_i \in \mathbb{R}^M$. We write $\mathbf{Z} \in \{0,1\}^{N\times C}$ for the one-hot label matrix, $\mathbf{X} \in \mathbb{R}^{N\times M}$ for the feature matrix, $\mathcal{N}(v_i)$ for the neighborhood of v_i, $\rho(\mathbf{A})$ for the spectral radius, $\mathbf{I}_C$ for the identity, and $\mathbf{1}_C$ for the $C \times C$ all-ones matrix.

Tri-Hom. Zheng et al. (2024) define homophily-driven factors $h^{\mathcal{G}}_{\text{Tri-Hom}}$ and $h^{\neg\mathcal{G}}_{\text{Tri-Hom}}$ for graph-aware models (e.g., GCN (Kipf & Welling, 2017)) and graph-agnostic models (e.g., MLP). Label homophily h_L is node homophily; structural homophily h_S measures how consistent the neighbor-class distributions of same-class nodes are; feature homophily h_F is the strength of the feature diffusion $\mathbf{X} = (\mathbf{I} - \frac{h_F}{\rho(\mathbf{A})}\mathbf{A})^{-1}\mathbf{X}^{(0)}$ that best explains the observed features. Both factors share the term

$$T_{\text{local}} = C\Big(\tfrac{1-h_L}{C-1}\Big)^2 + C\tfrac{(1-h_S)^2}{C-1} + \Big(\tfrac{h_L C-1}{C-1}\Big)^2, \tag{1}$$

which collects squared label and structural terms of CSBM-3H. T_{local} is not exactly the second moment of a node's neighbor-class vector: writing $q = \frac{1-h_L}{C-1}$ and $a = \frac{h_L C-1}{C-1}$, the vector in Eq. 6 is $q\mathbf{1} + a\mathbf{e}_{y_i} + \boldsymbol{\epsilon}_i$, whose expected squared norm is $T_{\text{local}} + 2aq$. What matters here is only that h_S enters T_{local} through $C(1-h_S)^2/(C-1)$, i.e., C times the per-entry variance of the node-wise perturbation. Full definitions are in Appendix F. *Why a factor for graph-agnostic models?* An MLP does not aggregate neighbors, but under CSBM-3H the observed features are themselves produced by diffusion over $\mathbf{A}$. The class separability of $\mathbf{X}$, and hence MLP accuracy, therefore depends on the topology through h_F and T_{local}; $h^{\neg\mathcal{G}}$ quantifies this dependence, and $h^{\mathcal{G}}$ adds the effect of explicit aggregation. *Normalization.* Following the implementation we build on, all Tri-Hom and glocal values for real graphs are computed with the fixed constants $C = 3$ and $\rho(\mathbf{A}) = 10$ for every dataset. The factors are unnormalized scores that can exceed 1 (up to 1.165 on 14 datasets); we use only their association with accuracy. Appendix D.6 repeats the analysis with dataset-specific C and $\rho(\mathbf{A})$.

4 DEGREE REGULARITY AND DEGREE-AWARE GLOCAL HOMOPHILY

Nodes with different degrees aggregate information from different numbers of neighbors, and real-world graphs have broad, often right-skewed degree distributions (Appendix Table 18). We seek a graph-level statistic of degree dispersion that can be combined with the local factors of Tri-Hom.

Degree regularity. With $\mu = \mathbb{E}[\mathbf{d}]$, $\mathrm{CV}^2(\mathbf{d}) = \mathrm{Var}(\mathbf{d})/\mu^2$ ($:= 0$ for edgeless graphs), we define

$$r_{\deg} = \exp\big(-\gamma \cdot \mathrm{CV}^2(\mathbf{d})\big), \qquad \gamma > 0. \tag{2}$$

By construction $r_{\deg} \in (0,1]$, and $r_{\deg} = 1$ if and only if the graph is regular (all degrees equal), which motivates the name. Smaller values indicate a more dispersed (heterogeneous) degree distribution; $r_{\deg}$ measures concentration of the degree distribution, not closeness to a uniform distribution. The exponential maps the unbounded CV^2 to the same $(0,1]$ scale as h_S, and γ sets its resolution. We use $\gamma = 0.2$; every $\gamma \in [0.01, 0.2]$ improves over Tri-Hom for all GNNs (Appendix D.5).

Degree dispersion and aggregation noise. Many dispersion statistics exist (e.g., the Gini coefficient). The following elementary result shows how CV^2 enters the noise of mean aggregation.

Proposition 1 (Aggregation noise)**.** *Let $\mathbf{x}_j = \mathbf{m}_{y_j} + \boldsymbol{\xi}_j$, with class means $\mathbf{m}_c$ and i.i.d. zero-mean noise of covariance $\sigma^2\mathbf{I}$, independent of the graph, and let $\bar{\mathbf{x}}_i = \frac{1}{d_i}\sum_{j\in\mathcal{N}(v_i)}\mathbf{x}_j$. For a graph with*

degrees in $[d_{\min}, d_{\max}]$, $d_{\min} \geq 1$, *the average per-coordinate noise variance* $\bar{v} = \frac{1}{N}\sum_i \mathrm{Var}(\bar{x}_i \mid \mathcal{G}, \mathbf{y}) = \frac{\sigma^2}{N}\sum_i d_i^{-1}$ *satisfies*

$$\frac{\sigma^2}{\mu}\Big(1 + \frac{\mu}{d_{\max}}\,\mathrm{CV}^2(\mathbf{d})\Big) \;\leq\; \bar{v} \;\leq\; \frac{\sigma^2}{\mu}\Big(1 + \frac{\mu}{d_{\min}}\,\mathrm{CV}^2(\mathbf{d})\Big), \tag{3}$$

with equality if and only if the graph is regular (for $\sigma^2 > 0$*). Moreover, at a fixed mean degree* μ*: (a) if* $\mathbf{d}'$ *is a mean-preserving spread of* $\mathbf{d}$ *(convex order), then both* $\bar{v}$ *and* CV^2 *are larger for* $\mathbf{d}'$*; (b) for log-normal degrees,* $\bar{v} = \frac{\sigma^2}{\mu}(1 + \mathrm{CV}^2)$ *exactly (Appendix A adds the Gamma case).*

The proof is in Appendix A. The bounds hold for every graph, but CV^2 alone does not order $\bar{v}$ across arbitrary degree sequences with the same mean, because $d_{\min}$ and $d_{\max}$ also change (counterexample in Appendix A). Monotonicity holds within these families, which include the log-normal model of DA-CSBM. On real graphs, however, CV^2 tracks neither the normalized excess noise $\nu = \mu \cdot \frac{1}{N}\sum_i d_i^{-1} - 1$ (Spearman 0.43) nor that of the actual GCN operator, $\frac{\sigma^2}{N}\|\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\|_F^2$ (0.15), and both work worse as global terms (Appendix D.5). Without degree correlations, the GCN noise is $\approx \sigma^2\big(\mathbb{E}[(d+1)^{-2}] + \mathbb{E}[\frac{d}{d+1}]^2/\mu\big)$, a function of inverse-degree moments, not of CV^2 (Remark 1; Spearman 0.93 with the exact value). Proposition 1 thus only shows that dispersion, summarized by CV^2, enters aggregation noise; it is not a mechanism for real graphs. Among dispersion statistics, $1 - \mathrm{Gini}$ and $1/(1 + \mathrm{CV}^2)$ give GNN correlations within 0.013 of ours (Appendix D.5); for MLP the CV^2-based variant is clearly better.

Degree-aware glocal homophily. In T_{local} (Eq. 1), structural inconsistency enters as a variance term, $C(1 - h_S)^2/(C-1)$. Finite and heterogeneous degrees add a second source of node-wise variation: node v_i observes an empirical neighbor distribution averaged over d_i neighbors, whose sampling variance scales with $1/d_i$ (Proposition 1). We add a term of the same form as the h_S term, using $1 - r_{\deg}$ as the dispersion level:

$$T_{\text{glocal}} = T_{\text{local}} + \underbrace{C\,\tfrac{(1-r_{\deg})^2}{C-1}}_{\text{global}}, \tag{4}$$

$$h^{\neg\mathcal{G}}_{\text{glocal}} = \frac{1 - \big(\frac{h_F}{\rho(\mathbf{A})}\big)^2 T_{\text{glocal}}}{\big[1 - \frac{h_F}{\rho(\mathbf{A})}\frac{h_L C - 1}{C-1}\big]^2}, \qquad h^{\mathcal{G}}_{\text{glocal}} = \frac{\big(\frac{h_L C - 1}{C-1}\big)^2}{T_{\text{glocal}}}\, h^{\neg\mathcal{G}}_{\text{glocal}}. \tag{5}$$

Setting $r_{\deg} = 1$ recovers Tri-Hom exactly. Eq. 4 is a heuristic extension, not a derivation. In particular, the squared form is not implied by Proposition 1: for small dispersion the added term is $\approx \frac{C}{C-1}\gamma^2(\mathrm{CV}^2)^2$, whereas the excess aggregation noise is first order in CV^2. A linear term $C(1 - r_{\deg})/(C-1)$ gives GNN correlations within 0.006 of the squared one (but lower for MLP; Appendix D.5), so the choice between them is not critical, and we keep the squared form for symmetry with the h_S term. A parameter-free term derived from neighborhood sampling, $(1 - \|\mathbf{p}\|^2)\mathbb{E}[1/d]$, barely changes the factor (Appendix Table 40): Propositions 1 and 2 explain qualitatively why dispersion can matter and why its effect can change sign, but do not justify Eq. 4 quantitatively. We therefore treat Eq. 4 as an empirical scoring rule, not as a scientific model of how degree dispersion affects GNNs; Section 6.1 shows that the true effect is regime-dependent, and Section 6.2 compares the rule with additive, multiplicative, regression-based, and regime-gated alternatives. The following stylized result shows why the effect of dispersion can change sign, which Eq. 4 cannot represent; it is a toy explanation, not a model of trained GNNs.

Proposition 2 (Regime-dependent effect of dispersion)**.** *Consider two balanced classes with features* $x_i = \pm m + \xi_i$*, i.i.d.* $\xi_i \sim \mathcal{N}(0, \sigma^2)$*, mean aggregation with a self-loop,* $\bar{x}_i = (x_i + \sum_{j \in \mathcal{N}(v_i)} x_j)/(d_i + 1)$*, a fraction* $\frac{1+a}{2}$ *of same-class neighbors for every node (*$a \in [0, 1]$*), and the sign rule. Node* v_i *is then classified correctly with probability* $f_a(d_i) = \Phi\big(s\,\frac{1 + a d_i}{\sqrt{d_i + 1}}\big)$*,* $s = m/\sigma$*. For* $a = 1$*,* f_a *is strictly concave, so any mean-preserving spread of the degrees lowers mean accuracy; for* $a = 0$ *and* $s^2 < 3(d_{\min} + 1)$*,* f_a *is strictly convex, so any mean-preserving spread raises it.*

When neighbors are uninformative, high-degree nodes average away their own feature, so dispersion helps. With C classes and centered means, the aggregated mean is $\mathbf{m}_{y_i}\frac{1 + a d_i}{d_i + 1}$ with $a = \frac{h_L C - 1}{C-1}$, so the model carries over; for $C = 3$ both statements hold in closed form (Appendix A). It reproduces

the crossover near $h_L \approx 0.5$ on our synthetic grid but *fails for* $a < 0$: at $h_L = 0$ it predicts -11.5 to -19.9 points vs. a measured $+0.2$, since the mean flips sign at $d_i = 1/|a|$ and the fixed sign rule, unlike a trained GNN, then misclassifies high-degree nodes. On real graphs its dispersion component adds nothing beyond Tri-Hom (Appendix Tables 6, 7). We keep a single-sign term because the reversal occurs near $h_L \approx 1/C$, where the Tri-Hom factor is small, and a regime-gated variant did not help out of sample (Appendix D.3). As in Tri-Hom, "homophily" in the name refers to the factor, not to a new notion of homophily; degree dispersion is distinct from class homophily. For graph-agnostic models, r_{deg} enters only through the feature-diffusion term, scaled by $(h_F/\rho(\mathbf{A}))^2$, so its effect on $h^{\neg\mathcal{G}}_{\text{glocal}}$ is small by design. This is consistent with the data: r_{deg} carries no information about MLP accuracy beyond Tri-Hom (partial $r = 0.03$, Appendix D.1).

5 DA-CSBM: A Degree-Aware CSBM for Controlled Analysis

To vary degree dispersion at fixed local-factor targets, we extend the CSBM-3H generator of Tri-Hom with node-specific degree weights, in the spirit of the DC-SBM (Karrer & Newman, 2011). Algorithm 1 (Appendix B.1) gives the full procedure; we state the model exactly as implemented.

Degree weights. Given a target r_{deg}, we set $\text{CV}_{\text{t}}^2 = -\log(r_{\text{deg}})/\gamma$ and draw degree weights $w_i \sim \text{LogNormal}(\mu', \sigma^2)$ with $\sigma^2 = \log(1 + \text{CV}_{\text{t}}^2)$ and $\mu' = \log \bar{w} - \sigma^2/2$, so that $\mathbb{E}[w_i] = \bar{w}$ and $\text{CV}^2(w_i) = \text{CV}_{\text{t}}^2$ before rounding and clipping to $[1, 0.2N]$. We write w_i for these generative weights to distinguish them from the realized degrees d_i of Section 4. Unlike a heavy-tailed Pareto law, the log-normal attains any target CV^2 with finite variance and describes many empirical networks well (Broido & Clauset, 2019).

Topology. Labels are uniform over C classes. The neighbor-distribution matrix is

$$\mathbf{D}^{\mathcal{N}} = \mathbf{ZS} + \mathbf{E}, \quad \mathbf{S} = h_L \mathbf{I}_C + \tfrac{1-h_L}{C-1}(\mathbf{1}_C - \mathbf{I}_C), \quad E_{ic} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \tfrac{(1-h_S)^2}{C-1}\big), \tag{6}$$

where the noise $\mathbf{E} \in \mathbb{R}^{N \times C}$ is node-wise, as in CSBM-3H, so that same-class nodes have different neighbor distributions. With $\mathbf{\Theta} = \text{diag}(\boldsymbol{\theta})$ and $\theta_i = w_i/\sqrt{\bar{w}}$, edge probabilities are

$$\mathbf{A}_p = \text{clip}_{[0,1]}\Big(\tfrac{C}{N}\, \mathbf{\Theta}\mathbf{D}^{\mathcal{N}}\mathbf{Z}^\top\mathbf{\Theta}\Big). \tag{7}$$

Clipping also removes the negative entries that the Gaussian noise can create in $\mathbf{D}^{\mathcal{N}}$. We sample $A_{ij} \sim \text{Bernoulli}((\mathbf{A}_p)_{ij})$ for $i \le j$, set $A_{ji} = A_{ij}$, which makes the graph undirected, and discard the diagonal, so the final graph is simple (GCNs add self-loops themselves). Because $\mathbf{D}^{\mathcal{N}}$ is node-specific, $\mathbf{A}_p$ is not symmetric, and this rule uses the perturbation of the lower-indexed endpoint of each pair: node v_i controls a fraction $(N-i)/(N-1)$ of its pairs. Indices are assigned independently of labels, weights, and noise, so this creates no systematic link between orientation and any node attribute, and a variant with the symmetric probability matrix $(\mathbf{A}_p + \mathbf{A}_p^\top)/2$ yields nearly identical realized statistics, including the weak h_S control reported below (Appendix Table 15); the orientation rule therefore does not affect our conclusions. Eq. 7 is a DC-SBM with $\theta_i \propto w_i$ (the DC weighting) and node-specific block affinities, so the weights set the expected degrees. Our preliminary version used $\theta_i = \sqrt{w_i}$ (the $\sqrt{w}$ *variant*), for which they do not:

Lemma 1 (Expected degrees)**.** *Ignoring clipping and symmetrization, and with labels independent of* $\mathbf{w}$*, Eq. 7 gives* $\mathbb{E}[d_i \mid \mathbf{w}] \approx w_i$*. The* $\sqrt{w}$ *variant gives* $\mathbb{E}[d_i \mid \mathbf{w}] \approx \sqrt{w_i}\,\overline{\sqrt{w}}$*,* $\overline{\sqrt{w}} = \frac{1}{N}\sum_j \sqrt{w_j}$*; for log-normal weights, its expected degrees have* $\text{CV}^2 = (1 + \text{CV}_{\text{t}}^2)^{1/4} - 1 < \text{CV}_{\text{t}}^2$ *and mean* $\bar{w}\,(1 + \text{CV}_{\text{t}}^2)^{-1/4}$*, so dispersion is compressed and coupled with density.*

Calibration. We measured realized statistics on 2,640 generated graphs (Table 1, Appendix B.2). With Eq. 7, realized r_{deg} is monotone in the target (Spearman 0.99) and spans $[0.48, 0.97]$ (clipping and the degree cap keep it above the target), and the mean degree varies little (10.1 to 12.4) but is not constant. The $\sqrt{w}$ variant compresses realized r_{deg} to $[0.84, 0.97]$, as Lemma 1 predicts to within 0.03, and *lower target regularity also lowers its mean degree* (12.4 to 6.8). Clipping negative entries of $\mathbf{A}_p$ to zero raises the mean degree above Lemma 1

Table 1: DA-CSBM calibration: target vs. realized r_{deg} and mean degree (3 seeds, all h_L, h_S). Pred.: Lemma 1 plus Bernoulli noise.

	Eq. 7 (default)		$\sqrt{w}$ variant		
target	realized	$\bar{d}_{\text{real}}$	realized	pred.	$\bar{d}_{\text{real}}$
0.1	0.476	10.1	0.841	0.814	6.8
0.4	0.582	11.3	0.878	0.876	8.1
0.7	0.724	12.0	0.916	0.923	9.5
1.0	0.973	12.4	0.973	0.984	12.4

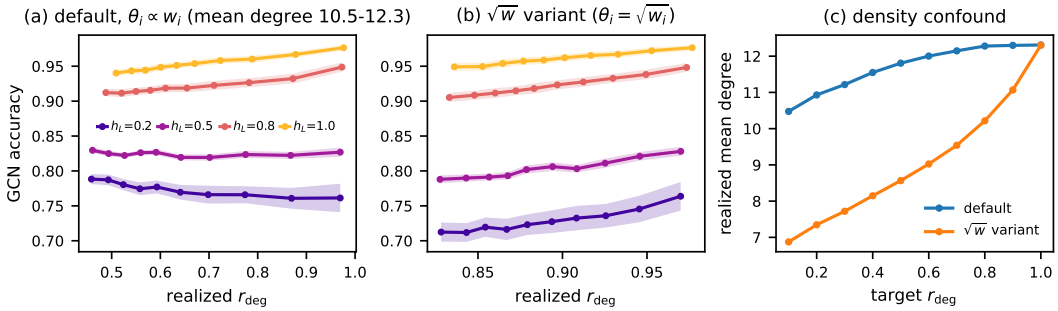


Figure 1: (a,b) GCN accuracy on DA-CSBM graphs vs. realized r_{deg}, averaged over h_S and h_F (bands: ± 1 s.e.), for Eq. 7 (a) and the $\sqrt{w}$ variant (b); (c) mean degree vs. target r_{deg}. h_L: targets.

(by 17–23%) and makes density depend on h_S and h_L (10.6 to 15.4 at target $r_{\text{deg}} = 1$), so their conditional correlations partly include density (Appendix Tables 11, 12). In both forms, realized h_L is monotone but compressed (0.11 to 0.79), realized h_F closely tracks its target, and the h_S knob inherited from CSBM-3H is weak (realized $h_S \in [0.48, 0.56]$, Spearman 0.27). DA-CSBM thus fixes targets, not realized factors: it is an approximate intervention. We keep the $\sqrt{w}$ variant because it exposes a density confound.

