WHEN DOES DIRECTION HELP? A SEPARATION THEOREM FOR DIRECTED POLYNOMIAL GRAPH FILTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many node-classification datasets are directed, and directed models improve accuracy on some benchmarks and not on others. It remains unclear when direction helps, and why. We study this question for polynomial graph filters. First, we prove a coefficient-budget separation for the self-looped walk that is deployed in practice. On complete binary trees of height H and from a shared root-marking input, a polynomial in the directed walk approximates depth to uniform error $1/4$ with monomial coefficient norm $O(H)$, whereas any polynomial in the undirected walk needs norm at least $2^H/6$, regardless of its degree: an over-squashing effect that the directed walk avoids. Second, from the constant input, symmetric polynomial filters cannot represent label components that are antisymmetric under an orientation-reversing symmetry. From this same input, a polynomial in the strict directed walk represents depth on rooted hierarchies and absorption-time potentials on DAGs exactly. Third, we introduce a supervised linear probe on Krylov features, scored on validation nodes before GNN training. Across 15 directed datasets with random features, its score correlates with the measured directed-minus-symmetric test-accuracy gap at Pearson $r = 0.89$ and agrees with its sign on 14 of 15 datasets; with dataset features, $r = 0.81$ and 12 of 15. Choosing the operator by that sign gives mean test accuracy close to that obtained by training both models and choosing on validation. On hierarchy tasks the directed filter improves over its symmetric counterpart and is competitive with the directed baselines MagNet and Dir-GNN. On most standard control graphs the symmetric filter remains the better choice. Direction helps when labels depend on orientation and are not already captured by node features.

1 INTRODUCTION

Many graph datasets are *directed*: citation networks, hypernym hierarchies, web link graphs, and dependency graphs all carry edges whose orientation is part of the signal. Yet the most widely used spectral graph neural networks and their polynomial spectral filters are built on the *symmetric* normalized Laplacian (Kipf & Welling, 2017; Defferrard et al., 2016). This operator discards orientation and depends only on the underlying undirected graph, because it gives a real-valued, self-adjoint operator with an orthogonal eigenbasis on which real-valued polynomial filters can be defined.

A recent line of work argues empirically that discarding direction loses information. MagNet (Zhang et al., 2021) encodes orientation in the phase of a complex Hermitian Laplacian, DiGCN (Tong et al., 2020) builds a directed propagation rule from PageRank, and Dir-GNN (Rossi et al., 2024) keeps separate in- and out-neighborhood aggregations, each reporting gains over symmetric baselines on some directed benchmarks. The gains are uneven: direction helps on some datasets and hurts on others. Existing theory addresses the expressive power of directed message passing (Rossi et al., 2024), of spectral filters (Wang & Zhang, 2022; Koke & Cremers, 2024) and of random node features (Abboud et al., 2021; Sato et al., 2021). **It does not say, for a given graph and label, *when* a directed filter should be preferred to a symmetric one.** A user cannot tell, before training, whether orientation carries label-relevant information or merely noise a directed operator will overfit.

Our main theoretical result quantifies the cost of symmetrizing a directed walk. With the same root-marking input, approximating depth on complete binary trees requires a monomial coefficient norm that is linear in the height for the directed walk but exponential for the undirected walk. We then study operator choice empirically, using matched polynomial filters (DIR-POLY, the directed

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.

instantiation of GPR-GNN (Chien et al., 2021), and its symmetric counterpart SYM-POLY) and a supervised probe that compares their validation performance. We make four contributions:

1. **A coefficient-budget separation for the deployed walk.** We consider complete binary trees of height H and a shared root-marking input. A polynomial in the self-looped directed walk approximates depth to uniform error $1/4$ with monomial coefficient norm $O(H)$, whereas any polynomial in the undirected walk needs norm at least $2^H/6$ (Theorem 4). The lower bound holds at every polynomial degree: more propagation steps do not remove the exponential requirement on the coefficient norm. Numerical linear programs show a larger separation.
2. **Same-input representability results.** We prove that *symmetric* polynomial filters cannot represent the orientation-antisymmetric part of a node label from a symmetry-invariant input. The same holds in expectation over exchangeable features for fixed equivariant maps (Theorem 1). From the same constant input, the strict directed walk represents depth on a rooted tree, and the absorption-time potential on a DAG, *exactly* (Theorems 2–3).
3. **A predictor of when direction helps.** We introduce a polynomial-expressibility score (a supervised linear probe on Krylov features, scored without test labels). It tracks the measured directed-minus-symmetric accuracy gain at Pearson $r = 0.89$ across 15 directed datasets, where four simpler predictors do not. Its sign selects about as well as training both models and choosing on validation, while training only the selected model. Its benefit lies in the structural regime of hierarchy and ordering labels.
4. **Experiments with matched operators and baselines.** In our experiments DIR-POLY and SYM-POLY differ only in whether the edge set is symmetrized. On four hierarchy and ordering benchmarks DIR-POLY is within one point of Dir-GNN and ahead of MagNet on three. On eight control graphs the score is negative throughout and the symmetric filter is better on seven. On a citation-era label Dir-GNN and MagNet are the strongest models. Making the walk self-adjoint removes its skew component, and the resulting filter trails the raw-walk filter on the hierarchies.

2 Related Work

Symmetric spectral filters. Graph neural networks follow message passing (Gilmer et al., 2017; Hamilton et al., 2017), with GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), and GIN (Xu et al., 2019) canonical. Our method belongs to the spectral-filter subfamily, where a layer applies a function of the graph Laplacian (Bruna et al., 2014; Shuman et al., 2013): ChebNet (Defferrard et al., 2016) uses a Chebyshev polynomial; APPNP (Gasteiger et al., 2019) a personalized-PageRank kernel; GPR-GNN (Chien et al., 2021) learnable coefficients; BernNet (He et al., 2021) a Bernstein basis for non-negative responses. Later designs learn the basis (ChebNetII (He et al., 2022a), Specformer (Bo et al., 2023)), and Wang & Zhang (2022) analyze expressive power. All operate on a *symmetric* operator M, so each output channel is a polynomial $\sum_k c_k M^k x$ in M applied to a node signal x. Our impossibility result (Theorem 1) applies to the whole family: whenever the undirected graph has an *involutive* automorphism σ and the label a σ-antisymmetric component y_-, any such filter applied to a σ-invariant input incurs error at least $\|y_-\|^2$, whatever its degree, basis, or coefficients. It concerns the *operator* as a map, independently of training. The theorem motivates our directed-versus-symmetric comparison. The accuracy gaps we measure on benchmarks are empirical: there its assumptions and a nonzero bound are not established.