Features are generated as in CSBM-3H: class-conditional Gaussians $\mathbf{x}_i^{(0)} \sim \mathcal{N}(\mathbf{m}_{y_i}, \mathbf{\Sigma}_{y_i})$, diffused as $\mathbf{X} = (\mathbf{I} - \frac{h_F}{\rho(\mathbf{A})}\mathbf{A})^{-1}\mathbf{X}^{(0)}$.

6 Experiments

6.1 Synthetic Graphs

Setting. We vary target $h_L \in \{0, 0.1, \ldots, 1\}$ and $r_{\text{deg}} \in \{0.1, \ldots, 1.0\}$ for $h_S \in \{0.2, 0.4, 0.6, 0.8\}$ and $h_F \in \{-0.8, -0.4, 0, 0.4, 0.8\}$ (2,200 configurations; $N = 1000$, $C = 3$, $M = 32$, $\bar{w} = 10.5$) with Eq. 7, whose mean degree varies little (10.5 to 12.3), and with the $\sqrt{w}$ variant, whose mean degree falls as dispersion grows. For each configuration we train 1- and 2-layer GCNs with three weight decays on 10 independently generated graphs (60/20/20 split), select the setting with the best mean validation accuracy, and report its mean test accuracy (Appendix C.2). Both grids are trained with the same batched implementation, which reproduces the accuracies of our original implementation (correlation 0.993; Appendix B.3). We report conditional correlations (factor vs. accuracy with the other factors fixed, averaged over slices) and within-slice partial correlations given the realized mean degree.

Degree dispersion at nearly fixed density. With Eq. 7, the mean degree stays within 10.5–12.3, and accuracy correlates with r_{deg} at 0.24 (0.58 for $h_L \geq 0.5$). Averaged over all configurations, accuracy changes by only 0.6 points between target $r_{\text{deg}} = 0.1$ and 1.0, because the effect changes sign with label homophily (Figure 1a, Appendix Table 14). On homophilic graphs ($h_L \geq 0.6$), regular graphs are easier by 2.4–3.7 points; for $0.1 \leq h_L \leq 0.4$, dispersed graphs are easier by 1.4–3.7 points. The former agrees with the decline reported by Li et al. (2023b) for homophilic SBM graphs. The positive partial correlation given realized mean degree (0.32; 0.60 for $h_L \geq 0.5$) confirms that dispersion itself matters. *The density confound.* In the $\sqrt{w}$ variant, accuracy correlates much more strongly with r_{deg} (0.77), but the realized mean degree rises from 6.9 to 12.3 along with it; once it is partialled out,

Table 2: Synthetic GCN grid (2,200 configurations per generator). Top: conditional Pearson correlation between accuracy and each factor, and partial correlation with realized r_{deg} given realized mean degree $\bar{d}$. Bottom: Spearman correlation of each composite score with accuracy; "realized" uses the realized r_{deg}. h_L thresholds refer to targets.

	Eq. 7 ($\bar{d}$ nearly constant)		$\sqrt{w}$ variant ($\bar{d}$ varies)	
	all h_L	$h_L \geq 0.5$	all h_L	$h_L \geq 0.5$
h_L	0.835	0.969	0.875	0.972
h_S	0.065	0.754	−0.005	0.723
h_F	−0.795	−0.711	−0.873	−0.812
r_{deg}	0.235	0.581	0.766	0.866
$r_{\text{deg}} \mid \bar{d}$ (partial)	0.324	0.595	0.021	0.047
$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.866	0.908	0.832	0.896
$h^{\mathcal{G}}_{\text{glocal}}$	0.860	0.908	0.844	0.916
$h^{\mathcal{G}}_{\text{glocal}}$ (realized)	0.873	0.943	0.836	0.908

Table 3: Pearson correlation (p-value) with accuracy on 31 datasets; best and second-best per model in red and blue; Rank: mean rank over the four models. Tri-Hom and glocal use $C = 3$, $\rho(\mathbf{A}) = 10$.

Metric	MLP		GCN		GraphSAGE		GAT		Rank
	Cor.	p-value	Cor.	p-value	Cor.	p-value	Cor.	p-value	
h_{edge}	0.4914	0.0050	0.6192	0.0002	0.5575	0.0011	0.6133	0.0002	7.50
h_{node}	0.4699	0.0077	0.5995	0.0004	0.5379	0.0018	0.6037	0.0003	9.00
h_{class}	0.6167	0.0002	0.6351	0.0001	0.6174	0.0002	0.6425	$< 10^{-4}$	3.00
h_{adj}	0.5029	0.0039	0.5648	0.0009	0.5296	0.0022	0.5712	0.0008	9.75
h_{unb}	0.3862	0.0319	0.4681	0.0079	0.4206	0.0185	0.4650	0.0084	14.50
h_{NLD}	0.6122	0.0003	0.4573	0.0097	0.4492	0.0112	0.4235	0.0176	11.50
h_{den}	0.0606	0.7461	0.2251	0.2234	0.1396	0.4540	0.1937	0.2965	23.00
h_{2hop}	0.4706	0.0076	0.5316	0.0021	0.5119	0.0032	0.5241	0.0025	11.00
h_{nei}	0.4035	0.0244	0.4789	0.0064	0.4740	0.0071	0.4725	0.0073	13.00
LI	0.5754	0.0007	0.5797	0.0006	0.5913	0.0005	0.5946	0.0004	7.50
h_{NS}	0.2883	0.1158	0.3394	0.0618	0.3478	0.0552	0.3648	0.0436	18.00
h_{agg}	0.5555	0.0012	0.5985	0.0004	0.6391	0.0001	0.6262	0.0002	6.25
h_S	0.1211	0.5165	0.2651	0.1496	0.2446	0.1848	0.3518	0.0523	20.75
h_{GE}	0.3572	0.0485	0.4101	0.0220	0.4179	0.0193	0.3800	0.0350	16.00
$h_{\text{LS-cos}}$	0.3444	0.0578	0.3953	0.0278	0.4094	0.0222	0.3788	0.0356	17.00
$h_{\text{LS-euc}}$	-0.1285	0.4909	-0.1101	0.5556	-0.1164	0.5329	-0.1168	0.5314	24.50
h_{attr}	-0.2164	0.2423	-0.1078	0.5637	-0.0778	0.6774	-0.1363	0.4647	24.50
h_{CF}	0.2543	0.1674	0.2531	0.1695	0.3005	0.1004	0.2907	0.1127	20.00
h_F	0.4169	0.0197	0.4860	0.0056	0.4967	0.0045	0.4704	0.0076	12.50
h_{KR}	-0.5229	0.0025	-0.3245	0.0749	-0.3540	0.0507	-0.3213	0.0780	27.00
h_{GNB}	-0.3625	0.0451	-0.1993	0.2824	-0.2418	0.1901	-0.1930	0.2982	26.00
h_{SVM}	0.2007	0.2790	0.2285	0.2162	0.3005	0.1005	0.2598	0.1581	21.00
$h^{\neg\mathcal{G}}_{\text{Tri-Hom}}$	0.5932	0.0004	0.6179	0.0002	0.6167	0.0002	0.6289	0.0002	5.25
$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.5593	0.0011	0.6657	$< 10^{-4}$	0.6510	$< 10^{-4}$	0.7036	$< 10^{-4}$	3.25
r_{deg}	0.1293	0.4882	0.2894	0.1143	0.2856	0.1194	0.3146	0.0848	20.25
$h^{\neg\mathcal{G}}_{\text{glocal}}$	0.5999	0.0004	0.6228	0.0002	0.6211	0.0002	0.6356	0.0001	3.75
$h^{\mathcal{G}}_{\text{glocal}}$	0.5655	0.0009	0.6922	$< 10^{-4}$	0.6799	$< 10^{-4}$	0.7168	$< 10^{-4}$	2.25

the correlation vanishes (0.02), so that effect is almost entirely one of density. **Other factors.** h_L dominates; h_S matters only for $h_L \geq 0.5$ (its generator control is weak); accuracy decreases with h_F, whereas Tri-Hom and our factor increase with it when $h_L > 1/C$. **Composite scores.** On the synthetic grids, the glocal factor and Tri-Hom predict accuracy about equally well: with the target r_{deg} the differences are at most 0.03 in either direction, and with the realized value the glocal factor is better in every setting, by 0.003–0.035 (Table 2, Appendix B.3). This is expected from the sign change above, which Proposition 2 explains: Eq. 4 always penalizes dispersion, which is correct for homophilic graphs but not for moderately heterophilic ones. In practice the penalty acts mostly where it is correct, since the Tri-Hom factor is near zero for $h_L \approx 1/C$ (Appendix Table 16). The synthetic study therefore supports the claim that degree dispersion affects GNN accuracy when the local-factor targets are fixed and density variation is reduced, but not the claim that the specific glocal formula captures this effect well; the evidence for the formula comes from real graphs.

6.2 Real-World Graphs

Setting. We use 31 real-world datasets covering homophilic and heterophilic regimes (Appendix C.1) and our reproduction of the MLP, GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), and GAT (Veličković et al., 2018) accuracies (Table 19); following the implementation of Platonov et al. (2023b), all three GNNs use residual connections and layer normalization (Appendix C.2); our conclusions are limited to them, and heterophily-oriented designs (e.g., H2GCN (Zhu et al., 2020), LINKX (Lim et al., 2021a)) may differ. We compare with 24 label-based, structural, feature-based, and classifier-based metrics and Tri-Hom (Appendix C.3). The fixed constants $C = 3$ and $\rho(\mathbf{A}) = 10$ are hyperparameters rather than graph properties, which weakens the theoretical reading of both factors; with dataset-specific C and $\rho(\mathbf{A})$ the glocal factor still exceeds Tri-Hom for all models (e.g., GCN: 0.652 vs. 0.612; Appendix D.6). We recomputed h_{edge}, h_{adj}, LI, and h_{agg} from their definitions after finding that the inherited values violated them (e.g., negative LI; Appendix C.3).

Table 4: Gain over Tri-Hom, $\Delta r = r(h^{\mathcal{G}}_{\text{glocal}}) - r(h^{\mathcal{G}}_{\text{Tri-Hom}})$. Family deletions: of the 6 single-family deletions, those with glocal $\geq$ Tri-Hom. Partial r: controls $\log N$, $\log C$, $\log M$, $\log \bar{d}$ and, for GNNs, MLP accuracy ($^*p < 0.01$). LOOCV / LOFO-CV R^2: out-of-sample R^2 of a linear calibration (leave-one-dataset-out / leave-one-family-out); "in-fold" parameters are selected on training folds.

		MLP	GCN	GraphSAGE	GAT
Significance	Δr	0.006	0.027	0.029	0.013
	Williams' test p (two-sided)	0.91	0.59	0.57	0.78
	95% bootstrap CI of Δr	[−0.09, 0.10]	[−0.03, 0.11]	[−0.02, 0.10]	[−0.04, 0.08]
Sensitivity	family deletions with glocal $\geq$ Tri-Hom	3/6	6/6	6/6	5/6
Partial r	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.544*	0.468	0.486	0.554*
	$h^{\mathcal{G}}_{\text{glocal}}$	0.548*	0.584*	0.585*	0.608*
	r_{deg}	0.147	0.559*	0.566*	0.519*
LOOCV R^2	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.215	0.365	0.346	0.425
	$h^{\mathcal{G}}_{\text{Tri-Hom}} + \alpha\, r_{\text{deg}}$ (α in-fold)	0.147	0.317	0.301	0.402
	OLS on $(h^{\mathcal{G}}_{\text{Tri-Hom}}, r_{\text{deg}})$	0.110	0.318	0.304	0.404
	$h^{\mathcal{G}}_{\text{glocal}}$, $\gamma = 0.2$	0.224	0.413	0.395	0.450
	$h^{\mathcal{G}}_{\text{glocal}}$, γ in-fold	0.194	0.395	0.369	0.445
LOFO-CV R^2	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.171	0.338	0.347	0.396
	$h^{\mathcal{G}}_{\text{glocal}}$, $\gamma = 0.2$	0.174	0.353	0.366	0.374
	$h^{\mathcal{G}}_{\text{glocal}}$, γ in-fold	−0.221	0.170	0.194	0.199

Correlation. $h^{\mathcal{G}}_{\text{glocal}}$ has the highest correlation with GCN, GraphSAGE, and GAT accuracy (0.692, 0.680, 0.717 vs. 0.666, 0.651, 0.704 for Tri-Hom; Table 3) and the best average rank, also under Spearman correlation; for MLP, h_{class} is best.

Significance and sensitivity. The gain over Tri-Hom is modest and, with 31 datasets and two highly correlated scores ($r = 0.93$), not statistically significant (Williams' test (Steiger, 1980) and bootstraps; Table 4). Deleting any one of the six dataset families leaves $h^{\mathcal{G}}_{\text{glocal}}$ at least as correlated as Tri-Hom in 17 of 18 cases, a sensitivity check rather than 18 independent confirmations. Non-significance is expected: Williams' test has 5–8% power with 31 datasets and would need roughly 400–800 datasets for 50–80% power, more for GAT (Appendix Table 25). Pooling the three GNNs does not change this (mean $\Delta r = 0.023$, 95% bootstrap CI $[-0.027, 0.099]$). Low power does not imply that an improvement exists; the data are simply uninformative about differences of this size.

Confounding. Datasets differ in size, number of classes, feature dimension, density, and intrinsic feature quality. Controlling for $\log N$, $\log C$, $\log M$, $\log \bar{d}$, and MLP accuracy, the advantage of the glocal factor widens (e.g., GCN: $0.468 \to 0.584$), and r_{deg} alone has a significant partial correlation with GNN accuracy (0.52–0.57), although its raw correlation is only 0.29–0.31. This result depends on the control set: across all 64 subsets of six candidate controls (Appendix D.2), the partial correlation of r_{deg} is always positive (0.21–0.60) but is significant in 91–100% of the specifications that include MLP accuracy and in at most 6% of those that do not; controlling for Tri-Hom alone gives 0.21–0.26 ($p > 0.17$). MLP accuracy measures what the features alone predict; it is unrelated to r_{deg} given Tri-Hom (partial $r = 0.03$), and adjusting only accuracy for it gives nearly the same values (semipartial 0.49–0.53), so it acts as a precision variable, not a suppressor (Appendix Table 28). Degree regularity is thus associated with how much a GNN gains beyond what the features predict; this is a conditional association, not a causal effect. The GNN−MLP gain as the target gives only weak support (Appendix Table 34).

Out-of-sample calibration. Is Eq. 4 better than using r_{deg} as a second covariate? Under LOOCV, the glocal factor gives the best out-of-sample R^2 for all four models, and adding r_{deg} linearly or through a two-variable regression is worse than Tri-Hom alone (Table 4); with 31 datasets a free coefficient costs more than it gains. The advantage survives selecting γ inside each training fold (e.g., GCN: 0.395 vs. 0.365). Across held-out families it is weaker (better for GCN and GraphSAGE, worse for GAT), and selecting γ from the remaining families is unstable, so the formula transfers only with γ in a reasonable range ($[0.01, 0.2]$; Appendix D.5). Since $\gamma = 0.2$ was set on the same 31 datasets, fixed-γ results are partly in-sample; the in-fold rows remove this advantage.

Design search and family structure. To correct for our design search, we permuted the degree statistics across datasets and reran the search over all 52 variants (Appendix Table 31). The gain

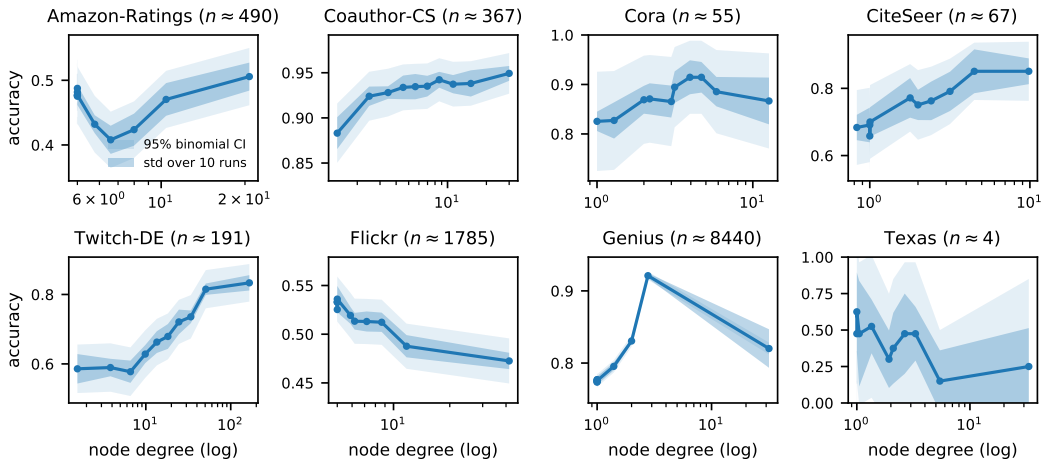


Figure 2: Degree-wise GCN accuracy from a separate experiment (Appendix Tables 44, 45) over ten equal-size degree groups of test nodes. Bands: std over 10 runs (dark), 95% binomial CI (light).

of our fixed variant beats its own null ($p = 0.003$) but only marginally the null maximum over all variants ($p = 0.06$; $p = 0.03$ for the best variant), so the evidence is suggestive rather than conclusive. Families make datasets far from independent: the family intraclass correlation of GNN accuracy is 0.37–0.44, i.e., roughly 11–12 effective datasets (Appendix Table 33). When degree statistics are permuted only within families, the gain is not significant ($p = 0.09$; 0.33 against the search maximum), and with family fixed effects r_{deg} alone is no longer a significant partial predictor, whereas the glocal factor still is and exceeds Tri-Hom (Table 32). Much of the degree signal therefore lies between benchmark families.

Is degree regularity special? Not uniquely. In the same slot, $1 - \text{Gini}$, $1/(1+\text{CV}^2)$, and the average clustering coefficient give GNN correlations within 0.013 of r_{deg} (Appendix D.5). As conditional predictors given Tri-Hom, dataset statistics, and MLP accuracy, however, clustering carries nothing ($|r| \leq 0.04$), whereas r_{deg} (0.52–0.56) and $1 - \text{Gini}$ (0.69–0.76) do, and with family fixed effects only $1 - \text{Gini}$ remains significant for all three GNNs (exploratory, post hoc; Appendix Table 39). The missing information is thus degree-based, consistent with Li et al. (2023b), and r_{deg} is a simple but not the strongest summary of it. We keep r_{deg} as pre-specified, since switching to Gini post hoc would enlarge the design search.

6.3 Does Degree Matter for Node Classification?

Degree matters, and not only through low-degree nodes. Figure 2 shows GCN accuracy per degree decile. A χ^2 test of equal accuracy across the groups (Bonferroni over datasets) rejects for six of eight datasets in at least half of the seeds, but not for Cora or Texas; it treats nodes as independent, so it is indicative only. Accuracy rises, falls, or peaks with degree depending on the dataset, so overall accuracy depends on the degree distribution.

7 Conclusion

We studied degree dispersion as a global complement to local homophily factors. A careful audit of 31 benchmarks suggests that degree dispersion adds information beyond local homophily. A calibrated degree-aware CSBM exposes a density confound in degree-varying generators and shows that, at nearly fixed density, the effect of dispersion changes sign with label homophily, as an elementary aggregation model explains. Added to Tri-Hom, degree regularity achieves the highest correlation with GCN, GraphSAGE, and GAT accuracy among 24 metrics. Global graph structure should therefore be considered alongside local homophily when characterizing datasets for GNNs.