Directed GNNs and directed spectral theory. A separate line keeps edge direction. MagNet (Zhang et al., 2021) and MSGNN (He et al., 2022b) encode direction through a complex Hermitian "magnetic" Laplacian with charge q. MagNet and Dir-GNN are our directed baselines. DiGCN (Tong et al., 2020) builds a PageRank digraph Laplacian, and Dir-GNN (Rossi et al., 2024) separates in- and out-neighbor aggregation (survey in App. C). These establish that direction can help, and Rossi et al. (2024) and Koke & Cremers (2024) analyze the expressive power of directed models. We differ in two ways: our filter is a polynomial in a *real* directed walk operator, needing no phase hyperparameter and no eigendecomposition. We also contribute theory: an impossibility and expressibility separation that identifies when direction helps, plus a predictor of the gain. Comparisons appear in §7 and App. C.

Relation to GPR-GNN and Dir-GNN. DIR-POLY is a learnable polynomial $\sum_{k=0}^{K} \gamma_k P^k z$ in the directed walk operator P (§3) applied to hidden node features z, structurally GPR-GNN (Chien

et al., 2021) over the *raw directed* walk. Its bidirectional variant BIDIR-POLY, adding $\sum_k \delta_k (P^\top)^k z$, is a polynomial rendering of Dir-GNN (Rossi et al., 2024). Both directed filters lie *outside* the symmetric impossibility of Theorem 1, which constrains only symmetric operators.

Oversmoothing, over-squashing and heterophily. Oversmoothing under repeated symmetric propagation (Li et al., 2018; Oono & Suzuki, 2020; Cai & Wang, 2020), with remedies such as normalization and oscillatory propagation (Zhao & Akoglu, 2020; Guo et al., 2023; Rusch et al., 2022), is orthogonal to orientation. Over-squashing (Alon & Yahav, 2021; Topping et al., 2022; Di Giovanni et al., 2023) is what Theorem 4 quantifies for polynomial filters on trees. Several of our control graphs are heterophily benchmarks (Pei et al., 2020; Zhu et al., 2020). Platonov et al. (2023) identify duplicated nodes in Chameleon and Squirrel, and we use the standard, unfiltered versions.

3 PRELIMINARIES

Directed graphs and the random walk. Let $G = (V, E)$ be a directed graph with $|V| = n$ nodes. Adding a self-loop per node and row-normalizing the adjacency gives a *row-stochastic* random walk $P \in \mathbb{R}^{n\times n}$ ($P_{ij} \geq 0$, $\sum_j P_{ij} = 1$): the probability of stepping from i to j along an out-edge. Unlike the symmetrized adjacency, P retains the *orientation* of E ($P \neq P^\top$ in general), so a hierarchical or acyclic graph induces a non-reversible walk.

Teleportation and primitivity. A directed walk can be reducible or periodic, leaving P without a unique positive stationary distribution. Following PageRank (Page et al., 1999), we restore primitivity by teleportation:

$$A_\alpha \;=\; \alpha P \;+\; (1-\alpha)\,\tfrac{1}{n}\,\mathbf{1}\mathbf{1}^\top, \qquad \alpha \in (0, 1], \tag{1}$$

with $\mathbf{1} \in \mathbb{R}^n$ the all-ones vector. For any $\alpha < 1$, A_α is primitive, so by Perron–Frobenius it has a unique strictly positive stationary π (App. A):

$$\pi > 0, \qquad \pi A_\alpha = \pi, \qquad \textstyle\sum_i \pi_i = 1, \tag{2}$$

collected into $\Pi = \mathrm{diag}(\pi)$. We use $\alpha = 0.9$ by default. Teleportation makes the stationary weighting well defined for every graph, including the rooted hierarchies below.

The stationary inner product $\ell^2(\pi)$. The measure π defines a reweighted inner product on $\mathbb{R}^n$,

$$\langle x, y\rangle_\pi \;=\; \sum_i \pi_i\, x_i\, y_i \;=\; x^\top \Pi\, y, \tag{3}$$

$\ell^2(\pi)$ being $\mathbb{R}^n$ under this product. Here A_α is generally *not* self-adjoint. Its adjoint is the time-reversed walk $\Pi^{-1} A_\alpha^\top \Pi$.

Reversibilization and the skew (orientation) part. We define the $\ell^2(\pi)$-self-adjoint *reversibilized* operator and its skew complement,

$$\hat{A} = \tfrac{1}{2}\left(A_\alpha + \Pi^{-1} A_\alpha^\top \Pi\right), \tag{4}$$

$$\hat{S} = \tfrac{1}{2}\left(A_\alpha - \Pi^{-1} A_\alpha^\top \Pi\right), \tag{5}$$

so $A_\alpha = \hat{A} + \hat{S}$. Under $\langle\cdot,\cdot\rangle_\pi$, $\hat{A}$ is self-adjoint and $\hat{S}$ is skew-adjoint ($\langle \hat{S}x, y\rangle_\pi = -\langle x, \hat{S}y\rangle_\pi$). In terms of the stationary edge flows $F = \Pi A_\alpha$, $\Pi\hat{A} = \frac{1}{2}(F + F^\top)$ is their symmetric part and $\Pi\hat{S} = \frac{1}{2}(F - F^\top)$ their antisymmetric part: passing from A_α to $\hat{A}$ keeps the symmetric flows and removes the net flow along each edge. The Laplacian-like $\tilde{L} = (I - \hat{A})/2$ has spectrum in $[0, 1]$ and is used only where a self-adjoint spectral filter is required.
The symmetric counterparts are built on the symmetrized edge set: P_{sym}, the self-looped random walk of the undirected graph (reversible, with no teleportation), and $\hat{A}_{\mathrm{sym}}$, the same teleport-and-reversibilize pipeline applied to it (App. C). Both depend only on the undirected graph, so both commute with its automorphisms.