AI USE STATEMENT

An LLM-based assistant was used for coding, running experiments, formatting LaTeX tables, reviewing proofs, and editing the manuscript. The authors independently verified all experimental results, mathematical arguments, and proofs, and take full responsibility for the final content.

ETHICS STATEMENT

This work studies how graph properties relate to the performance of graph neural networks, using publicly available benchmark datasets and synthetic graphs. It does not involve human subjects or the collection of new personal data.

REPRODUCIBILITY STATEMENT

Proofs are in Appendix A, numerical checks of the theory (DA-CSBM calibration against Lemma 1 and the aggregation model of Proposition 2) in Appendices B.2 and B.3, the DA-CSBM algorithm in Appendix B.1, datasets in Appendix C.1, training settings in Appendix C.2, and per-dataset metric values in Appendix G. All real-world analyses use only these values and the accuracies of Table 19, so they can be reproduced without retraining. Anonymized code to reproduce our experiments is provided in the supplementary material.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *international conference on machine learning*, pp. 21–29. PMLR, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1): 1017, 2019.

Andrea Cavallo, Claas Grohnfeldt, Michele Russo, Giulio Lovisotto, and Luca Vassio. 2-hop neighbor class similarity (2ncs): A graph structural metric indicative of graph neural network performance. *arXiv preprint arXiv:2212.13202*, 2022.

Yuhan Chen, Yihong Luo, Jing Tang, Liang Yang, Siya Qiu, Chuan Wang, and Xiaochun Cao. LSGNN: Towards general graph neural network in node classification by local similarity. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3550–3558, 2023.

Ernesto Estrada. Quantifying network heterogeneity. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(6):066102, 2010.

Shengbo Gong, Jiajun Zhou, Chenxuan Xie, and Qi Xuan. Neighborhood homophily-based graph convolutional network. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 3908–3912, 2023.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Di Jin, Rui Wang, Meng Ge, Dongxiao He, Xiang Li, Wei Lin, and Weixiong Zhang. Raw-gnn: Random walk aggregation based graph neural network. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pp. 2108–2114. International Joint Conferences on Artificial Intelligence, 2022.

Mingxuan Ju, Tong Zhao, Wenhao Yu, Neil Shah, and Yanfang Ye. Graphpatcher: mitigating degree bias for graph neural networks via test-time augmentation. *Advances in Neural Information Processing Systems*, 36:55785–55801, 2023.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 83(1):016107, 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Jérôme Kunegis and Julia Preusse. Fairness on the web: Alternatives to the power law. In *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 175–184, 2012.

Soo Yong Lee, Sunwoo Kim, Fanchen Bu, Jaemin Yoo, Jiliang Tang, and Kijung Shin. Feature distribution on graph topology mediates the effect of graph convolution: Homophily perspective. In *International Conference on Machine Learning*, pp. 26686–26714. PMLR, 2024.

Shouheng Li, Dongwoo Kim, and Qing Wang. Restructuring graph for higher homophily via adaptive spectral clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8622–8630, 2023a.

Ting Wei Li, Qiaozhu Mei, and Jiaqi Ma. A metadata-driven approach to understand graph neural networks. volume 36, pp. 15320–15340, 2023b.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in neural information processing systems*, 34:20887–20902, 2021a.

Derek Lim, Xiuyu Li, Felix Hohne, and Ser-Nam Lim. New benchmarks for learning on non-homophilous graphs. *arXiv preprint arXiv:2104.01404*, 2021b.

Renming Liu, Semih Cantürk, Frederik Wenkel, Sarah McGuire, Xinyi Wang, Anna Little, Leslie O'Bray, Michael Perlmutter, Bastian Rieck, Matthew Hirn, et al. Taxonomy of benchmarks in graph representation learning. In *Learning on Graphs Conference*, pp. 6–1. PMLR, 2022.

Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. Tail-gnn: Tail-node graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1109–1119, 2021.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. *Advances in neural information processing systems*, 35:1362–1375, 2022.

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. *Advances in Neural Information Processing Systems*, 36:28748–28760, 2023.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Mikhail Mironov and Liudmila Prokhorenkova. Revisiting graph homophily measures. In *Proceedings of the Third Learning on Graphs Conference*, volume 269 of *Proceedings of Machine Learning Research*, pp. 1:1–1:22. PMLR, 26–29 Nov 2025.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

John Palowitch, Anton Tsitsulin, Brandon Mayer, and Bryan Perozzi. Graphworld: Fake graphs bring real insights for gnns. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3691–3701, 2022.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 36:523–548, 2023a.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? 2023b.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2): 245, 1980.

Arjun Subramonian, Jian Kang, and Yizhou Sun. Theoretical and empirical insights into the origins of degree bias in graph neural networks. *Advances in Neural Information Processing Systems*, 37: 8193–8239, 2024.

Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Yiqi Wang, Jiliang Tang, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. Investigating and mitigating degree-related biases in graph convoltuional networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1435–1444, 2020.

Louis Emiel T Van Langendonck, Guillermo Bernárdez, Nina Miolane, and Pere Barlet-Ros. Graphuniverse: Synthetic graph generation for evaluating inductive generalization. In *International Conference on Learning Representations*, volume 2026, pp. 138611–138642, 2026.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Junfu Wang, Yuanfang Guo, Liang Yang, and Yunhong Wang. Understanding heterophily for graph neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50489–50529. PMLR, 21–27 Jul 2024.

Liang Yang, Mengzhe Li, Liyang Liu, Bingxin Niu, Chuan Wang, Xiaochun Cao, and Yuanfang Guo. Diverse message passing for attribute with heterophily. *Advances in Neural Information Processing Systems*, 34:4751–4763, 2021.

Renchi Yang, Jieming Shi, Xiaokui Xiao, Yin Yang, Sourav S Bhowmick, and Juncheng Liu. Pane: scalable and effective attributed network embedding. *The VLDB Journal*, 32(6):1237–1262, 2023.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Mustafa Yasir, John Palowitch, Anton Tsitsulin, Long Tran-Thanh, and Bryan Perozzi. Examining the effects of degree distribution and homophily in graph learning models. *arXiv preprint arXiv:2307.08881*, 2023.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJe8pkHFwS.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

Zhili Zhao, Li Wan, Xupeng Liu, Ruiyi Yan, and Shaomeng Wang. Node classification in gnns: Impact of neighborhood label distribution on homophily and heterophily. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Yilun Zheng, Sitao Luan, and Lihui Chen. What is missing for graph homophily? disentangling graph homophily for graph neural networks. *Advances in Neural Information Processing Systems*, 37:68406–68452, 2024.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

A PROOFS

Proof of Proposition 1. Conditioned on the graph and the labels, $\bar{x}_i - \frac{1}{d_i}\sum_{j\in\mathcal{N}(v_i)} m_{y_j}$ is an average of d_i independent noise terms, so $\mathrm{Var}(\bar{x}_i \mid \mathcal{G}, \mathbf{y}) = \sigma^2/d_i$ and $\bar{v} = \sigma^2\,\frac{1}{N}\sum_i d_i^{-1}$. For every $d > 0$ and $\mu > 0$,

$$\frac{1}{\mu} - \frac{d-\mu}{\mu^2} + \frac{(d-\mu)^2}{\mu^2 d} = \frac{\mu d - d(d-\mu) + (d-\mu)^2}{\mu^2 d} = \frac{\mu^2}{\mu^2 d} = \frac{1}{d}.$$

Averaging over nodes with $\mu = \frac{1}{N}\sum_i d_i$ removes the linear term, giving the exact identity

$$\frac{1}{N}\sum_i \frac{1}{d_i} = \frac{1}{\mu} + \frac{1}{\mu^2}\cdot\frac{1}{N}\sum_i \frac{(d_i-\mu)^2}{d_i}.$$

Since $d_{\min} \le d_i \le d_{\max}$, the last average lies between $\mathrm{Var}(\mathbf{d})/d_{\max}$ and $\mathrm{Var}(\mathbf{d})/d_{\min}$. Substituting $\mathrm{Var}(\mathbf{d}) = \mu^2\mathrm{CV}^2(\mathbf{d})$ gives Eq. 3. If the graph is regular, $\mathrm{Var}(\mathbf{d}) = 0$ and both bounds equal σ^2/μ. If it is not regular, some node has $d_i < \mu < d_{\max}$ and some node has $d_i > \mu > d_{\min}$, so both inequalities are strict. □

Parts (a), (b), and the Gamma case. Write $\bar{v} = \sigma^2\,\mathbb{E}[1/d]$ and $\mathrm{Var}(d) = \mathbb{E}[d^2] - \mu^2$, with expectations over the (empirical or population) degree distribution. (a) If d' dominates d in the convex order with $\mathbb{E}[d'] = \mathbb{E}[d] = \mu$, then $\mathbb{E}[\phi(d')] \ge \mathbb{E}[\phi(d)]$ for every convex ϕ; applying this to the convex functions $\phi(x) = 1/x$ on $(0,\infty)$ and $\phi(x) = x^2$ gives $\bar{v}' \ge \bar{v}$ and $\mathrm{CV}^2(d') \ge \mathrm{CV}^2(d)$. (b) For $d \sim \mathrm{LogNormal}(\mu', s^2)$ with $\mathbb{E}[d] = e^{\mu'+s^2/2} = \mu$ and $\mathrm{CV}^2 = e^{s^2} - 1$, we have $\mathbb{E}[1/d] = e^{-\mu'+s^2/2} = e^{s^2}/\mu = (1+\mathrm{CV}^2)/\mu$. Gamma case: for $d \sim \mathrm{Gamma}(k,\theta)$ with $k\theta = \mu$ and $\mathrm{CV}^2 = 1/k < 1$, $\mathbb{E}[1/d] = \frac{1}{\theta(k-1)} = \frac{1}{\mu(1-\mathrm{CV}^2)}$. □

No monotonicity in general. For $\mathbf{d} = (1,1,3,3,3,3)$ and $\mathbf{d}' = (1,2,2,2,2,5)$ (both graphical, mean $7/3$), $\mathrm{CV}^2(\mathbf{d}) = 8/49 < \mathrm{CV}^2(\mathbf{d}') = 2/7$, while $\frac{1}{N}\sum_i d_i^{-1} = 5/9 > 8/15 = \frac{1}{N}\sum_i d_i'^{-1}$. Neither sequence is a mean-preserving spread of the other, so this does not contradict part (a).

The same argument applies to sum aggregation followed by degree normalization with self-loops, with d_i replaced by $d_i + 1$.

Remark 1 (Noise of the GCN operator)**.** *Let* $\mathbf{P} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ *with* $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ *and* $\tilde{d}_i = d_i + 1$*, and let the feature noise be as in Proposition 1. The average per-coordinate noise variance of* $\mathbf{PX}$ *is exactly*

$$\bar{v}_{\mathrm{gcn}} = \frac{\sigma^2}{N}\|\mathbf{P}\|_F^2 = \frac{\sigma^2}{N}\sum_i \Big(\frac{1}{\tilde{d}_i^2} + \frac{1}{\tilde{d}_i}\sum_{j\in\mathcal{N}(v_i)}\frac{1}{\tilde{d}_j}\Big).$$

If the neighbors of a node are reached by edges whose other endpoint has the size-biased degree distribution $d\,p(d)/\mu$*, independently of the node's own degree (configuration-model approximation, no degree assortativity), then* $\mathbb{E}\big[\sum_{j\in\mathcal{N}(v_i)}\tilde{d}_j^{-1} \mid d_i\big] = d_i\,\mathbb{E}[d/(d+1)]/\mu$*, and hence*

$$\bar{v}_{\mathrm{gcn}} \approx \sigma^2 Q, \qquad Q = \mathbb{E}\big[(d+1)^{-2}\big] + \frac{\mathbb{E}[d/(d+1)]^2}{\mu},$$

with $Q = 1/(\mu+1)$ *for a regular graph.* Q *depends on the degree distribution through the inverse-degree moments* $\mathbb{E}[(d+1)^{-2}]$ *and* $\mathbb{E}[(d+1)^{-1}]$*, not through* CV^2*, which is dominated by the upper tail.*

Table 5: Normalized excess noise of the GCN operator: exact ν_{gcn} and the approximation $(\mu+1)Q-1$ of Remark 1.

Dataset	μ	exact	approx.	Dataset	μ	exact	approx.
Roman-Empire	2.9	0.01	0.01	Amazon-Photo	31.1	1.06	0.83
Amazon-Ratings	7.6	−0.04	0.02	Amazon-Computer	35.8	1.46	1.15
Minesweeper	7.9	0.00	0.00	Coauthor-CS	8.9	0.31	0.19
Tolokers	88.3	2.04	1.59	Coauthor-Physics	14.4	0.45	0.20
Questions	6.3	0.46	0.63	WikiCS	36.9	2.27	2.02
Squirrel	42.3	2.29	0.98	Blog-Catalog	66.1	0.03	0.02
Chameleon	19.9	1.08	0.48	Ogbn-Arxiv	13.7	0.89	0.54
Actor	7.0	0.34	0.34	Genius	4.4	0.11	0.44
Texas	3.0	−0.00	0.13	Twitch-DE	32.2	0.69	0.68
Cornell	3.0	0.01	0.14	Twitch-ENGB	9.9	0.44	0.45
Wisconsin	3.6	0.04	0.13	Twitch-ES	25.6	0.66	0.66
Cora	3.9	0.12	0.11	Twitch-FR	34.4	0.55	0.61
CoraFull	6.4	0.33	0.25	Twitch-PTBR	32.7	0.67	0.74
CiteSeer	2.7	0.21	0.14	Twitch-RU	17.0	0.55	0.59
PubMed	4.5	0.19	0.32	Twitch-TW	45.8	0.86	0.87
Flickr	10.1	−0.14	0.04				

Derivation. The first identity holds because $(\mathbf{P}\boldsymbol{\xi})_i = \sum_j P_{ij}\xi_j$ with independent ξ_j, so its variance is $\sigma^2 \sum_j P_{ij}^2 = \sigma^2 \sum_j \tilde{A}_{ij}/(\tilde{d}_i\tilde{d}_j)$. Under the approximation, $\mathbb{E}\big[\sum_{j\in\mathcal{N}(v_i)} \tilde{d}_j^{-1} \mid d_i\big] = d_i \sum_d \frac{d\,p(d)}{\mu}\frac{1}{d+1}$, and $\frac{1}{N}\sum_i \frac{d_i}{d_i+1}\cdot\frac{\mathbb{E}[d/(d+1)]}{\mu} = \frac{\mathbb{E}[d/(d+1)]^2}{\mu}$. Table 5 compares the normalized excess $(\mu+1)Q-1$ with its exact counterpart ν_{gcn} on the 31 graphs: Pearson 0.91, Spearman 0.93; the approximation underestimates the noise most on Squirrel and Chameleon, and across graphs the residual has Spearman 0.60 with degree assortativity, which the approximation ignores. Because the approximation is a function of inverse-degree moments, it is highly rank-correlated with ν (0.95) but not with CV^2 (0.36), which explains why CV^2 does not track the noise of the GCN operator on real graphs. □

Proof of Proposition 2. With exactly $\frac{1+a}{2}d_i$ same-class and $\frac{1-a}{2}d_i$ other-class neighbors, $\bar{x}_i$ given $y_i=+1$ is Gaussian with mean $m\,\frac{1+ad_i}{d_i+1}$ and variance $\frac{\sigma^2}{d_i+1}$, so the sign rule is correct with probability $f_a(d_i)=\Phi(g(d_i))$, $g(d)=s\frac{1+ad}{\sqrt{d+1}}$ (the case $y_i=-1$ is symmetric). Since $\Phi''(x)=-x\varphi(x)$, $f_a'' = \varphi(g)\,(g''-g\,g'^2)$. For $a=1$, $g=s\sqrt{d+1}>0$ and $g''=-\frac{s}{4}(d+1)^{-3/2}<0$, so $f_1''<0$. For $a=0$, $g=s(d+1)^{-1/2}$, $g'=-\frac{s}{2}(d+1)^{-3/2}$, $g''=\frac{3s}{4}(d+1)^{-5/2}$, hence $f_0'' = \varphi(g)\,\frac{s}{4}(d+1)^{-5/2}\big(3-\frac{s^2}{d+1}\big)>0$ whenever $s^2<3(d+1)$, in particular on $[d_{\min},\infty)$ if $s^2<3(d_{\min}+1)$. A mean-preserving spread lowers the mean of a concave function and raises the mean of a convex one (convex order, as in Proposition 1(a)). □

Table 6: Effect of degree regularity on GCN accuracy (percentage points; regular minus dispersed) measured on the Eq. 7 grid and predicted by the model of Proposition 2 with log-normal degrees at the realized dispersion levels ($r_{\mathrm{deg}}=0.47$ vs. 0.97).

h_L	a	measured	model ($s=0.5$)	model ($s=1$)
0.0	−0.50	0.2	−11.5	−19.9
0.1	−0.35	−1.4	−9.0	−17.7
0.2	−0.20	−2.7	−6.2	−12.7
0.3	−0.05	−3.7	−3.6	−6.8
0.4	0.10	−1.8	−1.1	−1.9
0.5	0.25	−0.3	1.3	2.2
0.6	0.40	2.4	3.6	4.3
0.7	0.55	3.4	5.4	4.5
0.8	0.70	3.7	6.7	3.8
0.9	0.85	3.7	7.3	3.0
1.0	1.00	3.6	7.5	2.2

Table 7: Exploratory: the model of Proposition 2 on the 31 real graphs (each graph's degree sequence, $a = (h_{\text{node}}C - 1)/(C - 1)$, one global s). Correlation of the predicted accuracy and of its dispersion component with GNN accuracy, and partial correlation of the dispersion component given $h^{\mathcal{G}}_{\text{Tri-Hom}}$.

s	C	Model	r(model acc.)	r(dispersion comp.)	partial given Tri-Hom
0.5	C=3	GCN	0.650	−0.485	−0.205
0.5	C=3	GraphSAGE	0.582	−0.427	−0.124
0.5	C=3	GAT	0.657	−0.512	−0.225
0.5	true C	GCN	0.716	−0.486	−0.105
0.5	true C	GraphSAGE	0.680	−0.458	−0.071
0.5	true C	GAT	0.735	−0.538	−0.160
1.0	C=3	GCN	0.627	−0.504	−0.336
1.0	C=3	GraphSAGE	0.557	−0.450	−0.261
1.0	C=3	GAT	0.626	−0.518	−0.350
1.0	true C	GCN	0.677	−0.335	−0.060
1.0	true C	GraphSAGE	0.630	−0.282	0.010
1.0	true C	GAT	0.696	−0.394	−0.129

Table 8: Effect of degree regularity (percentage points; regular minus dispersed) in the C-class model of Appendix A, at the degree distributions of Table 6; f_3 estimated by Monte Carlo (2×10^5 draws per grid point).