Polynomial filters and Krylov subspaces. All filters we study are degree-K polynomials in a fixed operator M applied to a node signal x, with reachable set the Krylov subspace

$$\mathcal{K}_K(M, x) \;=\; \mathrm{span}\{x,\; Mx,\; M^2x,\; \ldots,\; M^K x\}, \tag{6}$$

so expressibility reduces to membership in $\mathcal{K}_K(M, x)$.

Problem setting. We study *transductive node classification*: given G and labels on a training subset of V, predict the rest, with the whole graph visible at training. To isolate *directed topology* from node content, we assign every node a fixed i.i.d. *random* feature vector, so a node's class (a label such as depth or topological order encoded in the edge *orientation*) must be recovered from structure alone. Removing feature shortcuts isolates the expressibility of the two operators in the directed-vs-symmetric comparison.

4 The Orientation-Separation Theorem

We separate polynomial filters in a directed walk from those in its symmetrized counterpart, with both applied to the same input. From a symmetry-invariant input, symmetric filters cannot represent label components that are antisymmetric under an orientation-reversing symmetry (Theorem 1), whereas the strict directed walk represents depth exactly (Theorems 2–3). From a shared root-marking input, the deployed directed walk approximates depth on complete binary trees with a coefficient norm linear in the height, and the undirected walk requires an exponential one (Theorem 4, §4.3; our main result). On real hierarchies we measure expressibility (§4.4).

4.1 Orientation impossibility for symmetric filters

We fix an *involutive* automorphism σ of the *undirected* graph and write R for its orthogonal permutation matrix, so that $R^2 = I$. Since $\hat{A}_{\mathrm{sym}}$ depends only on the symmetric adjacency and degree, σ commutes with it:

$$R\,\hat{A}_{\mathrm{sym}} = \hat{A}_{\mathrm{sym}}\,R. \tag{7}$$

With $R^2 = I$, R has eigenvalues ± 1, and any label $y \in \mathbb{R}^n$ splits orthogonally as $y = y_+ + y_-$ with $Ry_+ = y_+$, $Ry_- = -y_-$. The component y_- is the part of the label that changes sign under σ, nonzero exactly when the label is not σ-invariant; $\mathrm{Fix}(R) = \{z : Rz = z\}$ is the σ-invariant subspace, and $\mathrm{dist}(y, \mathrm{Fix}(R))^2 = \|y_-\|^2$.

Theorem 1 (Orientation impossibility)**.** *Let σ be an involutive automorphism ($R^2 = I$) of the undirected graph, so that $R\hat{A}_{\mathrm{sym}} = \hat{A}_{\mathrm{sym}}R$.* (A, deterministic.) *If the seed x is σ-invariant ($Rx = x$), then any polynomial filter output $\sum_{k=0}^{K} c_k \hat{A}_{\mathrm{sym}}^k x$ lies in $\mathrm{Fix}(R)$. Hence the output $\hat{y}$ of any read-out applied node-wise with shared weights satisfies*

$$\|\hat{y} - y\|^2 \;\geq\; \mathrm{dist}\big(y, \mathrm{Fix}(R)\big)^2 \;=\; \|y_-\|^2.$$

(B, in expectation.) *If the features X are σ-exchangeable ($RX \overset{d}{=} X$; i.i.d. Gaussian features qualify), then any* fixed *node-wise-equivariant predictor built from $\hat{A}_{\mathrm{sym}}$, with parameters fixed independently of the feature realization, satisfies*

$$\mathbb{E}\|\hat{y} - y\|^2 \;\geq\; \tfrac{1}{4}\,\|y - R^{-1}y\|^2 \;=\; \|y_-\|^2.$$

Proof sketch. (A) For a σ-invariant seed, (7) gives $\hat{A}_{\mathrm{sym}}^k x \in \mathrm{Fix}(R)$, and a node-wise read-out preserves $\mathrm{Fix}(R)$, so $\hat{y} \in \mathrm{Fix}(R)$ and the residual is at least $\mathrm{dist}(y, \mathrm{Fix}(R))^2 = \|y_-\|^2$. *(B)* Node-wise equivariance with $RX \overset{d}{=} X$ gives $\mathbb{E}\|\hat{y} - y\|^2 = \mathbb{E}\|\hat{y} - R^{-1}y\|^2$, which by a parallelogram bound is $\geq \frac{1}{4}\|y - R^{-1}y\|^2 = \|y_-\|^2$. The full proof is in Appendix A. □

The obstruction follows from equivariance alone: it holds for *any* degree, coefficients and capacity. Part (B) replaces invariance of each input by invariance of the input *distribution*, and gives an expected-error bound for a predictor whose parameters are fixed independently of the feature realization. A model fitted on the same realized features, with asymmetric labels and masks, falls outside that bound. The directed walk breaks (7): for an orientation-reversing σ, $RPR^\top$ is the walk of the edge-reversed graph. Hence $RP \neq PR$ and the argument above no longer confines the output to $\mathrm{Fix}(R)$. Noncommutation alone does not make y_- reachable (P fixes the constant input). Theorems 2–3 identify a directed operator and an input that reach it.

4.2 The directed walk expresses depth exactly

The positive result is explicit on the canonical orientation label, and it uses the same input as Theorem 1(A): the constant vector $\mathbf{1}$, which is invariant under every automorphism. Let T be a rooted tree of height H and U its *strict* upward (child→parent) walk: $U_{v,\mathrm{pa}(v)} = 1$ off the root ($\mathrm{pa}(v)$ is the parent of v) and a *zero row at the root*, so $U\mathbf{1} = \mathbf{1} - e_{\mathrm{root}}$ (e_{root} is the root indicator). The depth function d ($d(\mathrm{root}) = 0$) carries the orientation signal. On a directed path, reflection makes centered

depth σ-antisymmetric. On a general rooted tree, whether Theorem 1 gives a nonzero bound for depth depends on the available automorphism.

Theorem 2 (Exact depth expressibility)**.** *On a rooted tree of height H with strict upward walk U, depth is an exact degree-H polynomial in U applied to the constant input:*

$$d \;=\; \sum_{k=1}^{H} U^k\,\mathbf{1} \;=\; \sum_{k=0}^{H-1} U^k\big(\mathbf{1}-e_{\mathrm{root}}\big) \;\in\; \mathcal{K}_H\big(U,\,\mathbf{1}\big), \qquad U^{H+1}\mathbf{1}=0.$$

The proof is in App. A.

The argument generalizes to arbitrary directed acyclic graphs (DAGs), the structure of real citation and hypernym hierarchies. On a multi-parent DAG the exactly expressible label becomes the parent-*averaged* depth potential, with truncation degree the longest directed path.

Theorem 3 (Exact potential expressibility on a DAG)**.** *Let G be a finite directed acyclic graph with edges oriented child→parent, let $d^+(v)$ be the out-degree of v, let $\mathcal{R}=\{v : d^+(v)=0\}$ be its set of* roots *(the sinks at which the upward walk absorbs), and let L be the length of its longest directed path. Let $U = D_{\mathrm{out}}^{-1}A$, with A the adjacency matrix and D_{out} the diagonal out-degree matrix, be the strict, self-loop-free* parent-averaging *operator, with $U_{vu}=1/d^+(v)$ for each parent u of a node $v\notin\mathcal{R}$ and $U_{v\cdot}=0$ for $v\in\mathcal{R}$ (hence substochastic), so that $b:=U\mathbf{1}=\mathbf{1}-\mathbf{1}_{\mathcal{R}}$ (1 off the roots, 0 on $\mathcal{R}$). Then U is nilpotent ($U^{L+1}=0$), and the expected-steps-to-root potential t, the unique solution of $t(v)=1+\mathrm{mean}_{u\in\mathrm{pa}(v)}\,t(u)$ off $\mathcal{R}$ with $t|_{\mathcal{R}}=0$, is exactly a degree-L polynomial in U applied to the constant input:*

$$t \;=\; \sum_{k=1}^{L} U^k\,\mathbf{1} \;=\; \sum_{k=0}^{L-1} U^k\big(\mathbf{1}-\mathbf{1}_{\mathcal{R}}\big) \;\in\; \mathcal{K}_L\big(U,\,\mathbf{1}\big).$$

On a single-parent DAG with one root this is Theorem 2.

The proof is in App. A.

Strict and deployed walks. (a) The zero-root convention is essential. A walk with a root self-loop, and the deployed self-looped walk P, are row-stochastic, so every polynomial in them maps the constant input to a constant. Our theorems establish expressibility for U. That a trained DIRPOLY, which applies P to non-constant features, recovers depth is an empirical finding (§7). (b) On a single-parent tree $P=\frac{1}{2}(I+U_{\mathrm{a}})$ with U_{a} the absorbing-root walk, so depth is also an exact polynomial in P applied to the root-marking seed $\mathbf{1}-e_{\mathrm{root}}$. The root marker breaks automorphisms that move the root, and a symmetric filter can use it too (on a three-node path it recovers depth exactly at degree two). With this input the separation is one of resources (Theorem 4). (c) On multi-parent DAGs the exact label is the averaging potential t, which differs from shortest- and longest-path depth.

Corollary 1 (Depth/order requires direction)**.** *On a directed path the orientation-reversing reflection $\sigma(i)=n+1-i$ is an involutive automorphism of the undirected path, the constant input is σ-invariant, and the centered position is σ-antisymmetric. From this one input, Theorem 1(A) confines every symmetric polynomial filter to $\mathrm{Fix}(R)$, with error at least $\|y_-\|^2$, while Theorem 2 represents position exactly with U. The positive side extends to any rooted tree and, by Theorem 3, to any DAG. The lower bound is nonzero only when an automorphism changes the label, so on general trees we instead measure expressibility (§4.4).*

4.3 A COEFFICIENT-BUDGET SEPARATION FOR THE DEPLOYED WALK

We compare the deployed self-looped walk P with its undirected counterpart on a shared non-constant input.

Theorem 4 (Coefficient-budget separation)**.** *Let T be the complete binary tree of height $H\ge 2$ with depth function d, let $b=\mathbf{1}-e_{\mathrm{root}}$, let $P=\frac{1}{2}(I+U_{\mathrm{a}})$ be the self-looped upward walk and P_{sym} the self-looped random walk of the undirected tree.* (i) *For $K\ge 4H$, $\big\|\frac{1}{2}\sum_{k<K}P^kb-d\big\|_\infty\le 5e^{-K/8}$: depth is approximated within any $\varepsilon>0$ by a polynomial in P with monomial coefficient norm $\|\gamma\|_1=\sum_k|\gamma_k|=K/2=O(H+\log(1/\varepsilon))$.* (ii) *If a polynomial $q(x)=\sum_k\gamma_kx^k$ of any degree and a constant c satisfy $\|q(P_{\mathrm{sym}})b+c\mathbf{1}-d\|_\infty\le\frac{1}{4}$, then $\|\gamma\|_1\ge 2^H/6$.*

Proof sketch. (i) $(P^kb)(v)$ is the probability that the lazy upward walk from v is not yet absorbed at the root after k steps. The absorption time is a sum of $d(v)$ independent Geometric$(\frac{1}{2})$ variables

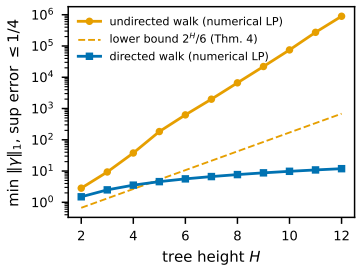


Figure 1: Smallest $\|\gamma\|_1$ approximating depth to uniform error $\frac{1}{4}$ on complete binary trees from the shared input $\mathbf{1} - e_{\text{root}}$ (numerical LPs on the exact level-chain reduction, degree $\leq 8H$) and the any-degree bound of Theorem 4.