		$C = 2$		$C = 3$	
h_L	a	$s = 0.5$	$s = 1$	$s = 0.5$	$s = 1$
0.0	−0.50	−11.5	−19.9	−9.1	−14.5
0.1	−0.35	−9.0	−17.7	−7.5	−14.1
0.2	−0.20	−6.2	−12.7	−5.5	−11.4
0.3	−0.05	−3.6	−6.9	−3.4	−7.0
0.4	0.10	−1.1	−1.9	−1.2	−2.2
0.5	0.25	1.4	2.2	1.3	2.8
0.6	0.40	3.6	4.3	3.9	6.6
0.7	0.55	5.5	4.5	6.4	8.2
0.8	0.70	6.7	3.8	8.6	8.0
0.9	0.85	7.3	3.0	10.3	7.0
1.0	1.00	7.5	2.2	11.4	5.7

C classes and the case $a < 0$. Let the class means be centered, $\sum_c \mathbf{m}_c = \mathbf{0}$, with equal norms and equal pairwise angles (a regular simplex), and let node v_i have $h_L d_i$ same-class neighbors and $\frac{1-h_L}{C-1} d_i$ neighbors in each other class. The mean of the other classes is $-\mathbf{m}_{y_i}/(C-1)$, so $\mathbb{E}[\bar{\mathbf{x}}_i] = \mathbf{m}_{y_i}\big(1 + d_i(h_L - \frac{1-h_L}{C-1})\big)/(d_i+1) = \mathbf{m}_{y_i}(1 + a d_i)/(d_i+1)$ with $a = \frac{h_L C - 1}{C-1}$, and the noise covariance is $\frac{\sigma^2}{d_i+1}\mathbf{I}$. The nearest-mean rule is therefore correct with probability $f_C\big(s\frac{1+ad_i}{\sqrt{d_i+1}}\big)$ for a function f_C that depends only on C, with $f_2 = \Phi$. *A general criterion.* Write $F = f_C$ and $g_a(d) = s\frac{1+ad}{\sqrt{d+1}}$. F is strictly increasing on $[0,\infty)$: the region where the rule is correct, $K = \{\mathbf{h} : \langle \mathbf{h}, \mathbf{m}_y - \mathbf{m}_c \rangle > 0 \ \forall c \neq y\}$, is a convex cone containing $\mathbf{m}_y$, so for every noise draw the set of t with $t\mathbf{m}_y/\|\mathbf{m}_y\| + \boldsymbol{\zeta} \in K$ is an interval (t_0, ∞). Since $(F \circ g)'' = F'(g)\,[g'' + (F''/F')(g)\,g'^2]$ with $g_1''/g_1'^2 = -1/g_1$ and $g_0''/g_0'^2 = 3/g_0$, the accuracy is strictly concave in d for $a = 1$ if and only if $\kappa_C(t) := -tF''(t)/F'(t) > -1$ for $t \geq s\sqrt{d_{\min}+1}$, and strictly convex for $a = 0$ if and only if $\kappa_C(t) < 3$ for $0 < t \leq s/\sqrt{d_{\min}+1}$. For $C = 2$, $F = \Phi$ and $\kappa_2(t) = t^2$, which recovers Proposition 2.

The case $C = 3$ in closed form. Let $s = \|\mathbf{m}_c\|/\sigma$ and $\boldsymbol{\zeta}$ be standard normal. The rule is correct if and only if $\langle \boldsymbol{\zeta}, \mathbf{u}_c \rangle > -\tau$ for both other classes, where $\mathbf{u}_c = (\mathbf{m}_y - \mathbf{m}_c)/\|\mathbf{m}_y - \mathbf{m}_c\|$, $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \frac{1}{2}$, and $\tau = \frac{\sqrt{3}}{2}t$. Hence $F_3(t) = \Phi_2(\tau, \tau; \frac{1}{2})$, the bivariate normal distribution function with correlation $\frac{1}{2}$, $F_3'(t) \propto \varphi(\tau)\,\Phi(k\tau)$ with $k = 1/\sqrt{3}$, and

$$\kappa_3(t) = \tau^2 - \frac{k\tau\,\varphi(k\tau)}{\Phi(k\tau)}.$$

Since $x\varphi(x)/\Phi(x) < 0.3$ for $x \geq 0$, $\kappa_3 > \tau^2 - 0.3 > -1$ for every $t \geq 0$, so the $a = 1$ statement holds for every s; and since $\kappa_3 < \tau^2 < 3$ whenever $t < 2$, the $a = 0$ statement holds if $s^2 < 4(d_{\min}+1)$ (compared with $3(d_{\min}+1)$ for $C = 2$). This closed form of F_3 agrees with

Monte Carlo to within 4×10^{-4}, and Table 8 shows the resulting effects, with the same sign pattern and crossover as for $C = 2$. For $C > 3$, F_C is an orthant probability of an equicorrelated Gaussian, and we have not verified the criterion. For $a < 0$ the sign of the aggregated mean flips at $d_i = 1/|a|$, so a fixed nearest-mean (or sign) rule misclassifies high-degree nodes systematically. A trained GNN is not bound to this rule, and the model overstates the negative effect of regularity for $a < 0$ by an order of magnitude (Table 6).

Proof of Lemma 1. Taking the expectation over the zero-mean noise $\mathbf{E}$ and over labels drawn uniformly and independently of $\mathbf{w}$, $\mathbb{E}[D^{\mathcal{N}}_{i,y_j}] = \frac{1}{C}\sum_c S_{y_i c} = \frac{1}{C}$ because every row of $\mathbf{S}$ sums to one. Without clipping, Eq. 7 therefore gives $\mathbb{E}[d_i \mid \mathbf{w}] = \sum_j \frac{C}{N\bar{w}} w_i w_j \cdot \frac{1}{C} = w_i \cdot \frac{1}{N\bar{w}} \sum_j w_j \approx w_i$. For the $\sqrt{w}$ variant, $\mathbb{E}[A_{ij} \mid \mathbf{w}, \mathbf{y}, \mathbf{E}] = \frac{C}{N}\sqrt{w_i w_j}\, D^{\mathcal{N}}_{i,y_j}$, hence $\mathbb{E}[d_i \mid \mathbf{w}] = \sum_j \frac{C}{N}\sqrt{w_i w_j} \cdot \frac{1}{C} = \sqrt{w_i}\,\overline{\sqrt{w}}$. These expected degrees are proportional to $\sqrt{w_i}$, so their squared coefficient of variation is $\mathrm{Var}(\sqrt{w})/\mathbb{E}[\sqrt{w}]^2 = \mathbb{E}[w]/\mathbb{E}[\sqrt{w}]^2 - 1$. For $w \sim \mathrm{LogNormal}(\mu', \sigma^2)$, $\mathbb{E}[w] = e^{\mu' + \sigma^2/2}$ and $\mathbb{E}[\sqrt{w}] = e^{\mu'/2 + \sigma^2/8}$, so the ratio is $e^{\sigma^2/4} = (1 + \mathrm{CV}_{\mathrm{t}}^2)^{1/4}$, and the mean expected degree is $\mathbb{E}[\sqrt{w}]^2 = \bar{w}\, e^{-\sigma^2/4}$. Bernoulli sampling adds variance approximately equal to the mean degree, which contributes about $1/\bar{d}_{\text{real}}$ to the realized CV^2; this term is included in the prediction in Table 1. □

B Synthetic Experiments

B.1 DA-CSBM Details

Algorithm 1 Degree-Aware Contextual Stochastic Block Model (DA-CSBM)

Require: Label homophily h_L, structural homophily h_S, feature homophily h_F, degree regularity $r_{\deg}$, number of nodes N, number of classes C, feature dimension M, mean degree weight $\bar{w}$, random seed ϕ
Ensure: Adjacency matrix $\mathbf{A}$, node features $\mathbf{X}$, and node labels $\mathbf{Y}$

Initialize node labels $\mathbf{Y} \in \mathbb{R}^N$ and node features $\mathbf{X} \in \mathbb{R}^{N \times M}$
Set random seed ϕ
for $n = 1$ to N **do**
  $k \sim \lfloor \mathrm{Uniform}(0, C) \rfloor$
  $Y_n \leftarrow k$
end for ▷ Sample node labels
$\mathbf{Z} \leftarrow \mathrm{OneHot}(\mathbf{Y})$ ▷ One-hot encoding of labels
$\mathrm{CV}_{\mathrm{t}}^2 \leftarrow -\log(r_{\deg})/\gamma$; $\sigma^2 \leftarrow \log(1 + \mathrm{CV}_{\mathrm{t}}^2)$; $\mu' \leftarrow \log \bar{w} - \sigma^2/2$
$w_n \sim \mathrm{LogNormal}(\mu', \sigma^2)$, rounded and clipped to $[1, 0.2N]$, for $n = 1, \ldots, N$ ▷ Degree weights (Section 5)
$\mathbf{\Theta} \leftarrow \mathrm{diag}(\mathbf{w})/\sqrt{\bar{w}}$ ▷ $\sqrt{w}$ variant: $\mathbf{\Theta} \leftarrow \mathrm{diag}(\mathbf{w})^{1/2}$
$\mathbf{S} \leftarrow h_L \mathbf{I}_C + \frac{1-h_L}{C-1}(\mathbf{1}_C - \mathbf{I}_C)$ ▷ Introduce label homophily
$\mathbf{D}^{\mathcal{N}} \leftarrow \mathbf{ZS}$ ▷ Construct neighborhood distributions
for $D^{\mathcal{N}}_{u,c} \in \mathbf{D}^{\mathcal{N}}$ **do**
  $\epsilon \sim \mathrm{Normal}\left(0, \frac{(1-h_S)^2}{C-1}\right)$
  $D^{\mathcal{N}}_{u,c} \leftarrow D^{\mathcal{N}}_{u,c} + \epsilon$
end for ▷ Introduce structural homophily
$\mathbf{A}_p \leftarrow \frac{C}{N}\mathbf{\Theta}\mathbf{D}^{\mathcal{N}}\mathbf{Z}^\top\mathbf{\Theta}$
$\mathbf{A}_p \leftarrow \max(0, \min(1, \mathbf{A}_p))$ ▷ Bound probabilities to $[0, 1]$
$\mathbf{A} \leftarrow \langle \mathrm{Sym} \circ \mathrm{Binarize} \rangle(\mathbf{A}_p)$ ▷ Bernoulli for $i \le j$, mirrored; diagonal discarded
for $m = 1$ to M **do**
  Sample class-wise Gaussian means $\mathbf{m}$ and variances $\mathbf{s}^2$
  $\mathbf{X}^{(0)}_{:,m} \sim \mathrm{Normal}(\mathbf{Zm}, \mathrm{diag}(\mathbf{Zs}^2))$
end for ▷ Sample structural-agnostic node features
$\mathbf{X} \leftarrow \left(\mathbf{I} - \frac{h_F}{\rho(\mathbf{A})}\mathbf{A}\right)^{-1} \mathbf{X}^{(0)}$ ▷ Introduce feature homophily
return $\mathbf{A}, \mathbf{X}, \mathbf{Y}$

Relation to DC-SBM. The DC-SBM (Karrer & Newman, 2011) sets $\mathbb{E}[A_{ij}] = \theta_i \theta_j \omega_{y_i y_j}$. Eq. 7 has this form with $\theta_i = w_i/\sqrt{\bar{w}}$ (and the $\sqrt{w}$ variant with $\theta_i = \sqrt{w_i}$), except that the block affinity $D^{\mathcal{N}}_{i,y_j}$ depends on node i through the node-wise noise that implements structural homophily. It differs

from GraphWorld (Palowitch et al., 2022) in parameterizing degree dispersion by a target $r_{\deg}$ and in generating features by graph diffusion (feature homophily), as in CSBM-3H.

Negative entries and symmetrization. The Gaussian perturbation can make entries of $\mathbf{D}^{\mathcal{N}}$ negative or rows not sum to one. We do not renormalize $\mathbf{D}^{\mathcal{N}}$; negative probabilities are set to zero by the clipping in Eq. 7, which makes the realized structural homophily differ from its target. The graph is made undirected by sampling the upper triangle and mirroring it. Both effects are included in the calibration below, which measures statistics on the final simple graph.

B.2 CALIBRATION

Table 9: DA-CSBM calibration of degree regularity (mean $\pm$ std over h_L, h_S, and 3 seeds; 132 graphs per cell).

target $r_{\deg}$	realized (Eq. 7)	$\bar{d}_{\text{real}}$	realized ($\sqrt{w}$ variant)	$\bar{d}_{\text{real}}$
0.1	0.476 ± 0.021	10.1	0.841 ± 0.012	6.8
0.2	0.512 ± 0.019	10.6	0.854 ± 0.011	7.3
0.3	0.546 ± 0.018	11.0	0.866 ± 0.009	7.7
0.4	0.582 ± 0.016	11.3	0.878 ± 0.009	8.1
0.5	0.621 ± 0.014	11.5	0.890 ± 0.008	8.5
0.6	0.666 ± 0.013	11.8	0.902 ± 0.008	9.0
0.7	0.724 ± 0.012	12.0	0.916 ± 0.007	9.5
0.8	0.796 ± 0.010	12.2	0.932 ± 0.006	10.2
0.9	0.880 ± 0.009	12.3	0.950 ± 0.005	11.1
1.0	0.973 ± 0.005	12.4	0.973 ± 0.005	12.4

Table 10: DA-CSBM calibration of the local factors (Eq. 7; h_F measured on the $\sqrt{w}$ variant). Realized h_L: node homophily; h_S: Tri-Hom structural homophily; h_F: graph-level estimate of Zheng et al. (2024).

	target / realized value					
target h_L	0.0	0.2	0.4	0.6	0.8	1.0
realized h_L	0.107	0.229	0.381	0.534	0.675	0.787
target h_S	0.2	0.4	0.6	0.8		
realized h_S	0.482	0.493	0.517	0.557		
target h_F	-0.8	-0.4	0.0	0.4	0.8	
realized h_F	-1.00	-0.53	0.02	0.46	0.82	

For every target $r_{\deg} \in \{0.1, \ldots, 1.0\}$, $h_L \in \{0, 0.1, \ldots, 1\}$, $h_S \in \{0.2, 0.4, 0.6, 0.8\}$, and 3 seeds, we generated graphs with Eq. 7 and with the $\sqrt{w}$ variant (2,640 graphs), and 45 further graphs with $h_F \in \{-0.8, \ldots, 0.8\}$ to measure realized feature homophily. Tables 9 and 10 report the results. The generator code with both parametrizations and the calibration script are included in the supplementary material.

Table 11: Realized mean degree of DA-CSBM graphs ($\sqrt{w}$ variant; $\bar{w} = 10.5$, $N = 1000$; mean over 3 seeds) by target h_L (rows) and target h_S (columns). Eq. 7 gives nearly the same pattern.

	target $r_{\deg} = 0.1$, $h_S =$				target $r_{\deg} = 1.0$, $h_S =$			
h_L	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
0.0	7.7	7.1	6.5	6.1	14.2	12.9	11.8	11.1
0.2	7.5	6.8	6.1	5.8	13.8	12.4	11.3	10.7
0.3	7.4	6.7	6.1	5.8	13.7	12.2	11.2	10.6
0.4	7.5	6.7	6.1	5.8	13.7	12.3	11.2	10.6
0.6	7.6	6.9	6.2	5.9	14.0	12.7	11.5	10.7
0.8	8.0	7.2	6.6	6.0	14.6	13.3	12.0	11.0
1.0	8.4	7.7	7.1	6.4	15.4	14.1	12.9	11.7

Table 12: Gap between the mean degree of Lemma 1 and the realized mean degree ($\sqrt{w}$ variant). Columns add, from left to right: rounding and clipping of the weights to $[1, 0.2N]$; clipping of $\mathbf{A}_p$ to $[0, 1]$ (almost entirely from negative entries); Bernoulli sampling, symmetrization, and removal of the diagonal.

target $r_{\deg}$	Lemma 1	+ weight rounding	+ clipping $\mathbf{A}_p$	realized
0.1	5.58	5.83	6.80	6.85
0.4	6.83	6.90	8.06	8.12
0.7	8.13	8.09	9.48	9.54
1.0	10.50	10.49	12.41	12.51

Table 11 reports the realized mean degree by target h_L and h_S, and Table 12 decomposes the gap between Lemma 1 and the realized mean degree. Setting negative entries of $\mathbf{A}_p$ to zero raises the expected edge probability; the amount clipped grows with the noise level $1 - h_S$ and with the number of zero entries of $\mathbf{S}$ (two for $h_L = 1$, one for $h_L = 0$, none for $h_L = 1/C$), which explains why density falls with h_S and is U-shaped in h_L. Probabilities above one are negligible, and the ratio of realized to predicted mean degree varies only between 1.17 and 1.23.

B.3 Re-Runs with Both Degree Parametrizations

Table 13: Additional statistics of the synthetic re-runs: Spearman correlations of the factors and Pearson correlations of the composite scores with accuracy.

	Eq. 7		$\sqrt{w}$ variant	
	all h_L	$h_L \geq 0.5$	all h_L	$h_L \geq 0.5$
h_L (Spearman)	0.801	0.994	0.845	0.998
h_S (Spearman)	0.064	0.763	−0.008	0.731
h_F (Spearman)	−0.781	−0.684	−0.874	−0.798
$r_{\deg}$ (Spearman)	0.220	0.573	0.759	0.860
$\bar{d} \mid r_{\deg}$ (partial, realized)	−0.141	−0.125	0.185	0.172
corr(target $r_{\deg}$, realized $\bar{d}$)	0.434	0.445	0.833	0.829
$h^{\mathcal{G}}_{\text{Tri-Hom}}$ (Pearson)	0.844	0.895	0.809	0.879
$h^{\mathcal{G}}_{\text{glocal}}$ (Pearson)	0.817	0.873	0.791	0.868
$h^{\mathcal{G}}_{\text{glocal}}$, realized (Pearson)	0.847	0.920	0.812	0.887

Table 14: Effect of degree regularity on GCN accuracy by label homophily: accuracy at target $r_{\deg} = 1.0$ minus accuracy at 0.1 (percentage points, averaged over h_S and h_F; positive: regular graphs are easier).

h_L	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Eq. 7	0.2	−1.4	−2.7	−3.7	−1.8	−0.3	2.4	3.4	3.7	3.7	3.6
$\sqrt{w}$ variant	8.6	7.6	5.1	2.6	3.3	4.0	4.9	5.0	4.3	3.4	2.7

The synthetic grid was trained twice with the same batched implementation: once with Eq. 7 (DC weighting, $\theta_i \propto w_i$) and once with the $\sqrt{w}$ variant ($\theta_i = \sqrt{w_i}$). The $\sqrt{w}$ variant was the parametrization of our preliminary version, whose original run is shown in Appendix B.4. The batched implementation trains the 60 models of one configuration (6 hyperparameter settings $\times$ 10 graphs) simultaneously with dense batched matrix products; because Adam and the coupled L_2 penalty are elementwise, this is equivalent to training the models one by one. The batched $\sqrt{w}$-variant run reproduces the accuracies of the original implementation on all 2,200 configurations (correlation 0.993, mean absolute difference 0.009). Graphs are generated with per-configuration seeds, the realized mean degree and $r_{\deg}$ of every graph are recorded, and all other settings follow Appendix C.2. Table 13 gives additional statistics, and Table 14 the effect of $r_{\deg}$ by label homophily. The full grids of the original run are shown in Appendix B.4.

Table 15: Eq. 7 with the upper-triangle rule vs. a symmetric edge-probability matrix: realized $r_{\deg}$ and mean degree at four targets, range of realized h_S, and Spearman(target, realized h_S).