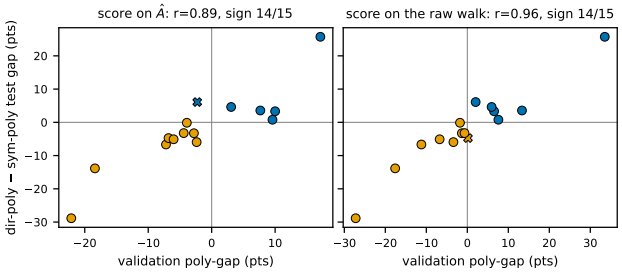


Figure 2: Validation poly-gap against the measured DIR-POLY minus SYM-POLY test gap on 15 directed datasets (random features): score on $\hat{A}$ (left) and raw-walk variant (right). Blue: directed filter better; orange: symmetric. Crosses: sign disagreements.

with mean $2d(v)$, and a Hoeffding bound controls the tail beyond $K \geq 4H$. (ii) $P_{\text{sym}}^k \mathbf{1} = \mathbf{1}$, so the output is a constant minus $\sum_k \gamma_k P_{\text{sym}}^k e_{\text{root}}$. By reversibility (stationary distribution $\propto \deg +1$) and the symmetry of each level, $P_{\text{sym}}^k(v, \text{root}) \leq \frac{3}{\deg(v)+1} 2^{-d(v)}$. A leaf and its parent must be separated by at least $\frac{1}{2}$, which forces $3\|\gamma\|_1 2^{-H} \geq \frac{1}{2}$. The full proof is in App. A. □

Part (ii) is an over-squashing statement (Alon & Yahav, 2021; Topping et al., 2022; Di Giovanni et al., 2023) specialized to polynomial filters: the root's marker is diluted over the 2^j nodes at depth j. The upward walk has out-degree one, so nothing is diluted. He et al. (2025) study reachability-preserving rewiring for directed GNNs. We instead compare directed and undirected polynomial filters on a fixed tree. Numerical linear programs on the exact level-chain reduction (degree up to $8H$) show a larger separation than the bound (Figure 1): at $H = 12$ the smallest $\|\gamma\|_1$ reaching error $\frac{1}{4}$ is 11.9 for the directed walk and 8.9×10^5 for the undirected one. The theorem concerns scalar polynomial regression with a specified input and the monomial coefficient norm, a resource that training does not constrain explicitly. Optimization, generalization, parameter count, runtime and basis choice are outside its scope, as are trained models with random features or nonlinear readouts.

4.4 Expressibility measured on real hierarchies

Root-preserving automorphisms of a rooted tree preserve depth, so on general trees the lower bound of Theorem 1 can be zero. We therefore *measure* expressibility as the held-out coefficient of determination R^2 of degree-10 Krylov fits in double precision (Table 10). From the constant input, U fits depth at $R^2 = 0.98$–1.00 on a random tree and the three WordNet hierarchies, while every row-stochastic symmetric operator returns a constant. From the root-marking seed the symmetric walk reaches 0.43–0.50 against 0.86–1.00 for U and P. From the 16-dimensional random features our experiments use, a ridge probe on the Krylov basis of P reaches 0.49–0.92 against at most 0.19 for the symmetric walk (App. B, Table 7).

5 Predicting When Direction Helps

We want to estimate the directed advantage for a *given* graph and label before training a GNN. We develop a GNN-free, eigendecomposition-free score (a supervised linear probe on Krylov features) and test it as a predictor of the measured gap and as a model-selection rule.

The polynomial-expressibility score. For operator M and features X we form the order-K Krylov basis $\mathcal{K}_K(M, X) = [X, MX, \ldots, M^K X]$, fit a multinomial logistic-regression readout on the training nodes, and record its accuracy on the *validation* nodes. No test label is used. We define the *polynomial-expressibility gap* as

$$\text{poly-gap} \;=\; \text{acc}\Big(\mathcal{K}_K(\hat{A}, X)\Big) \;-\; \text{acc}\Big(\mathcal{K}_K(\hat{A}_{\text{sym}}, X)\Big)\,,$$

with $K = 8$ and the same teleportation $\alpha = 0.9$ on both sides: two self-adjoint operators. We also report a *raw-walk* variant on P and P_{sym}, the operators the trained filters use. Either score costs K operator applications per side and two logistic-regression fits.

Table 1: Operator selection on 15 directed datasets (random features, five splits): mean test accuracy (%) of the model chosen within each split.

selection strategy	mean test accuracy
always directed (DIR-POLY)	61.2
always symmetric (SYM-POLY)	63.1
poly-gap rule, score on $\hat{A}$	65.6
train both, choose on validation	65.7
oracle (better model on each split's test nodes)	66.7

The scores track the measured improvement. Our target is the test-accuracy gap between trained DIR-POLY and SYM-POLY (the comparison of §7; 5 seeded splits, random features). We measure it on 15 directed datasets: three WORDNET hypernym subtrees, two DSBM settings ($k = 5, 8$ blocks), two food webs and eight control graphs (citation, social and heterophilous web graphs). cit-HepTh is evaluated separately (App. B). Across dataset means the score on $\hat{A}$ tracks the gap at Pearson $r = 0.89$ (bootstrap 95% CI $[0.69, 0.97]$), Spearman $\rho = 0.86$, with sign agreement on $14/15$ (two-sided binomial $p = 10^{-3}$). The raw-walk score reaches $r = 0.96$ ($[0.81, 0.99]$), $\rho = 0.90$, and $14/15$ (Figure 2). Leave-one-family-out keeps $r \geq 0.84$ and $r \geq 0.93$. Without the two largest gaps (WordNet *animal*, Telegram) $r = 0.81$ and 0.87 (App. B). Within groups the association is weaker: among the five hierarchy and ordering datasets $r = 0.78$ but $\rho = 0.10$, and among the eight controls $r = 0.94$ falls to 0.46 without Telegram (Table 5). Simpler predictors scored the same way do not track the gap: static π-alignment ($r = -0.07$), label propagation (-0.50), edge reciprocity (0.03) and cyclicity (-0.42).