Rule	realized $r_{\deg}$				realized $\bar{d}$				h_S range	$\rho_S(h_S)$
upper triangle (used)	0.478	0.582	0.724	0.974	9.8	11.0	11.8	12.7	0.489–0.572	0.27
$(\mathbf{A}_p + \mathbf{A}_p^\top)/2$	0.471	0.576	0.722	0.973	9.2	10.3	11.0	11.9	0.476–0.583	0.32

Table 16: Where the degree term acts (Eq. 7 grid, realized $r_{\deg}$): mean Tri-Hom factor, mean reduction of the factor by the degree term, and the accuracy effect of degree regularity (percentage points).

h_L	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	reduction by degree term	accuracy effect
0.0	0.181	0.023	0.2
0.1	0.112	0.018	−1.4
0.2	0.046	0.009	−2.7
0.3	0.004	0.001	−3.7
0.4	0.017	0.005	−1.8
0.5	0.115	0.032	−0.3
0.6	0.276	0.073	2.4
0.7	0.440	0.099	3.4
0.8	0.572	0.105	3.7
0.9	0.666	0.096	3.7
1.0	0.731	0.082	3.6

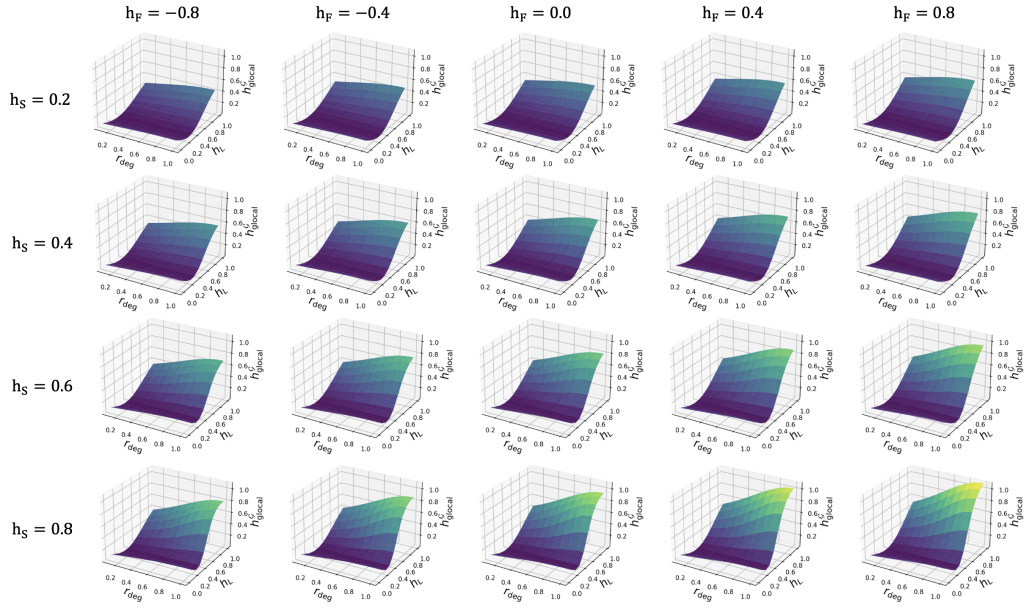
Table 15 compares the upper-triangle rule with a symmetric edge-probability matrix for Eq. 7: realized degree regularity, mean degree, and h_S are nearly identical. Table 16 shows where the degree term changes the score on the synthetic grid.

Table 6 (Appendix A) compares the measured effect of degree regularity with the aggregation model of Proposition 2, and Table 7 applies the model to the real graphs.

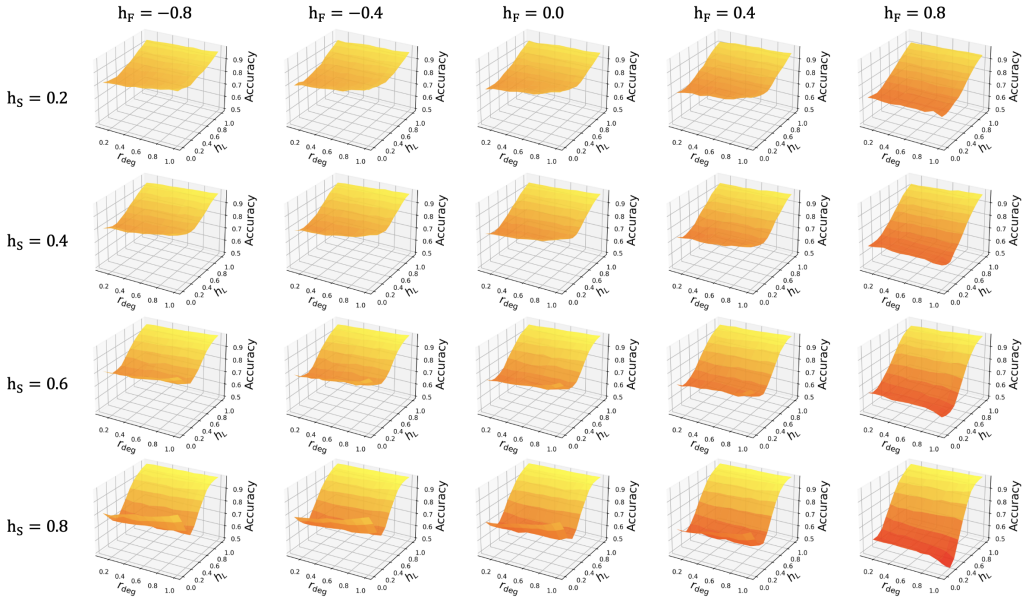
B.4 Numerical Behavior and Full Grids of the Original Run ($\sqrt{w}$ Variant)

Table 17: Numerical sanity check: conditional correlations between $h^{\mathcal{G}}_{\text{glocal}}$ (from the target factors) and its own inputs. They follow from Eq. 5 by construction and are not evidence of predictive validity.

Factor	Pearson	Spearman
h_L	0.827	0.731
h_S	0.997	1.000
h_F	0.283	0.288
$r_{\deg}$	0.992	1.000



(a) $h^{\mathcal{G}}_{\text{glocal}}$ ($\rho = 10$).



(b) GCN accuracy ($\sqrt{w}$ variant, original run).

Figure 3: Full synthetic grids. Rows: $h_S \in \{0.2, 0.4, 0.6, 0.8\}$; columns: $h_F \in \{-0.8, -0.4, 0, 0.4, 0.8\}$; each surface is over target (h_L, r_{deg}).

Table 17 is a numerical sanity check of Eq. 5, and Figure 3 shows the full synthetic grids of the original run of the $\sqrt{w}$ variant; Figure 1 summarizes the re-runs of Appendix B.3.

C Real-World Data, Training, and Baseline Metrics

C.1 Datasets

Table 18: Statistics of the 31 real-world datasets after symmetrization and removal of self-loops and duplicate edges. $\bar{d} = 2|\mathcal{E}|/|\mathcal{V}|$; $\rho(\mathbf{A})$: spectral radius. The Texas file (`texas-4-classes`) has five label values, one of which is assigned to a single node.

Dataset	#Nodes	#Edges	#Features	#Classes	$\bar{d}$	$\mathrm{CV}^2(\mathbf{d})$	$\rho(\mathbf{A})$
Roman-Empire	22,662	32,927	300	18	2.91	0.12	4.58
Amazon-Ratings	24,492	93,050	300	5	7.60	0.62	20.39
Minesweeper	10,000	39,402	7	2	7.88	0.01	7.99
Tolokers	11,758	519,000	10	2	88.28	3.36	392.36
Questions	48,921	153,540	301	2	6.28	18.66	95.31
Squirrel	2,223	46,998	2,089	5	42.28	3.15	206.02
Chameleon	890	8,854	2,325	5	19.90	1.99	78.05
Actor	7,600	26,659	932	5	7.02	6.42	37.37
Texas	183	279	1,703	5	3.05	6.55	10.98
Cornell	183	277	1,703	5	3.03	5.37	10.08
Wisconsin	251	450	1,703	5	3.59	4.89	11.88
Cora	2,708	5,278	1,433	7	3.90	1.80	14.39
CoraFull	19,793	63,421	8,710	70	6.41	1.88	25.63
CiteSeer	3,327	4,552	3,703	6	2.74	1.53	13.74
PubMed	19,717	44,324	500	3	4.50	2.73	23.24
Flickr	89,250	449,878	500	7	10.08	9.92	83.06
Amazon-Photo	7,650	119,081	745	8	31.13	2.31	122.54
Amazon-Computer	13,752	245,861	767	10	35.76	3.87	169.71
Coauthor-CS	18,333	81,894	6,805	15	8.93	1.04	24.60
Coauthor-Physics	34,493	247,962	8,415	5	14.38	1.17	51.18
WikiCS	11,701	215,603	300	10	36.85	3.67	149.77
Blog-Catalog	5,196	171,743	8,189	6	66.11	0.69	114.01
Ogbn-Arxiv	169,343	1,157,799	128	40	13.67	25.20	180.27
Genius	421,961	922,868	12	2	4.37	114.89	212.82
Twitch-DE	9,498	153,138	2,514	2	32.25	6.92	149.92
Twitch-ENGB	7,126	35,324	2,545	2	9.91	5.01	43.41
Twitch-ES	4,648	59,382	2,148	2	25.55	3.74	89.82
Twitch-FR	6,549	112,666	2,275	2	34.41	5.07	130.24
Twitch-PTBR	1,912	31,299	1,449	2	32.74	2.91	99.09
Twitch-RU	4,385	37,304	2,224	2	17.01	6.00	76.26
Twitch-TW	2,772	63,462	1,288	2	45.79	3.17	143.43

We conduct experiments on 31 real-world graph datasets: Roman-Empire, Amazon-Ratings, Minesweeper, Tolokers, and Questions from Platonov et al. (2023b); Squirrel, Chameleon, Actor, Texas, Cornell, and Wisconsin originally from Pei et al. (2020), in the filtered versions of Platonov et al. (2023b) (for Texas we use their `texas-4-classes` file, which in our copy still contains five label values, one of them assigned to a single node, so $C = 5$ in Table 18 and in the dataset-specific-C analysis); Cora, PubMed, and CiteSeer from Yang et al. (2016); CoraFull, Amazon-Photo, Amazon-Computer, Coauthor-CS, and Coauthor-Physics from Shchur et al. (2018); Flickr from Zeng et al. (2020); WikiCS from Mernyei & Cangea (2020); Blog-Catalog from Yang et al. (2023); Ogbn-Arxiv from Hu et al. (2020); and Genius, Twitch-DE, Twitch-ENGB, Twitch-ES, Twitch-FR, Twitch-PTBR, Twitch-RU, and Twitch-TW from Lim et al. (2021a). These datasets include both homophilic and heterophilic graphs spanning citation networks, webpage networks, purchase networks, coauthor networks, actor networks, and social networks. The diversity of these datasets enables comprehensive evaluation across a wide range of graph structures and homophily regimes. Dataset statistics are summarized in Table 18.

C.2 TRAINING DETAILS

Table 19: Node classification accuracy (%, mean±std over 10 runs), reproduced for this paper.

Dataset	MLP	GCN	GraphSAGE	GAT
	Acc±Std	Acc±Std	Acc±Std	Acc±Std
Actor	35.28±1.46	33.95±0.98	34.90±1.13	34.26±0.83
Amazon-Computer	83.68±0.90	91.55±0.37	90.87±0.62	91.49±0.47
Amazon-Photo	90.52±0.45	94.27±0.67	95.32±0.47	94.92±0.58
Amazon-Ratings	43.93±0.54	48.24±0.66	51.63±0.64	48.19±0.48
Blog-Catalog	93.51±0.65	96.00±0.47	95.72±0.50	95.19±0.46
Chameleon	40.21±3.62	39.98±3.34	39.15±4.95	39.00±4.64
CiteSeer	71.65±0.87	76.12±1.28	76.77±0.85	75.80±1.36
Coauthor-CS	94.22±0.34	95.46±0.26	95.28±0.28	95.21±0.30
Coauthor-Physics	94.90±0.34	96.86±0.25	96.75±0.31	96.71±0.27
Cora	68.89±0.76	85.80±1.26	86.89±1.45	86.93±1.48
CoraFull	57.79±1.01	68.19±0.74	68.35±1.00	69.07±0.61
Cornell	64.86±8.83	63.78±7.67	66.22±5.73	59.19±8.96
Flickr	46.53±0.37	51.66±0.33	51.97±0.42	52.93±0.49
Genius	86.47±0.07	88.32±1.02	88.07±0.30	76.83±1.53
Minesweeper	50.94±1.16	89.79±0.53	90.62±0.60	90.28±0.57
Ogbn-Arxiv	54.74±0.22	68.27±0.27	69.18±0.37	69.22±0.34
PubMed	86.25±0.46	88.45±0.43	88.26±0.44	87.93±0.60
Questions	71.87±1.15	72.77±1.51	76.27±0.88	77.75±0.94
Roman-Empire	65.21±0.54	75.34±0.58	79.04±0.74	79.11±0.35
Squirrel	39.44±1.88	39.04±1.75	39.17±2.96	38.81±2.20
Texas	71.47±6.75	65.48±6.30	72.82±4.98	70.11±4.79
Tolokers	73.19±0.82	83.47±0.79	82.76±0.47	83.75±0.37
Twitch-DE	67.25±1.07	72.69±1.06	70.61±1.05	72.82±1.24
Twitch-ENGB	60.91±1.11	63.06±0.97	61.73±1.16	62.83±1.17
Twitch-ES	60.88±2.16	66.93±1.46	64.32±1.47	66.90±1.35
Twitch-FR	60.58±1.55	65.77±1.72	62.14±1.93	65.82±2.40
Twitch-PTBR	61.71±2.33	67.74±1.49	64.59±2.22	67.39±1.68
Twitch-RU	50.66±2.55	52.41±4.86	51.80±2.87	53.65±1.81
Twitch-TW	59.37±2.07	62.81±2.31	60.52±1.61	64.45±1.56
WikiCS	80.45±0.70	84.09±0.82	84.86±0.44	84.73±0.84
Wisconsin	75.88±5.23	71.96±7.85	77.65±6.68	74.31±4.93

For all datasets, we randomly split the train, validation, and test sets with a ratio of 50%:25%:25% and report the average performance over 10 runs. We use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001. The maximum number of training epochs is set to 1000 with a patience of 40 for early stopping. To improve training stability and performance, we employ skip connections (He et al., 2016) and layer normalization (Ba et al., 2016) in each layer, following Platonov et al. (2023b). The GCN, GraphSAGE, and GAT models are therefore residual, layer-normalized variants of the original architectures, and the MLP uses the same residual blocks without aggregation; conclusions for the vanilla architectures may differ. All experiments were conducted on a single workstation equipped with an AMD Ryzen 9 9900X CPU, an NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU with 96GB memory, and 64GB RAM. For hyperparameter tuning, we perform grid search on the validation set over the following search space: (i) number of layers $\in \{1, 2\}$, (ii) hidden dimension $\in \{64, 128, 256\}$, (iii) dropout rate $\in \{0.2, 0.4, 0.6, 0.8\}$, and (iv) weight decay $\in \{10^{-3}, 10^{-4}, 10^{-5}\}$.

Table 19 reports the resulting accuracies, which we reproduced for this paper and use in all real-world analyses.

Synthetic experiments (Appendix B). For each configuration, we generate 10 graphs (seeds 0–9) with $N = 1000$, $C = 3$, $M = 32$, mean degree weight $\bar{w} = 10.5$, and $\gamma = 0.2$, and split nodes 60/20/20. We train GCNs with 64 hidden units, dropout 0.5, Adam with learning rate 10^{-3}, and 200 epochs, for each combination of depth $\in \{1, 2\}$ and weight decay $\in \{10^{-3}, 10^{-4}, 10^{-5}\}$. For each

graph we record the test accuracy at the epoch with the best validation accuracy, select the setting with the best mean validation accuracy over the 10 graphs, and report its mean test accuracy.

C.3 Baseline Metrics

Table 20: Audit of baseline metrics: theoretical range, observed range over the 31 datasets, and changes relative to the inherited values. h_{node}, h_S, and h_F were not affected.

Metric	range	observed	note
h_{edge}	[0,1]	[0.047, 0.931]	Recomputed from the definition on the undirected simple graph; the stored values of 5 datasets (Questions, Flickr, Ogbn-Arxiv, Genius, CiteSeer) differed.
h_{adj}	≤ 1	$[-0.294, 0.872]$	Recomputed from the definition; the stored implementation used degrees reduced by one (it assumed a self-loop at every node).
LI	[0,1]	[0.000, 0.722]	Recomputed as $2 - H(y_\xi, y_\eta)/H(y_\xi)$; the stored values were negative for 16 datasets because of the same degree error.
h_{agg}	[0,1]	[0.000, 0.947]	Modified aggregation homophily $[2S_{\mathrm{agg}} - 1]_+$ as defined by Luan et al. (2022); the stored values were not clipped at 0.
h_{unb}	[-1,1]	$[-0.805, 0.972]$	Added: unbiased homophily of Mironov & Prokhorenkova (2025), Eq. 4, computed on the undirected simple graph.
h_{NLD}	[0,1]	[0.703, 0.990]	Added: neighborhood-label-distribution distinguishability of Zhao et al. (2026), computed with the routine `cal_nei_dis` of the authors' public code applied to ground-truth labels (our $O(N)$ re-implementation matches the original routine to 10^{-6}).
h_{node}	[0,1]	[0.046, 0.915]	Verified against a direct implementation (identical for all datasets).
h_{class}	[0,1]	[0.000, 0.847]	Verified against a direct implementation (identical for all datasets).
h_{den}, h_{2hop}, h_{nei}, h_{NS}, h_S, h_{GE}, $h_{\mathrm{LS\text{-}cos}}$, $h_{\mathrm{LS\text{-}euc}}$, h_{attr}, h_{CF}, h_F, h_{KR}, h_{GNB}, h_{SVM}	–	–	As computed by the public implementations of Luan et al. (2023) and Zheng et al. (2024); ranges: h_{den} [0.44, 0.50]; h_{2hop} [0.07, 0.91]; h_{nei} [0.25, 0.93]; h_{NS} [1.03, 51.76]; h_S [0.09, 0.76]; h_{GE} [0.01, 0.84]; $h_{\mathrm{LS\text{-}cos}}$ [0.01, 0.86]; $h_{\mathrm{LS\text{-}euc}}$ $[-277.15, -0.05]$; h_{attr} $[-0.02, 0.62]$; h_{CF} $[-0.41, 1.00]$; h_F [0.00, 1.00]; h_{KR} [0.00, 1.00]; h_{GNB} [0.00, 1.00]; h_{SVM} [0.00, 1.00]

We compare with 24 existing metrics: label-based (h_{edge} (Zhu et al., 2020), h_{node} (Pei et al., 2020), h_{class} (Lim et al., 2021b), h_{adj} (Platonov et al., 2023a), h_{unb} (Mironov & Prokhorenkova, 2025), h_{NLD} (Zhao et al., 2026), h_{den} (Li et al., 2023a), h_{2hop} (Cavallo et al., 2022), h_{nei} (Gong et al., 2023), LI (Platonov et al., 2023a)), structural (h_{NS} (Ma et al., 2021), h_{agg} (Luan et al., 2022), h_S (Zheng et al., 2024)), feature-based (h_{GE} (Jin et al., 2022), $h_{\mathrm{LS\text{-}cos}}$, $h_{\mathrm{LS\text{-}euc}}$ (Chen et al., 2023), h_{attr} (Yang et al., 2021), h_{CF} (Lee et al., 2024), h_F (Zheng et al., 2024)), classifier-based (h_{KR}, h_{GNB}, h_{SVM} (Luan et al., 2023)), and Tri-Hom (Zheng et al., 2024).

We inherited per-dataset values of the 22 metrics other than h_{unb} and h_{NLD} from the public implementations of Luan et al. (2023) and Zheng et al. (2024). Checking them against their definitions revealed four problems, which we fixed (Table 20). First, LI is defined as $2 - H(y_\xi, y_\eta)/H(y_\xi) \in [0, 1]$, where (y_ξ, y_η) are the labels at the endpoints of a uniformly random directed edge (Platonov et al., 2023a). The inherited implementation subtracted one from every degree, assuming a self-loop at every node, so the edge-class distribution did not sum to one on graphs without self-loops, and 16 datasets received negative values. Second, the same error affected h_{adj}. Third, h_{edge} differed from its definition on five datasets. Fourth, h_{agg} was reported as $2S_{\mathrm{agg}} - 1$ without the clipping at zero that defines the modified aggregation homophily $H^M_{\mathrm{agg}} = [2S_{\mathrm{agg}} - 1]_+$ (Luan et al., 2022). We recomputed h_{edge}, h_{adj}, and LI directly on the undirected simple graphs, clipped h_{agg}, and added the unbiased homophily h_{unb} of Mironov & Prokhorenkova (2025), which is designed for comparing graphs with different numbers and sizes of classes. The inputs of Tri-Hom and of our factor (h_{node}, h_S, h_F) were verified and are unaffected. For the remaining metrics we use the values produced by the public implementations and report their observed ranges. We also added the neighborhood-label-distribution distinguishability h_{NLD} of Zhao et al. (2026). The paper is closed access and we found no open copy, so we use the routine `cal_nei_dis` from the authors' public code (in their `models.py`), which combines the within-class cosine similarity of neighbor-label distributions with the singular-value entropy of the class-mean distributions, and apply it to ground-truth labels with their normalized adjacency; our linear-time re-implementation matches the original routine to 10^{-6} on small graphs. If the paper's definition differs from this routine, h_{NLD} should be read as the authors' code rather than their paper.