The score as a model-selection rule. We also test the score as a selection rule with no fitted parameter: within each split, we use the directed filter if that split's validation poly-gap is positive and the symmetric filter otherwise (Table 1). The rule comes within 0.1 points of training both models to completion and choosing on validation, at about one eighth of the selection cost (the probe also requires training the selected model; full validation selection already includes that training). Its gain over always using DIR-POLY is $+4.3$ (hierarchical-bootstrap 95% interval $[0.7, 8.9]$) and over the symmetric default $+2.5$ ($[0.0, 6.3]$). The raw-walk score gives 65.4, and choosing after a short training run of the probe's cost 64.3 (App. B). The benefit is confined to the structural regime: on the eight control graphs with their dataset features (App. B) the rule is below the symmetric default (71.4 vs. 72.5). A post-hoc family rule (directed for WordNet and DSBM, symmetric elsewhere) selects as well as the probe (65.6 with random features), so these results do not show that the probe adds information beyond recognizing the dataset family. On the standard arxiv-year benchmark (169,343 nodes) the rule matches training both models in half the time (App. B).

6 METHOD: DIR-POLY

DIR-POLY applies a learnable polynomial in the self-looped directed walk P to an MLP transformation of the node features. The relation between P and the strict walk U of the theorems is in §4.

The filter. Given $X \in \mathbb{R}^{n \times d_0}$, we compute $Z = \mathrm{MLP}(X) \in \mathbb{R}^{n \times w}$ and aggregate with a degree-K polynomial of the raw directed walk P (Section 3):

$$\text{DIR-POLY}(X) \;=\; \sum_{k=0}^{K} \gamma_k \, P^k \, Z, \qquad Z = \mathrm{MLP}(X), \tag{8}$$

followed by a linear node-wise readout and softmax cross-entropy. The coefficients $\gamma = (\gamma_0, \ldots, \gamma_K) \in \mathbb{R}^{K+1}$ are learned jointly with the MLP. Teleportation is used only to define π for $\hat{A}$ and the predictor (Section 3). Inside DIR-POLY the walk is un-teleported. Replacing P by the teleported walk A_α changes accuracy by at most 0.3 points on the four hierarchy and ordering tasks, within seed-to-seed variation.

Evaluation by Horner's rule. Equation (8) is evaluated by Horner's rule in exactly K sparse matrix products, with no eigendecomposition, $O(K\,|E|\,w)$ time and $O(n\,w)$ memory (App. C, Figure 4).

Hyperparameters. The polynomial filters in the main experiments use $K = 10$ (the probe of §5 uses $K = 8$). App. C.3 sweeps K and teleportation for BIDIR-POLY on cit-HepTh.

A bidirectional variant BIDIR-POLY adds a channel that propagates with $P^\top$ (a polynomial rendering of Dir-GNN) for labels depending on both edge directions (App. C).

Choice of the raw walk. Reversibilizing the walk into the self-adjoint $\hat{A} = A_\alpha - \hat{S}$ supplies a real spectrum and a Dirichlet-energy interpretation, and it removes the skew component $\hat{S}$ by construction (App. A.1). How much label-relevant information that costs is an empirical question: the reversibilized filter trails DIR-POLY by 15.2, 3.4 and 2.3 points on the WordNet tasks and is 0.3 above it on DSBM (Table 2). DIR-POLY therefore filters the raw walk P.

7 EXPERIMENTS

The experiments test whether a directed polynomial filter helps when the label is encoded in directed hierarchical structure and not otherwise.

7.1 SETUP

Datasets. We use four hierarchy and ordering benchmarks, the cit-HepTh semantic-label benchmark, and eight standard benchmark graphs as controls. **WordNet**: three noun hypernym subtrees (*animal*, *plant*, *artifact*; 1500 nodes each) labeled by a depth-quantile bin (ties in integer depth leave 3 classes for *animal* and 2 for *plant* and *artifact*) (Miller, 1995). Depth is an orientation-dependent structural proxy for the potential of Theorem 3, and coincides with that potential in the single-parent tree case. **DSBM**: an ordered directed stochastic block model (k=5 blocks) with reciprocal within-block edges and one-way edges from each block to the next, so block membership is also visible to a symmetric model while the block order is carried by orientation. **cit-HepTh**: the SNAP arXiv citation graph (Leskovec et al., 2005; Leskovec & Krevl, 2014), a 4000-node dense core (54,269 edges, directedness 0.99) labeled by a 4-class publication-era bin (external metadata). This is the semantic-label benchmark, recoverable from citation direction since citations point backward in time. The eight **controls** are Telegram, Cora_ml, Citeseer (Zhang et al., 2021), Chameleon, Squirrel, Cornell, Texas, and Wisconsin (Pei et al., 2020), in their standard PyTorch Geometric versions.

Training and evaluation. We use *random node features* on every dataset to isolate structural information. App. B reports results with dataset features. The polynomial filters share one *architecture-matched* setup ($Z = \mathrm{MLP}(X)$, a degree-K operator polynomial with K=10, Horner evaluation, no eigendecomposition; §6; linear node-wise readout under softmax cross-entropy), *differing only in the propagation operator*. SYM-POLY uses the symmetrized walk, REV the reversibilized self-adjoint $\hat{A}$ (teleportation α=0.9), and DIR-POLY (ours) the raw directed walk P. A DIR-POLY vs. SYM-POLY gap is therefore due to the operator alone. We give the external directed baselines MAGNET (Zhang et al., 2021) and Dir-GNN (Rossi et al., 2024) the same training budget, and we select MagNet's configuration and charge q and Dir-GNN's depth once per dataset by mean validation accuracy. We report mean test accuracy over 5 seeded runs. The real-data graphs are fixed, and the DSBM graph is generated independently per seed (App. C).

7.2 HIERARCHY AND ORDERING BENCHMARKS

Table 2 reports the comparison on the four hierarchy and ordering benchmarks. DIR-POLY reaches 99.5–99.8 on WordNet, and the largest gap appears on *animal* (+25.7 over SYM-POLY). REV, which keeps the symmetric part of the stationary flows and removes the skew part, reaches $84.3_{\pm 2.1}$ there (SYM-POLY: 73.8) and trails DIR-POLY by about 15 points. On DSBM REV is 0.3 points above DIR-POLY. The bidirectional BIDIR-POLY (App. C.1) essentially matches DIR-POLY here (animal 99.1, plant 99.7, artifact 99.3, DSBM 98.7).