D Additional Real-World Analyses

All analyses in this section use the per-dataset metric values of Appendix G and the accuracies of Table 19; no model is retrained.

D.1 Significance, Partial Correlations, Rank Correlations, and Family Deletion

Table 21: Dependent correlations $r(h^{\mathcal{G}}_{\text{glocal}}, \text{Acc})$ vs. $r(h^{\mathcal{G}}_{\text{Tri-Hom}}, \text{Acc})$ ($n = 31$): Williams' test, and bootstrap over datasets or whole families (10,000 resamples; one-sided $p = \Pr(\Delta r \le 0)$).

Model	r_{glocal}	$r_{\text{Tri-Hom}}$	r_{between}	t	p	bootstrap CI	p_{boot}	family CI	p_{fam}
MLP	0.565	0.559	0.934	0.11	0.91	$[-0.092, 0.096]$	0.45	$[-0.145, 0.107]$	0.47
GCN	0.692	0.666	0.934	0.54	0.59	$[-0.029, 0.110]$	0.19	$[-0.019, 0.165]$	0.16
GraphSAGE	0.680	0.651	0.934	0.58	0.57	$[-0.021, 0.105]$	0.15	$[-0.017, 0.155]$	0.14
GAT	0.717	0.704	0.934	0.28	0.78	$[-0.036, 0.084]$	0.31	$[-0.022, 0.133]$	0.25

Table 22: Partial correlations with accuracy (p in parentheses). "stats": $\log N$, $\log C$, $\log M$, $\log \bar{d}$; "+MLP": also MLP accuracy; "Tri-Hom": $h^{\mathcal{G}}_{\text{Tri-Hom}}$ only; "target: GNN$-$MLP": correlation with the accuracy gain over MLP.

Controls	Score	MLP	GCN	GraphSAGE	GAT
stats	Tri-Hom	0.544 (0.003)	0.653 (0.000)	0.653 (0.000)	0.692 (0.000)
stats	glocal	0.548 (0.003)	0.694 (0.000)	0.686 (0.000)	0.716 (0.000)
stats	$r_{\deg}$	0.147 (0.464)	0.349 (0.074)	0.336 (0.087)	0.357 (0.067)
Tri-Hom	$r_{\deg}$	0.025 (0.894)	0.219 (0.246)	0.214 (0.257)	0.255 (0.174)
stats+MLP	Tri-Hom	–	0.468 (0.016)	0.486 (0.012)	0.554 (0.003)
stats+MLP	glocal	–	0.584 (0.002)	0.585 (0.002)	0.608 (0.001)
stats+MLP	$r_{\deg}$	–	0.559 (0.003)	0.566 (0.003)	0.519 (0.007)
target: GNN$-$MLP	Tri-Hom	–	0.285 (0.121)	0.262 (0.155)	0.330 (0.070)
target: GNN$-$MLP	glocal	–	0.331 (0.069)	0.314 (0.086)	0.345 (0.057)
target: GNN$-$MLP	$r_{\deg}$	–	0.367 (0.042)	0.365 (0.044)	0.387 (0.031)

Table 23: Spearman rank correlation with accuracy on 31 datasets.

Score	MLP	GCN	GraphSAGE	GAT
Tri-Hom$^{\neg\mathcal{G}}$	0.567	0.658	0.629	0.648
Tri-Hom$^{\mathcal{G}}$	0.498	0.665	0.618	0.694
$r_{\deg}$	0.137	0.362	0.339	0.383
glocal$^{\neg\mathcal{G}}$	0.577	0.662	0.633	0.651
glocal$^{\mathcal{G}}$	0.512	0.689	0.644	0.719

Table 24: Leave-one-family-out Pearson correlations, $h^{\mathcal{G}}_{\text{Tri-Hom}}$ / $h^{\mathcal{G}}_{\text{glocal}}$. Families: Platonov (Roman-Empire, Amazon-Ratings, Minesweeper, Tolokers, Questions), WebKB/Wiki (Squirrel, Chameleon, Actor, Texas, Cornell, Wisconsin), Planetoid, Shchur (CoraFull, Amazon, Coauthor), LINKX (Genius, Twitch), Other (Flickr, WikiCS, Blog-Catalog, Ogbn-Arxiv).

Held-out family	n	MLP	GCN	GraphSAGE	GAT
LINKX	23	0.626 / 0.616	0.729 / 0.741	0.708 / 0.717	0.733 / 0.733
Other	27	0.646 / 0.636	0.743 / 0.766	0.722 / 0.750	0.782 / 0.792
Planetoid	28	0.561 / 0.567	0.652 / 0.677	0.636 / 0.662	0.691 / 0.702
Platonov	26	0.591 / 0.627	0.675 / 0.679	0.656 / 0.667	0.712 / 0.709
Shchur	26	0.266 / 0.237	0.502 / 0.564	0.488 / 0.553	0.566 / 0.610
WebKB/Wiki	25	0.517 / 0.545	0.608 / 0.683	0.617 / 0.690	0.675 / 0.732

Table 25: Power of Williams' test (two-sided, $\alpha = 0.05$) to detect the observed advantage of $h^{\mathcal{G}}_{\text{glocal}}$ over $h^{\mathcal{G}}_{\text{Tri-Hom}}$ with n datasets (4,000 Monte Carlo draws with the observed correlations).

n	GCN	GraphSAGE	GAT
31	0.07	0.08	0.05
62	0.12	0.13	0.06
100	0.18	0.19	0.08
200	0.30	0.32	0.12
400	0.53	0.59	0.19
800	0.82	0.86	0.32
1600	0.98	0.99	0.55

Table 26: Exploratory regime-gated term, applied only to datasets with $h_{\text{node}} \geq \tau$. Left: Pearson correlation for fixed τ ($\tau = 0$ is Eq. 4). Right: out-of-sample R^2 with τ selected inside each training fold.

Model	$\tau = 0.0$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	LOOCV	LOOCV (gated)	LOFO	LOFO (gated)
MLP	0.565	0.571	0.569	0.569	0.569	0.562	0.532	0.224	0.231	0.174	0.178
GCN	0.692	0.691	0.688	0.688	0.688	0.687	0.675	0.413	0.364	0.353	0.282
GraphSAGE	0.680	0.683	0.680	0.680	0.680	0.666	0.650	0.395	0.398	0.366	0.295
GAT	0.717	0.717	0.714	0.714	0.714	0.712	0.704	0.450	0.382	0.374	0.299

Tables 21–24 give the details of the significance tests, partial correlations, rank correlations, and family deletion summarized in Section 6.2. Table 25 shows that the observed advantage over Tri-Hom cannot be detected reliably with any existing benchmark collection: the two scores are highly correlated ($r = 0.93$), so the standard error of their difference is large relative to the difference itself.

D.2 Specification Curve

Table 27: Specification curve: partial correlation of $r_{\deg}$ with GNN accuracy over all 64 subsets of the controls $\{\log N, \log C, \log M, \log \bar{d}, \text{MLP accuracy}, h^{\mathcal{G}}_{\text{Tri-Hom}}\}$; last column: share with $p < 0.05$.

Model	Specifications	#	min	median	max	$p < 0.05$
GCN	all 64 specifications	64	0.219	0.363	0.581	45%
GCN	with MLP accuracy	32	0.370	0.515	0.581	91%
GCN	without MLP accuracy	32	0.219	0.278	0.357	0%
GCN	with Tri-Hom	32	0.219	0.325	0.569	41%
GCN	without Tri-Hom	32	0.275	0.371	0.581	50%
GraphSAGE	all 64 specifications	64	0.210	0.350	0.596	45%
GraphSAGE	with MLP accuracy	32	0.349	0.504	0.596	91%
GraphSAGE	without MLP accuracy	32	0.210	0.268	0.351	0%
GraphSAGE	with Tri-Hom	32	0.210	0.310	0.595	41%
GraphSAGE	without Tri-Hom	32	0.267	0.360	0.596	50%
GAT	all 64 specifications	64	0.255	0.388	0.548	53%
GAT	with MLP accuracy	32	0.405	0.499	0.548	100%
GAT	without MLP accuracy	32	0.255	0.298	0.371	6%
GAT	with Tri-Hom	32	0.255	0.352	0.546	50%
GAT	without Tri-Hom	32	0.295	0.388	0.548	56%

Table 28: First column: correlation between MLP and GNN accuracy. Other columns: partial / semipartial correlation of $r_{\deg}$ with GNN accuracy for three control sets ("stats": $\log N$, $\log C$, $\log M$, $\log \bar{d}$); the semipartial correlation residualizes only the accuracy.

Model	r(MLP, GNN)	given MLP	given stats + MLP	given Tri-Hom + stats + MLP
GCN	0.895	0.393 / 0.390	0.559 / 0.528	0.553 / 0.515
GraphSAGE	0.898	0.388 / 0.385	0.566 / 0.534	0.563 / 0.525
GAT	0.877	0.422 / 0.419	0.519 / 0.490	0.520 / 0.485

Table 27 reports the specification curve. Table 28 examines the role of the MLP control, on which the significance of the partial correlations depends. Partial and semipartial values agree, so the control acts on the outcome (a precision variable), not by altering $r_{\deg}$; MLP accuracy is also not plausibly a collider, since an MLP does not use the graph.

D.3 Out-of-Sample Calibration

Table 29: Leave-one-out cross-validation of linear predictors: correlation between out-of-sample predictions and accuracy / out-of-sample R^2. The additive $\alpha \in [0, 2]$ is tuned inside each training fold; the last row is the in-sample correlation of the product.

Predictor	MLP	GCN	GraphSAGE	GAT
$h^{\mathcal{G}}_{\text{Tri-Hom}}$ (1-var OLS)	0.471 / 0.215	0.607 / 0.365	0.591 / 0.346	0.654 / 0.425
$h^{\mathcal{G}}_{\text{glocal}}$ (1-var OLS)	0.480 / 0.224	0.646 / 0.413	0.631 / 0.395	0.674 / 0.450
$h^{\mathcal{G}}_{\text{Tri-Hom}} + \alpha\, r_{\deg}$ (α tuned in-fold)	0.413 / 0.147	0.578 / 0.317	0.564 / 0.301	0.642 / 0.402
$h^{\mathcal{G}}_{\text{Tri-Hom}} \times r_{\deg}$ (1-var OLS)	0.400 / 0.147	0.624 / 0.386	0.602 / 0.359	0.641 / 0.408
OLS($h^{\mathcal{G}}_{\text{Tri-Hom}}, r_{\deg}$)	0.385 / 0.110	0.579 / 0.318	0.565 / 0.304	0.643 / 0.404
$h^{\mathcal{G}}_{\text{Tri-Hom}} \times r_{\deg}$ (raw corr)	0.513	0.670	0.651	0.685

Table 30: Leave-one-family-out cross-validated R^2 for all predictors.

Predictor	MLP	GCN	GraphSAGE	GAT
$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.171	0.338	0.347	0.396
$h^{\mathcal{G}}_{\text{Tri-Hom}} + \alpha\, r_{\deg}$ (α in-fold)	0.102	0.301	0.321	0.389
OLS on ($h^{\mathcal{G}}_{\text{Tri-Hom}}, r_{\deg}$)	0.008	0.299	0.324	0.387
$h^{\mathcal{G}}_{\text{glocal}}$, $\gamma = 0.2$	0.174	0.353	0.366	0.374
$h^{\mathcal{G}}_{\text{glocal}}$, γ in-fold	-0.221	0.170	0.194	0.199

Table 4 in the main text reports leave-one-dataset-out and leave-one-family-out R^2, with γ either fixed or selected on the training folds from $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2\}$. Table 29 adds the correlation between out-of-sample predictions and accuracy and the multiplicative combination, and Table 30 gives leave-one-family-out R^2 for all predictors.

Regime-gated term. Motivated by the sign change in the synthetic study, we tried applying the degree term only to datasets with $h_{\text{node}} \geq \tau$. Table 26 shows that correlations are nearly flat in τ up to 0.5, and that selecting τ inside the folds lowers out-of-sample R^2 for all GNNs. One reason is that 23 of the 31 datasets have $h_{\text{node}} \geq 1/3$, and near $h_L = 1/C$ the Tri-Hom factor is small, so the degree term has little effect there anyway (Table 16).

D.4 Design Search and Family Structure

Table 31: Correction for the design search over all 52 degree-based variants of the global term (23 statistics, including $\exp(-\gamma \mathrm{CV}^2)$ with 9 values of γ, Gini, entropy, clustering, assortativity, and aggregation- and GCN-noise statistics, each linear or squared, plus 6 regime gates), each scored by its mean gain over Tri-Hom across the three GNNs. Null: degree-derived statistics permuted across datasets, freely or within families (5,000 permutations), with the whole search rerun. p_{fixed}: our variant against its own null; $p_{\text{fixed}|\max}$, $p_{\text{best}|\max}$: our variant and the best variant against the null maximum.

Permutation	gain (ours)	best gain	null max (median)	null max (95%)	p_{fixed}	$p_{\text{fixed}\|\max}$	$p_{\text{best}\|\max}$
all datasets	0.023	0.027	0.005	0.023	0.003	0.056	0.028
within families	0.023	0.027	0.013	0.044	0.091	0.329	0.266

Table 32: Partial correlations with accuracy given dataset-family fixed effects, optionally with MLP accuracy and dataset statistics (p in parentheses).

Model	Controls	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	$h^{\mathcal{G}}_{\text{glocal}}$	$r_{\deg}$
GCN	family FE	0.363 (0.068)	0.469 (0.016)	0.056 (0.787)
GCN	family FE + MLP	0.289 (0.161)	0.490 (0.013)	0.245 (0.238)
GCN	family FE + dataset stats + MLP	0.205 (0.373)	0.496 (0.022)	0.412 (0.064)
GraphSAGE	family FE	0.352 (0.078)	0.446 (0.022)	0.019 (0.929)
GraphSAGE	family FE + MLP	0.272 (0.189)	0.453 (0.023)	0.175 (0.403)
GraphSAGE	family FE + dataset stats + MLP	0.257 (0.261)	0.510 (0.018)	0.350 (0.119)
GAT	family FE	0.432 (0.027)	0.510 (0.008)	0.063 (0.760)
GAT	family FE + MLP	0.429 (0.032)	0.561 (0.004)	0.246 (0.237)
GAT	family FE + dataset stats + MLP	0.419 (0.059)	0.597 (0.004)	0.188 (0.413)

Table 33: Family clustering of the 31 datasets (six families, unbalanced average family size $m_0 = 5.07$). ICC: one-way intraclass correlation by family; design effect $1 + (m_0 - 1)\,\mathrm{ICC}$; $n_{\text{eff}} = 31/\text{design effect}$.

Quantity	ICC	design effect	n_{eff}
$r_{\deg}$	0.185	1.75	17.7
$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.603	3.46	9.0
GCN accuracy	0.411	2.67	11.6
GCN accuracy residual given Tri-Hom, stats, MLP	0.045	1.18	26.2
GraphSAGE accuracy	0.368	2.50	12.4
GraphSAGE accuracy residual given Tri-Hom, stats, MLP	0.032	1.13	27.5
GAT accuracy	0.441	2.80	11.1
GAT accuracy residual given Tri-Hom, stats, MLP	0.141	1.57	19.7

Table 34: Association with the accuracy gain of each GNN over MLP: Pearson and Spearman correlation, and partial correlation given family fixed effects (p in parentheses).

Model	Score	Pearson	Spearman	partial (family FE)
GCN	$r_{\deg}$	0.367 (0.042)	0.284 (0.122)	0.251 (0.217)
GCN	$h^{\mathcal{G}}_{\text{glocal}}$	0.331 (0.069)	0.369 (0.041)	0.326 (0.104)
GCN	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.285 (0.121)	0.349 (0.054)	0.157 (0.444)
GraphSAGE	$r_{\deg}$	0.365 (0.044)	0.265 (0.150)	0.186 (0.363)
GraphSAGE	$h^{\mathcal{G}}_{\text{glocal}}$	0.314 (0.086)	0.368 (0.042)	0.327 (0.103)
GraphSAGE	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.262 (0.155)	0.375 (0.037)	0.170 (0.407)
GAT	$r_{\deg}$	0.387 (0.031)	0.167 (0.371)	0.249 (0.221)
GAT	$h^{\mathcal{G}}_{\text{glocal}}$	0.345 (0.057)	0.415 (0.020)	0.358 (0.073)
GAT	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.330 (0.070)	0.430 (0.016)	0.255 (0.208)

Table 31 corrects the comparison with Tri-Hom for the design search, and Table 32 adds family fixed effects. For future external validation we lock the specification used throughout the paper: CV^2 of the simple-graph degrees, $\gamma = 0.2$, the squared term of Eq. 4, and $C = 3$, $\rho(\mathbf{A}) = 10$.

Table 33 quantifies how strongly the datasets cluster by family (raw accuracies strongly, the accuracy residuals given Tri-Hom, dataset statistics, and MLP accuracy much less), and Table 34 uses the accuracy gain of each GNN over MLP as the target.

D.5 Alternative Degree Statistics, Functional Form, and Sensitivity to γ

Table 35: Alternative statistics $s \in (0, 1]$ in place of $r_{\deg}$: Pearson correlation with accuracy of s alone / of $h^{\mathcal{G}}_{\text{glocal}}$ with s in the global term. Local CV^2: mean over nodes of the CV^2 of degrees in the closed 1-hop neighborhood.

Statistic s	MLP	GCN	GraphSAGE	GAT
exp(-0.2 CV^2) [ours]	0.129 / 0.565	0.289 / 0.692	0.286 / 0.680	0.315 / 0.717
1/(1+CV^2)	0.039 / 0.503	0.288 / 0.679	0.303 / 0.675	0.317 / 0.709
1 - Gini	−0.048 / 0.524	0.199 / 0.684	0.247 / 0.679	0.232 / 0.717
degree entropy	0.137 / 0.619	−0.061 / 0.638	−0.104 / 0.617	−0.034 / 0.669
mean/max degree	−0.165 / 0.371	0.193 / 0.622	0.202 / 0.622	0.205 / 0.656
exp(-0.2 local CV^2)	0.069 / 0.561	0.191 / 0.670	0.224 / 0.659	0.197 / 0.707
avg clustering	−0.138 / 0.598	−0.017 / 0.686	−0.021 / 0.677	0.004 / 0.718
(1+assortativity)/2	−0.002 / 0.568	0.230 / 0.676	0.220 / 0.669	0.223 / 0.710

Table 36: Sensitivity to γ: Pearson correlation of $h^{\mathcal{G}}_{\text{glocal}}$ with accuracy, and median $r_{\deg}$ over the 31 datasets.

γ	median $r_{\deg}$	MLP	GCN	GraphSAGE	GAT
0.01	0.967	0.562	0.668	0.652	0.705
0.02	0.935	0.569	0.674	0.657	0.710
0.05	0.845	0.580	0.687	0.669	0.718
0.1	0.715	0.581	0.695	0.678	0.721
0.2	0.511	0.565	0.692	0.680	0.717
0.5	0.186	0.510	0.671	0.666	0.698
1.0	0.035	0.444	0.652	0.652	0.684
2.0	0.001	0.382	0.630	0.633	0.666
5.0	0.000	0.356	0.617	0.619	0.653
Tri-Hom (gamma=0)	–	0.559	0.666	0.651	0.704

Table 37: Functional form of the global term. Squared: $C(1-s)^2/(C-1)$ (Eq. 4); linear: $C(1-s)/(C-1)$. $r_{\text{agg}} = \exp(-\gamma\nu)$ with $\nu = \mu \cdot \frac{1}{N}\sum_i d_i^{-1} - 1$ (isolated nodes excluded).

Global term	MLP	GCN	GraphSAGE	GAT
$h^{\mathcal{G}}_{\text{Tri-Hom}}$ (no degree term)	0.559	0.666	0.651	0.704
$r_{\deg}$, squared term (ours)	0.565	0.692	0.680	0.717
$r_{\deg}$, linear term	0.525	0.686	0.679	0.714
r_{agg} (γ=0.1), squared term	0.552	0.662	0.653	0.698
r_{agg} (γ=0.1), linear term	0.535	0.670	0.667	0.705
r_{agg} (γ=0.2), squared term	0.543	0.664	0.661	0.698
r_{agg} (γ=0.2), linear term	0.509	0.668	0.669	0.702
r_{agg} (γ=0.5), squared term	0.497	0.661	0.664	0.692
r_{agg} (γ=0.5), linear term	0.455	0.658	0.663	0.693

Table 38: Global term from the noise of the GCN operator $\mathbf{P} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$: $r_{\text{gcn}} = \exp(-\gamma\nu_{\text{gcn}})$ with $\nu_{\text{gcn}} = (\bar{d}+1)\frac{1}{N}\|\mathbf{P}\|_F^2 - 1$, the excess over a regular graph with the same mean degree.

Global term	MLP	GCN	GraphSAGE	GAT
glocal with r_{gcn} (γ=0.1)	0.556	0.664	0.649	0.702
glocal with r_{gcn} (γ=0.2)	0.548	0.658	0.645	0.697
glocal with r_{gcn} (γ=0.5)	0.522	0.640	0.631	0.678
glocal with r_{gcn} (γ=1.0)	0.492	0.626	0.624	0.663

Table 39: Exploratory: partial correlation of global statistics with GNN accuracy (p in parentheses; "stats": $\log N$, $\log C$, $\log M$, $\log \bar{d}$). The negative partial correlation of clustering given $r_{\deg}$ reflects the correlation between the two and should not be interpreted.

Controls	Statistic	GCN	GraphSAGE	GAT
Tri-Hom + stats + MLP	$r_{\deg}$	0.553 (0.004)	0.563 (0.003)	0.520 (0.008)
Tri-Hom + stats + MLP	$1 - \mathrm{Gini}$	0.685 (0.000)	0.761 (0.000)	0.708 (0.000)
Tri-Hom + stats + MLP	clustering	0.013 (0.952)	0.040 (0.848)	−0.025 (0.905)
Tri-Hom + stats + MLP	assortativity	0.390 (0.054)	0.410 (0.042)	0.292 (0.157)
Tri-Hom + MLP + family FE	$r_{\deg}$	0.356 (0.088)	0.273 (0.196)	0.424 (0.039)
Tri-Hom + MLP + family FE	$1 - \mathrm{Gini}$	0.580 (0.003)	0.653 (0.001)	0.646 (0.001)
Tri-Hom + MLP + family FE	clustering	0.152 (0.480)	0.055 (0.799)	0.120 (0.575)
Tri-Hom + MLP + family FE	assortativity	0.474 (0.019)	0.363 (0.081)	0.320 (0.127)
Tri-Hom + stats + MLP + clustering	$r_{\deg}$	0.731 (0.000)	0.721 (0.000)	0.722 (0.000)
Tri-Hom + stats + MLP + $r_{\deg}$	clustering	−0.575 (0.003)	−0.547 (0.006)	−0.587 (0.003)

Table 35 replaces $r_{\deg}$ in Eq. 4 by other statistics mapped to $(0, 1]$. CV^2-based statistics and $1 - \mathrm{Gini}$ give nearly identical results for GNNs, which is expected because all three measure dispersion. Entropy and the mean-to-max ratio are clearly worse, and the local neighborhood variant is worse than the global statistic. The average clustering coefficient, a different structural property, performs similarly in this slot, but it has no partial correlation with GNN accuracy given Tri-Hom, dataset statistics, and MLP accuracy (Table 39); its similar slot performance therefore does not mean that it carries the same information as the dispersion statistics. Table 36 shows that every $\gamma \in [0.01, 0.2]$ improves over Tri-Hom for all three GNNs, and up to $\gamma = 0.5$ for GCN and GraphSAGE. The value $\gamma = 0.2$ is in the flat region, but it was not selected on held-out data, so it should be read as a representative value in this region rather than as a tuned optimum.

Table 37 compares functional forms of the degree term, and Table 38 uses the noise of the GCN operator as the global term; Remark 1 and Table 5 explain why this noise tracks inverse-degree moments rather than CV^2.

Table 40: Pearson correlation with accuracy when the degree term of Eq. 4 is replaced by a derived, parameter-free term (Williams' p against Tri-Hom in parentheses; $\mathbb{E}[1/d]$ over non-isolated nodes).

Score	MLP	GCN	GraphSAGE	GAT
$h^{\mathcal{G}}_{\text{Tri-Hom}}$	0.559	0.666	0.651	0.704
$h^{\mathcal{G}}_{\text{glocal}}$ (Eq. 4)	0.565 (0.91)	0.692 (0.59)	0.680 (0.57)	0.717 (0.78)
derived: $(1 - \|\mathbf{p}\|^2)\mathbb{E}[1/d]$	0.571 (0.41)	0.665 (0.94)	0.650 (0.93)	0.700 (0.79)
derived: $(1 - \|\mathbf{p}\|^2)(\mathbb{E}[1/d] - 1/\mu)$	0.563 (0.73)	0.671 (0.55)	0.658 (0.48)	0.708 (0.64)

A derived degree term. Under the neighbor model of Tri-Hom, node v_i observes the classes of d_i neighbors drawn with probabilities $\mathbf{p} = q\mathbf{1} + a\mathbf{e}_{y_i}$, so its empirical neighbor-class vector has covariance $(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)/d_i$ and total variance $(1 - \|\mathbf{p}\|^2)/d_i$, with $\|\mathbf{p}\|^2 = h_L^2 + (1 - h_L)^2/(C - 1)$. This is the same kind of quantity as the h_S term (C times the per-entry variance of the node-wise perturbation). Averaging over nodes gives the parameter-free term $(1 - \|\mathbf{p}\|^2)\,\mathbb{E}[1/d]$, and its excess over a regular graph with the same mean degree is $(1 - \|\mathbf{p}\|^2)(\mathbb{E}[1/d] - 1/\mu)$. Table 40 uses each in place of the degree term of Eq. 4. Neither changes the correlations with GNN accuracy by more than 0.007 (0.012 for MLP), because the derived terms are small (at most 0.48) relative to T_{local}, whereas the heuristic term reaches 1.5. The weight that Eq. 4 gives to dispersion is therefore larger than this sampling argument implies, and we do not claim that it is derived.

Which global statistics carry conditional information? Table 39 reports partial correlations of four global statistics with GNN accuracy. Across datasets, $r_{\deg}$ has Spearman correlation 0.62 with $1 - \mathrm{Gini}$, 0.58 with clustering, and 0.54 with assortativity, whereas $1 - \mathrm{Gini}$ and clustering are nearly unrelated (0.10). This analysis was done after the main analysis and is exploratory.

D.6 Structural-Homophily Term and Normalization Constants

Table 41: Ablation of the structural-homophily term ("w/o h_S" sets $h_S = 1$).

Score	MLP	GCN	GraphSAGE	GAT
Tri-Hom$^{\mathcal{G}}$	0.559	0.666	0.651	0.704
Tri-Hom$^{\mathcal{G}}$ w/o h_S	0.545	0.664	0.634	0.692
glocal$^{\mathcal{G}}$	0.565	0.692	0.680	0.717
glocal$^{\mathcal{G}}$ w/o h_S	0.546	0.693	0.678	0.714
glocal$^{\mathcal{G}}$ w/o h_S, w/o $r_{\deg}$	0.545	0.664	0.634	0.692

Table 42: Fixed normalization constants ($C = 3$, $\rho(\mathbf{A}) = 10$) vs. dataset-specific number of classes and spectral radius.

Normalization	Score	MLP	GCN	GraphSAGE	GAT
fixed $C = 3$, $\rho = 10$ (paper)	Tri-Hom$^{\mathcal{G}}$	0.559	0.666	0.651	0.704
fixed $C = 3$, $\rho = 10$ (paper)	glocal$^{\mathcal{G}}$	0.565	0.692	0.680	0.717
fixed $C = 3$, $\rho = 10$ (paper)	Tri-Hom$^{\neg\mathcal{G}}$	0.593	0.618	0.617	0.629
fixed $C = 3$, $\rho = 10$ (paper)	glocal$^{\neg\mathcal{G}}$	0.600	0.623	0.621	0.636
true C, $\rho = 10$	Tri-Hom$^{\mathcal{G}}$	0.549	0.612	0.612	0.640
true C, $\rho = 10$	glocal$^{\mathcal{G}}$	0.598	0.656	0.650	0.672
true C, $\rho = 10$	Tri-Hom$^{\neg\mathcal{G}}$	0.519	0.535	0.539	0.560
true C, $\rho = 10$	glocal$^{\neg\mathcal{G}}$	0.517	0.533	0.537	0.561
$C = 3$, true $\rho(\mathbf{A})$	Tri-Hom$^{\mathcal{G}}$	0.542	0.667	0.651	0.706
$C = 3$, true $\rho(\mathbf{A})$	glocal$^{\mathcal{G}}$	0.544	0.690	0.677	0.715
$C = 3$, true $\rho(\mathbf{A})$	Tri-Hom$^{\neg\mathcal{G}}$	0.253	0.420	0.413	0.414
$C = 3$, true $\rho(\mathbf{A})$	glocal$^{\neg\mathcal{G}}$	0.253	0.421	0.413	0.415
true C, true $\rho(\mathbf{A})$	Tri-Hom$^{\mathcal{G}}$	0.540	0.612	0.612	0.640
true C, true $\rho(\mathbf{A})$	glocal$^{\mathcal{G}}$	0.584	0.652	0.647	0.668
true C, true $\rho(\mathbf{A})$	Tri-Hom$^{\neg\mathcal{G}}$	0.277	0.395	0.399	0.397
true C, true $\rho(\mathbf{A})$	glocal$^{\neg\mathcal{G}}$	0.276	0.395	0.400	0.398

Removing the h_S term from the glocal factor changes the GNN correlations by at most 0.003 (Table 41), so on real data the degree term, not h_S, accounts for the gain over Tri-Hom; we keep h_S so that the factor reduces exactly to Tri-Hom at $r_{\deg} = 1$. With dataset-specific C and $\rho(\mathbf{A})$ (Table 42), the graph-aware glocal factor remains above Tri-Hom for all four models, and the gain for MLP becomes larger. The graph-agnostic factors, however, degrade substantially with the true spectral radius. Because $h_F \in [-1, 1]$, a large $\rho(\mathbf{A})$ makes $h_F/\rho(\mathbf{A})$ nearly zero and the factor nearly constant, which is why we follow the fixed normalization in the main text.

D.7 SCATTER PLOTS

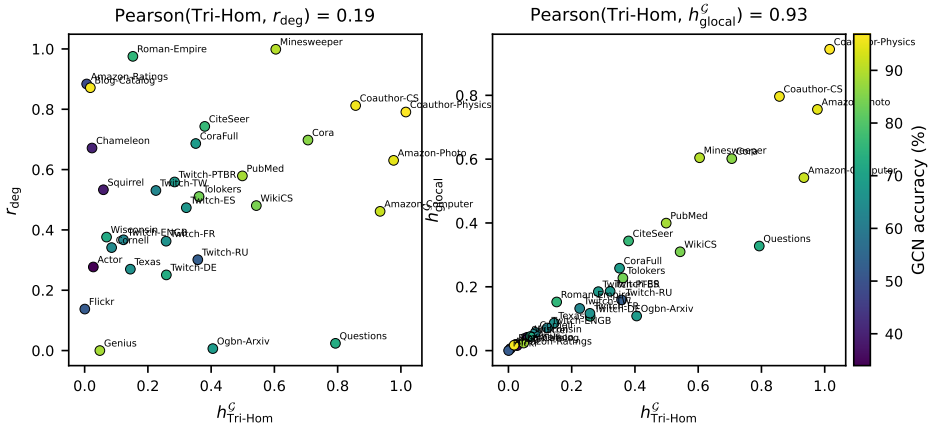


Figure 4: Real-world datasets colored by GCN accuracy. Left: $h^{\mathcal{G}}_{\text{Tri-Hom}}$ vs. r_{deg}. Right: $h^{\mathcal{G}}_{\text{Tri-Hom}}$ vs. $h^{\mathcal{G}}_{\text{glocal}}$.

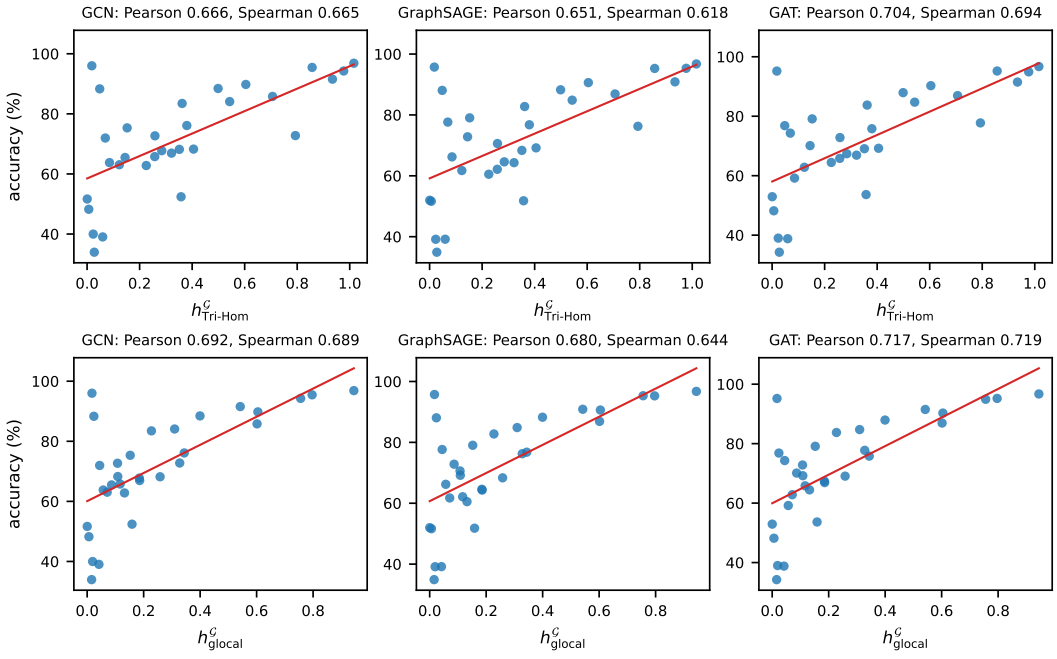


Figure 5: Factor versus accuracy for GCN, GraphSAGE, and GAT (top: Tri-Hom, bottom: glocal), with least-squares lines; rank correlations in the titles.

Figure 4 places the datasets in the plane of Tri-Hom and degree regularity, and Figure 5 plots each factor against accuracy. Few datasets have high factor values, so the least-squares lines are dominated by the many low-value datasets; rank correlations (Table 23) and family deletion (Table 24) are less sensitive to this.

D.8 SUMMARY OF CLAIMS

Table 43: Claims and the evidence for each.

Claim	Evidence	Verdict
Global degree structure carries information beyond local homophily factors	$r(r_{\deg}, h^{\mathcal{G}}_{\text{Tri-Hom}}) = 0.19$; partial correlation of $r_{\deg}$ positive in all 64 control specifications and significant when MLP accuracy is controlled (Table 27); permutation test $p = 0.003$, 0.03–0.06 after correcting for the design search, but $p = 0.09$ when permuting only within families (Table 31); $r_{\deg}$ alone is not significant with family fixed effects (Table 32)	Suggestive; conditional on MLP control and mostly between families
Degree regularity is the unique missing factor	In the slot, $1 - \text{Gini}$ and clustering perform within 0.01 (Table 35); as conditional predictors, clustering carries nothing and $1 - \text{Gini}$ is stronger than $r_{\deg}$ (Table 39, exploratory)	Not supported; the information is degree-based
Eq. 4 is theoretically justified	Proposition 1 is motivational; the squared form and CV^2 are not implied; GCN-operator noise does not track CV^2 (Table 38)	Not supported
Eq. 4 improves on Tri-Hom descriptively	Highest correlation for GCN, GraphSAGE, GAT; 17/18 family deletions; best LOOCV R^2, also with in-fold γ; higher partial correlation than Tri-Hom within families	Mostly yes
Eq. 4 improves on Tri-Hom significantly	Williams' $p = 0.57$–0.78; power 5–8% at $n = 31$ (Table 25)	No (underpowered)
Eq. 4 transfers to unseen families	LOFO-CV: better for GCN and GraphSAGE, worse for GAT; in-fold γ selection fails	Weak
Degree dispersion affects accuracy when local-factor targets are fixed and density variation is reduced	DC synthetic grid: partial $r = 0.32$ given mean degree; effect of $\pm$2–4 points whose sign depends on h_L (Table 14)	Supported (synthetic); not captured by Eq. 4

Table 43 summarizes each claim with its evidence.

E DEGREE-WISE ACCURACY DETAILS

Table 44: Degree-wise GCN accuracy (Figure 2): per dataset, mean degree (first row) and mean accuracy over 10 runs as a fraction (second row) of each of the ten equal-size test-node groups; n: test nodes per group and run.

Dataset	n	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Amazon-Ratings	490	5.0	5.0	5.0	5.0	5.0	5.8	6.6	8.0	10.4	20.7
		0.475	0.488	0.477	0.482	0.477	0.432	0.408	0.424	0.470	0.506
Coauthor-CS	367	1.4	2.5	3.5	4.5	5.6	7.0	8.6	11.0	15.2	30.0
		0.883	0.924	0.928	0.934	0.935	0.935	0.942	0.937	0.938	0.949
Cora	55	1.0	1.3	2.0	2.2	3.0	3.2	3.9	4.7	5.8	12.7
		0.825	0.827	0.869	0.871	0.865	0.895	0.915	0.915	0.885	0.867
CiteSeer	67	0.8	1.0	1.0	1.0	1.8	2.0	2.4	3.2	4.5	9.8
		0.684	0.691	0.658	0.700	0.772	0.751	0.763	0.791	0.851	0.851
Twitch-DE	191	1.6	3.8	6.6	9.8	13.5	18.2	24.5	33.9	50.9	165.8
		0.586	0.590	0.577	0.628	0.663	0.679	0.721	0.736	0.815	0.834
Flickr	1785	5.0	5.0	5.0	5.0	5.9	6.2	7.2	8.6	11.7	41.6
		0.533	0.525	0.533	0.536	0.519	0.513	0.513	0.512	0.488	0.473
Genius	8440	1.0	1.0	1.0	1.0	1.0	1.0	1.4	2.0	2.8	31.8
		0.774	0.777	0.774	0.775	0.777	0.776	0.795	0.831	0.921	0.820
Texas	4	1.0	1.0	1.1	1.4	1.9	2.1	2.6	3.3	5.4	33.1
		0.475	0.625	0.475	0.525	0.300	0.375	0.475	0.475	0.150	0.250

Table 45: Degree-wise experiment: a separate, simplified GCN (PyG 2-layer GCN, 64 hidden units, dropout 0.5, Adam, learning rate 10^{-3}, 200 epochs, weight decay and depth selected on validation, random 60/20/20 splits, 10 seeds). χ^2 test of equal accuracy across the ten groups per seed: median p over seeds and number of seeds with $p < 0.05/8$ (Bonferroni over 8 datasets).