7.3 COMPARISON TO DIRECTED BASELINES

MagNet encodes orientation through the charge q of its complex Hermitian Laplacian. Dir-GNN aggregates in- and out-neighbors separately. We give MagNet its standard two-layer, order-2 form and two forms whose propagation reach matches the degree-10 polynomial. Validation selects order 2 with five layers on every task. DIR-POLY exceeds MagNet on the three WordNet tasks, by 5.6 points on *animal* and 0.9–1.0 on the other two. MagNet is 0.8 ahead on DSBM (App. C, Tab. 9). A four-layer Dir-GNN is within one point of DIR-POLY on every task, so on these tasks the gain comes from using direction and is shared across directed architectures. On the semantic citation-era label Dir-GNN (77.0) and MagNet (76.7) lead by a wide margin over BIDIR-POLY (65.1), DIR-POLY (61.3) and SYM-POLY (52.2). The full table is in App. C.

Table 2: Main result. SYM-POLY, REV, and DIR-POLY are architecture-matched degree-K polynomial filters (K=10), differing only in the propagation operator (SYM-POLY: symmetrized; REV: reversibilized self-adjoint; DIR-POLY: raw directed walk, ours). Dir-GNN and MAGNET are external directed baselines with the same training budget (depth, and MagNet's order, layers and charge q, selected on validation). Random features, 5 seeded runs (fixed real-data graphs; DSBM regenerated per seed). The first three rows are WordNet hypernym subtrees. Mean$_{\pm}$sample standard deviation of test accuracy (%).

dataset	SYM-POLY	REV	DIR-POLY	Dir-GNN	MAGNET
animal	$73.8_{\pm 6.5}$	$84.3_{\pm 2.1}$	$\mathbf{99.5}_{\pm 0.6}$	$98.8_{\pm 1.0}$	$93.9_{\pm 0.9}$
plant	$95.2_{\pm 0.9}$	$96.4_{\pm 1.5}$	$\mathbf{99.8}_{\pm 0.2}$	$\mathbf{99.8}_{\pm 0.2}$	$98.8_{\pm 0.8}$
artifact	$96.0_{\pm 0.5}$	$97.2_{\pm 0.8}$	$99.5_{\pm 0.5}$	$\mathbf{99.7}_{\pm 0.3}$	$98.5_{\pm 0.5}$
DSBM (k=5)	$97.1_{\pm 0.8}$	$99.2_{\pm 0.6}$	$98.9_{\pm 0.3}$	$99.4_{\pm 0.6}$	$\mathbf{99.7}_{\pm 0.1}$

7.4 ORIENTATION AND NODE FEATURES ARE COMPLEMENTARY

We vary how informative the node features are on WordNet *animal*. As feature informativeness rises the directed advantage falls from $+25.7$ to 0 (App. C, Fig. 5). The same complementarity holds on a real-feature large graph (ogbn-arxiv; App. C).

7.5 WHEN DIRECTION DOES NOT HELP

On the eight control graphs DIR-POLY underperforms SYM-POLY on seven (mean -5.7; Chameleon is the exception at $+6.1$), versus $+25.7$ on *animal* and $+8.9$ on average over the four hierarchy and ordering benchmarks (App. C). The validation poly-gap on $\hat{A}$ is negative on all eight.
For trophic-level prediction, DIR-POLY reaches 37.4 and 30.3 accuracy on the dry- and wet-season Florida Bay food webs, compared with 51.3 and 36.9 for SYM-POLY. Food webs are dense and cyclic, so the acyclicity behind Theorem 3 fails, and the poly-gap is negative on both ($-18/-7$). Figure 6 (App. C) shows the gap on all 15 datasets, the pattern the poly-gap score of §5 tracks.

8 LIMITATIONS

(i) *Structure only.* All our main experiments use random features. They isolate structural information and do not establish an advantage on attributed graphs. When features already determine the label, the directed advantage shrinks toward zero (§7.4; dataset features in App. B). (ii) *Direction-determined labels.* On seven of eight control graphs DIR-POLY costs accuracy (mean -5.7), and on the two cyclic food webs it falls below SYM-POLY. Trophic-level evaluation covers two food webs. (iii) *Theory and deployment.* Theorems 2–3 hold for the strict walk U and the constant input, and Theorem 4 covers the deployed P with a specified root marker on complete binary trees. The link to a trained DIR-POLY on random features is empirical (§4). (iv) *Semantic labels.* Comparisons with external directed baselines on metadata-derived labels use one dataset, cit-HepTh, where Dir-GNN and MagNet are well ahead of the polynomial filters (App. C).

9 CONCLUSION

We characterized when direction helps a polynomial graph filter. For the deployed self-looped walk and a shared root-marking input, we showed that, to approximate depth to uniform error $1/4$ on a complete binary tree, a monomial coefficient norm linear in the height suffices for the directed walk, while the undirected walk requires an exponential norm regardless of polynomial degree. From a symmetry-invariant input, symmetric filters cannot represent the part of a label that an automorphism changes, whereas a polynomial in the strict directed walk represents depth exactly on rooted hierarchies. Our supervised poly-gap probe tracks the directed-versus-symmetric gain at $r = 0.89$ across 15 directed datasets, and its sign selects the operator about as well as training both models. The observed gains are concentrated in hierarchy and ordering labels. Symmetric propagation remains stronger on most control graphs, and the selection rule does not improve on the symmetric default on the control graphs with dataset features. Direction helps when labels depend on orientation and are not already captured by node features.

AI USE STATEMENT

We used generative AI tools to assist with manuscript phrasing, shortening and polishing, grammar and consistency checks, and coding. The authors developed the original research ideas, reviewed

the manuscript and code, and analyzed and interpreted the results. The authors take responsibility for the final manuscript and accompanying materials.

Reproducibility statement

Proofs of all theorems are in Appendix A, together with numerical checks of each statement. Training details, hyperparameters, dataset construction and splits are in Appendix C. Trained-model results come from result files written by seeded scripts that save per-seed values, and the theorem checks and linear programs are deterministic double-precision computations. The WordNet subtrees are fixed by running with a fixed hash seed, and the DSBM graphs are regenerated per seed. The code, the result files, and a table mapping each table and figure to its script and result file are available at `https://anonymous.4open.science/r/repo-2D70/`, and will be released publicly on acceptance.