Dataset	n/group	acc. (this exp.)	acc. (Table 19)	median p	seeds sig.
Amazon-Ratings	490	46.4	48.24	0.0037	7/10
Coauthor-CS	367	93.1	95.46	0.0067	5/10
Cora	55	87.3	85.80	0.6205	0/10
CiteSeer	67	75.0	76.12	0.0133	5/10
Twitch-DE	191	68.2	72.69	0.0000	10/10
Flickr	1785	51.5	51.66	0.0000	8/10
Genius	8440	80.2	88.32	0.0000	10/10
Texas	4	42.1	65.48	0.3238	0/10

The degree of each node is the number of distinct neighbors in the undirected simple graph. Table 44 lists the mean degree and accuracy of each degree group in Figure 2, and Table 45 describes this experiment and the test of equal accuracy.

F Detailed Definitions of Homophily Metrics

Label Homophily. Label homophily h_L measures the consistency of node labels across the topology. The two most commonly used metrics are edge homophily (Zhu et al., 2020; Abu-El-Haija et al., 2019) and node homophily (Pei et al., 2020), defined as:

$$h_{\text{edge}}(G, \mathbf{y}) = \frac{|\{e_{uv} \in \mathcal{E} \mid \mathbf{y}_u = \mathbf{y}_v\}|}{|\mathcal{E}|},\ h_{\text{node}}(G, \mathbf{y}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{|\{u \in \mathcal{N}(v) \mid \mathbf{y}_u = \mathbf{y}_v\}|}{|\mathcal{N}(v)|}. \tag{8}$$

$h^{\mathcal{G}}_{\text{Tri-Hom}}$ and $h^{\neg\mathcal{G}}_{\text{Tri-Hom}}$ adopt $h_{\text{node}}(G, \mathbf{y})$ as the measure of label homophily.

Structural Homophily. Structural homophily measures the consistency of structural information among nodes within the same class. Instead of directly considering label agreement across edges, it captures whether nodes belonging to the same class exhibit similar structural characteristics.

Formally, the structural homophily of a graph is defined as:

$$h_S(G, S, \mathbf{y}) = \frac{1}{C} \sum_{c=1}^{C} \left(1 - \frac{\sigma\left(\{S(u) \mid u \in \mathcal{V},\ \mathbf{y}_u = c\}\right)}{\sigma_{\max}}\right),$$

where $\sigma(\cdot)$ denotes the standard deviation and $\sigma_{\max}$ is the corresponding maximum value for normalization. In $h_{\text{Tri-Hom}}$ (Zheng et al., 2024), the structural information of v_i is quantified using its neighbor distribution, defined as the class distribution of its local neighbors: $\mathbf{D}_i^{\mathcal{N}} = [p_{i,1}, p_{i,2}, \ldots, p_{i,C}]$, where $p_{i,k} = \frac{|\{v_j \in \mathcal{N}(v_i) | \mathbf{y}_j = k\}|}{|\mathcal{N}(v_i)|}$ denotes the proportion of neighbors of v_i that belong to class k.

Feature Homophily. Feature homophily measures the degree of feature dependencies across the graph topology, reflecting how node features are influenced by graph connectivity. For the linear case of the graph diffusion process, the feature homophily for the m-th feature dimension is defined as:

$$\mathbf{X}_{:,m} = \left(\mathbf{I} - \frac{h_{F,m}}{\rho(\mathbf{A})}\mathbf{A}\right)^{-1} \mathbf{X}^{(0)}_{:,m}, \tag{9}$$

where $\rho(\mathbf{A})$ denotes the spectral radius of $\mathbf{A}$, $\mathbf{X}^{(0)} \sim p(\mathbf{X} \mid \mathbf{y})$ represents structural-agnostic node features, and $\mathbf{X} \sim p(\mathbf{X} \mid \mathbf{y}, \mathbf{A})$ denotes structural-aware node features. The overall feature homophily is defined as:

$$h_F(\mathcal{G}, \mathbf{X}, \mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} h_{F,m}. \tag{10}$$

In practice, since both $h_{F,m}$ and $\mathbf{X}^{(0)}$ are unknown, feature homophily is estimated via the inverse formulation proposed by Zheng et al. (2024). Specifically, for a given $h_{F,m}$, the corresponding

structural-agnostic features are recovered as $\mathbf{X}^{(0)}_{:,m} = \left(\mathbf{I} - \frac{h_{F,m}}{\rho(\mathbf{A})}\mathbf{A}\right)\mathbf{X}_{:,m}$. Then, $h_{F,m}$ is estimated by minimizing the intra-class variation of the recovered features:

$$h^*_{F,m}(\mathcal{G}, \mathbf{X}_{:,m}, \mathbf{y}) = \arg\min_{h_{F,m}} \sum_{\substack{v_i, v_j \in \mathcal{V} \\ \mathbf{y}_i = \mathbf{y}_j}} \left\| X^{(0)}_{i,m} - X^{(0)}_{j,m} \right\|^2. \tag{11}$$

Tri-Hom. $h^{\mathcal{G}}_{\text{Tri-Hom}}$ for graph-aware model and $h^{\neg\mathcal{G}}_{\text{Tri-Hom}}$ for graph-agnostic model are defined as:

$$h^{\mathcal{G}}_{\text{Tri-Hom}} = \frac{\left(\frac{h_L C-1}{C-1}\right)^2}{C\left(\frac{1-h_L}{C-1}\right)^2 + C\frac{(1-h_S)^2}{C-1} + \left(\frac{h_L C-1}{C-1}\right)^2} h^{\neg\mathcal{G}}_{\text{Tri-Hom}} \tag{12}$$

$$h^{\neg\mathcal{G}}_{\text{Tri-Hom}} = \frac{1 - \left(\frac{h_F}{\rho(\mathbf{A})}\right)^2 \left(C\left(\frac{1-h_L}{C-1}\right)^2 + C\frac{(1-h_S)^2}{C-1} + \left(\frac{h_L C-1}{C-1}\right)^2\right)}{\left[1 - \left(\frac{h_F}{\rho(\mathbf{A})}\right)\left(\frac{h_L C-1}{C-1}\right)\right]^2}. \tag{13}$$

G Per-Dataset Metric Values

Table 46: Per-dataset metric values (label-based).

Dataset	h_{edge}	h_{node}	h_{class}	h_{adj}	h_{unb}	h_{NLD}	h_{den}	h_{2hop}	h_{nei}	LI
Roman-Empire	0.0469	0.0460	0.0208	−0.0468	−0.4913	0.8492	0.4994	0.0750	0.2925	0.1101
Amazon-Ratings	0.3804	0.3757	0.1266	0.1402	0.2769	0.8530	0.5000	0.3686	0.5132	0.0398
Minesweeper	0.6828	0.6829	0.0094	0.0094	0.0145	0.7032	0.5000	0.6809	0.7999	0.0001
Tolokers	0.5945	0.6344	0.1801	0.0926	0.1016	0.8467	0.4984	0.6510	0.7390	0.0066
Questions	0.8396	0.8980	0.0790	0.0207	0.0583	0.9248	0.4999	0.9097	0.9342	0.0007
Squirrel	0.2072	0.1905	0.0398	0.0086	0.0197	0.8211	0.4907	0.2239	0.3191	0.0013
Chameleon	0.2361	0.2441	0.0444	0.0295	0.0812	0.8491	0.4917	0.2739	0.4141	0.0139
Actor	0.2167	0.2199	0.0064	0.0028	0.0053	0.8906	0.4999	0.2135	0.3091	0.0002
Texas	0.0609	0.0567	0.0000	−0.2936	−0.8050	0.8307	0.4444	0.5429	0.6946	0.1923
Cornell	0.1227	0.1110	0.0383	−0.2201	−0.4675	0.8780	0.4828	0.3989	0.5819	0.1557
Wisconsin	0.1778	0.1552	0.0461	−0.1733	−0.2851	0.8613	0.4838	0.4239	0.5959	0.1311
Cora	0.8100	0.8252	0.7657	0.7711	0.9200	0.9768	0.5015	0.7407	0.8118	0.5904
CoraFull	0.5670	0.5861	0.4959	0.5558	0.9723	0.9430	0.5012	0.4383	0.5661	0.5221
CiteSeer	0.7355	0.7166	0.6267	0.6707	0.8520	0.9582	0.5010	0.6849	0.8389	0.4508
PubMed	0.8024	0.7924	0.6641	0.6860	0.7653	0.9811	0.5002	0.7435	0.8179	0.4093
Flickr	0.3195	0.3221	0.0698	0.0941	0.2264	0.8766	0.5000	0.3362	0.4657	0.0130
Amazon-Photo	0.8272	0.8493	0.7722	0.7850	0.9329	0.9817	0.5036	0.6576	0.7112	0.6662
Amazon-Computer	0.7772	0.8017	0.7002	0.6823	0.9055	0.9769	0.5015	0.5656	0.6439	0.5279
Coauthor-CS	0.8081	0.8320	0.7547	0.7845	0.9582	0.9811	0.5008	0.6862	0.7274	0.6467
Coauthor-Physics	0.9314	0.9153	0.8474	0.8724	0.9480	0.9899	0.5005	0.8369	0.8655	0.7222
WikiCS	0.6543	0.6774	0.5675	0.5790	0.8569	0.9652	0.5022	0.3710	0.4306	0.3692
Blog-Catalog	0.4011	0.3914	0.2680	0.2724	0.5256	0.8919	0.5029	0.2061	0.2479	0.0918
Ogbn-Arxiv	0.6542	0.6353	0.4211	0.5877	0.9316	0.9488	0.5001	0.5013	0.6251	0.4535
Genius	0.5932	0.5087	0.0229	−0.0527	−0.0705	0.9790	0.5000	0.7216	0.8078	0.0025
Twitch-DE	0.6322	0.5958	0.1394	0.1394	0.1566	0.8856	0.5001	0.5431	0.6568	0.0153
Twitch-ENGB	0.5560	0.5452	0.0852	0.0852	0.0874	0.9085	0.5000	0.5235	0.6060	0.0053
Twitch-ES	0.5800	0.6186	0.1468	0.1036	0.1092	0.8573	0.4995	0.5794	0.6712	0.0081
Twitch-FR	0.5595	0.5739	0.0855	0.0848	0.0876	0.8448	0.5000	0.5349	0.6105	0.0053
Twitch-PTBR	0.5708	0.5949	0.1196	0.1069	0.1104	0.8451	0.4996	0.5418	0.6213	0.0085
Twitch-RU	0.6176	0.6383	0.0424	0.0263	0.0330	0.8652	0.4998	0.6276	0.7470	0.0006
Twitch-TW	0.5332	0.5500	0.0339	0.0328	0.0339	0.8231	0.5000	0.5212	0.5884	0.0008

Table 47: Per-dataset metric values (structure- and feature-based).

Dataset	h_{NS}	h_{agg}	h_S	h_{GE}	$h_{\mathrm{LS\text{-}cos}}$	$h_{\mathrm{LS\text{-}euc}}$	h_{attr}	h_{CF}	h_F
Roman-Empire	3.2151	0.5874	0.5271	0.0257	0.0255	−8.0595	0.0585	0.7781	0.1300
Amazon-Ratings	1.7253	0.4191	0.5256	0.1146	0.1192	−24.40	0.6159	0.5976	1.0000
Minesweeper	2.1402	0.3937	0.7070	0.3330	0.3332	−0.9430	0.1856	0.0930	0.5000
Tolokers	2.1232	0.0000	0.5451	0.8093	0.7705	−0.2957	0.1099	0.2228	1.0000
Questions	23.30	0.0000	0.5576	0.4267	0.5544	−0.1423	0.0050	0.9060	0.7100
Squirrel	1.2125	0.0000	0.6030	0.0138	0.0219	−0.5155	0.0007	−0.4113	0.0000
Chameleon	1.1318	0.1685	0.5367	0.0141	0.0135	−0.5673	0.0006	0.0819	0.0000
Actor	1.0833	0.0000	0.3841	0.1594	0.1566	−0.6630	0.0011	0.0229	0.0000
Texas	4.3899	0.4863	0.5158	0.3487	0.3448	−0.1400	0.0008	−0.0266	0.0000
Cornell	2.5518	0.4098	0.3676	0.3322	0.3391	−0.1455	0.0009	0.1393	0.0000
Wisconsin	3.0825	0.6574	0.4687	0.3414	0.3373	−0.1444	0.0009	0.2975	0.0000
Cora	14.97	0.8789	0.6164	0.1677	0.1780	−0.3338	0.0066	0.1410	0.1500
CoraFull	44.73	0.8626	0.6980	0.1447	0.1604	−0.2263	0.0037	0.0052	0.0000
CiteSeer	6.4521	0.6495	0.3909	0.1906	0.2229	−0.2208	0.0067	0.1487	0.7700
PubMed	3.9050	0.6707	0.3792	0.2719	0.2658	−0.2384	0.0099	0.2326	1.0000
Flickr	2.9056	0.1328	0.6086	0.3815	0.4084	−0.0950	0.0027	0.0924	0.3000
Amazon-Photo	22.04	0.9312	0.7559	0.4878	0.4450	−0.0843	0.0016	0.1062	0.9100
Amazon-Computer	19.77	0.9236	0.7628	0.4897	0.4402	−0.0829	0.0015	0.0337	1.0000
Coauthor-CS	43.58	0.9474	0.7213	0.2944	0.3194	−0.1763	0.0056	0.2355	0.4400
Coauthor-Physics	26.92	0.9424	0.7330	0.3513	0.3723	−0.2160	0.0069	0.3741	0.8800
WikiCS	9.3530	0.7431	0.6366	0.3342	0.3916	−277.15	−0.0209	0.9998	1.0000
Blog-Catalog	1.5237	0.5277	0.6987	0.1223	0.1228	−0.1818	0.0001	−0.0430	0.0000
Ogbn-Arxiv	51.76	0.7367	0.5939	0.8389	0.8649	−0.0546	0.0028	0.2352	1.0000
Genius	2.9629	0.0000	0.0931	0.6656	0.5909	−0.2162	0.1585	0.1898	0.7500
Twitch-DE	1.1644	0.2426	0.5364	0.1961	0.1919	−0.2860	0.0006	0.0782	0.0000
Twitch-ENGB	1.0328	0.1485	0.3853	0.2101	0.2001	−0.2836	0.0007	0.0232	0.0000
Twitch-ES	1.5146	0.0000	0.5700	0.2039	0.1976	−0.2939	0.0007	0.0278	0.0000
Twitch-FR	1.1885	0.0000	0.6009	0.2251	0.2117	−0.2850	0.0007	0.0716	0.0000
Twitch-PTBR	1.2712	0.0000	0.5801	0.2135	0.2030	−0.2860	0.0009	0.0911	0.0000
Twitch-RU	1.7610	0.0000	0.5698	0.1925	0.1902	−0.2829	0.0006	0.0256	0.0000
Twitch-TW	1.1133	0.0000	0.6248	0.1893	0.1914	−0.3037	0.0010	0.0271	0.0000

Table 48: Per-dataset metric values (classifier-based metrics, Tri-Hom, and ours).

Dataset	h_{KR}	h_{GNB}	h_{SVM}	$h^{\neg\mathcal{G}}_{\text{Tri-Hom}}$	$h^{\mathcal{G}}_{\text{Tri-Hom}}$	$r_{\deg}$	$h^{\neg\mathcal{G}}_{\text{glocal}}$	$h^{\mathcal{G}}_{\text{glocal}}$
Roman-Empire	0.0000	0.0000	0.0000	0.9887	0.1525	0.9754	0.9887	0.1524
Amazon-Ratings	0.9992	0.0253	1.0000	1.0064	0.0064	0.8838	1.0062	0.0062
Minesweeper	1.0000	1.0000	1.0000	1.0533	0.6044	0.9989	1.0533	0.6044
Tolokers	0.9982	0.0000	0.0000	1.0901	0.3618	0.5107	1.0862	0.2276
Questions	0.0000	0.6304	1.0000	1.1262	0.7931	0.0240	1.1181	0.3277
Squirrel	1.0000	1.0000	0.0000	1.0000	0.0593	0.5331	1.0000	0.0417
Chameleon	0.9893	1.0000	0.0000	1.0000	0.0233	0.6714	1.0000	0.0193
Actor	0.0000	0.0000	0.0000	1.0000	0.0274	0.2771	1.0000	0.0157
Texas	0.0000	0.0000	1.0000	1.0000	0.1446	0.2698	1.0000	0.0865
Cornell	0.0000	0.0000	1.0000	1.0000	0.0853	0.3414	1.0000	0.0569
Wisconsin	0.0000	0.0000	1.0000	1.0000	0.0693	0.3762	1.0000	0.0442
Cora	1.0000	1.0000	1.0000	1.0223	0.7061	0.6979	1.0223	0.6016
CoraFull	1.0000	1.0000	0.0000	1.0000	0.3513	0.6864	1.0000	0.2582
CiteSeer	0.0000	1.0000	0.0000	1.0886	0.3798	0.7432	1.0880	0.3437
PubMed	0.0000	0.0000	0.0000	1.1409	0.4988	0.5791	1.1378	0.3995
Flickr	0.0000	0.0000	0.0000	0.9985	0.0005	0.1376	0.9975	0.0002
Amazon-Photo	0.0000	0.0000	0.0034	1.1505	0.9769	0.6305	1.1485	0.7559
Amazon-Computer	0.0000	0.0000	1.0000	1.1498	0.9342	0.4615	1.1448	0.5420
Coauthor-CS	0.0000	0.0000	1.0000	1.0678	0.8569	0.8124	1.0677	0.7965
Coauthor-Physics	0.0000	0.0000	1.0000	1.1654	1.0157	0.7909	1.1648	0.9443
WikiCS	0.0000	0.0000	0.0000	1.1058	0.5429	0.4806	1.1013	0.3097
Blog-Catalog	0.0000	0.0000	0.0000	1.0000	0.0180	0.8714	1.0000	0.0170
Ogbn-Arxiv	0.0057	0.0000	0.0000	1.0911	0.4053	0.0065	1.0748	0.1085
Genius	0.0001	0.0000	0.0000	1.0320	0.0481	0.0000	1.0232	0.0237
Twitch-DE	0.0390	0.9831	0.0000	1.0000	0.2584	0.2508	1.0000	0.1075
Twitch-ENGB	0.0571	0.9191	0.0231	1.0000	0.1227	0.3672	1.0000	0.0710
Twitch-ES	0.3051	0.9999	0.0001	1.0000	0.3215	0.4736	1.0000	0.1858
Twitch-FR	0.9968	0.2459	0.0000	1.0000	0.2576	0.3626	1.0000	0.1168
Twitch-PTBR	0.3399	0.9971	0.0009	1.0000	0.2843	0.5589	1.0000	0.1847
Twitch-RU	0.9293	0.0307	0.4450	1.0000	0.3578	0.3012	1.0000	0.1589
Twitch-TW	0.3110	0.0705	0.0001	1.0000	0.2253	0.5307	1.0000	0.1321

Tables 46–48 list every metric value used in the paper. Together with the accuracies in Table 19, they suffice to recompute all real-world correlations (up to rounding).