References

Ralph Abboud, İsmail İlkan Ceylan, Martin Grohe, and Thomas Lukasiewicz. The surprising power of graph neural networks with random node initialization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations (ICLR)*, 2021.

Dominique Beaini, Saro Passaro, Vincent Létourneau, William L. Hamilton, Gabriele Corso, and Pietro Liò. Directional graph networks. In *International Conference on Machine Learning (ICML)*, 2021.

Deyu Bo, Chuan Shi, Lele Wang, and Renjie Liao. Specformer: Spectral graph neural networks meet transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. ICML Graph Representation Learning and Beyond (GRL+) Workshop; arXiv:2006.13318, 2020.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations (ICLR)*, 2021.

Jeongwhan Choi, Hyowon Wi, Jayoung Kim, Yehjin Shin, Kookjin Lee, Nathaniel Trask, and Noseong Park. Graph convolutions enrich the self-attention in transformers! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9 (1):1–19, 2005.

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

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Liò, and Michael M. Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning (ICML)*, 2023.

Michaël Fanuel, Carlos M. Alaíz, Ángela Fernández, and Johan A. K. Suykens. Magnetic eigenmaps for the visualization of directed networks. *Applied and Computational Harmonic Analysis*, 44(1): 189–199, 2018.

Satoshi Furutani, Toshiki Shibahara, Mitsuaki Akiyama, Kunio Hato, and Masaki Aida. Graph signal processing for directed graphs based on the Hermitian Laplacian. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, volume 11906 of *Lecture Notes in Computer Science*, pp. 447–463, 2020.

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

Simon Geisler, Yujia Li, Daniel Mankowitz, Ali Taylan Cemgil, Stephan Günnemann, and Cosmin Paduraru. Transformers meet directed graphs. In *International Conference on Machine Learning (ICML)*, 2023.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, 2017.

Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. ContraNorm: A contrastive learning perspective on oversmoothing and beyond. In *International Conference on Learning Representations (ICLR)*, 2023.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. BernNet: Learning arbitrary graph spectral filters via Bernstein approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mingguo He, Zhewei Wei, and Ji-Rong Wen. Convolutional neural networks on graphs with Chebyshev approximation, revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Yixuan He, Michael Perlmutter, Gesine Reinert, and Mihai Cucuringu. MSGNN: A spectral graph neural network based on a novel magnetic signed Laplacian. In *Learning on Graphs Conference (LoG)*, 2022b.

Yu He, Ishani Karmarkar, and Ellen Vitercik. Overcoming information bottlenecks in directed graph neural networks through rewiring. In *Learning on Graphs Conference (LoG), Extended Abstract*, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

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

Christian Koke and Daniel Cremers. HoloNets: Spectral convolutions do extend to directed graphs. In *International Conference on Learning Representations (ICLR)*, 2024.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)*, 2005.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20887–20902, 2021.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations (ICLR)*, 2019.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Federico Monti, Karl Otness, and Michael M. Bronstein. MotifNet: A motif-based graph convolutional network for directed graphs. In *IEEE Data Science Workshop (DSW)*, 2018.

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

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

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 (ICLR)*, 2020.

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? In *International Conference on Learning Representations (ICLR)*, 2023.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M. Bronstein. Edge directionality improves learning on heterophilic graphs. In *Learning on Graphs Conference (LoG 2023)*, volume 231 of *Proceedings of Machine Learning Research*, pp. 25:1–25:27, 2024.

T. Konstantin Rusch, Benjamin P. Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael M. Bronstein. Graph-coupled oscillator networks. In *International Conference on Machine Learning (ICML)*, 2022.

Aliaksei Sandryhaila and José M. F. Moura. Discrete signal processing on graphs. *IEEE Transactions on Signal Processing*, 61(7):1644–1656, 2013.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Random features strengthen graph neural networks. In *SIAM International Conference on Data Mining (SDM)*, 2021.

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

Zekun Tong, Yuxuan Liang, Changsheng Sun, Xinke Li, David S. Rosenblum, and Andrew Lim. Digraph inception convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Xiyuan Wang and Muhan Zhang. How powerful are spectral graph neural networks. In *International Conference on Machine Learning (ICML)*, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, and Matthew Hirn. MagNet: A neural network for directed graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Lingxiao Zhao and Leman Akoglu. PairNorm: Tackling oversmoothing in GNNs. In *International Conference on Learning Representations (ICLR)*, 2020.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A PROOFS AND NUMERICAL VERIFICATION

Primitivity (existence of π). For any $\alpha < 1$ the teleported walk A_α is strictly positive, hence primitive and irreducible, so by Perron–Frobenius it has a *unique, strictly positive* stationary measure $\pi > 0$ with $\pi A_\alpha = \pi$ and $\sum_i \pi_i = 1$, used throughout for the $\ell^2(\pi)$ construction.

Theorem 1 (orientation impossibility). *(A, deterministic.)* By (7), $R\,\hat{A}^k_{\mathrm{sym}} x = \hat{A}^k_{\mathrm{sym}}(Rx) = \hat{A}^k_{\mathrm{sym}} x$ for every k, so $\hat{y} = \sum_k c_k \hat{A}^k_{\mathrm{sym}} x \in \mathrm{Fix}(R)$. A node-wise read-out is itself σ-equivariant and maps $\mathrm{Fix}(R)$ into $\mathrm{Fix}(R)$, so $\hat{y} \in \mathrm{Fix}(R)$. Because $y_+ = \frac{1}{2}(I+R)y \in \mathrm{Fix}(R)$ and $y_- = \frac{1}{2}(I-R)y \perp \mathrm{Fix}(R)$, the best approximation in $\mathrm{Fix}(R)$ leaves residual $\|y_-\|^2$. *(B, in expectation.)* Let $f(\hat{A}_{\mathrm{sym}}, X)$ be a node-wise predictor; node-wise equivariance gives $f(\hat{A}_{\mathrm{sym}}, RX) = R\,f(\hat{A}_{\mathrm{sym}}, X)$. Because $RX \overset{d}{=} X$, $\hat{y} = f(\hat{A}_{\mathrm{sym}}, X)$ and $R\hat{y} = f(\hat{A}_{\mathrm{sym}}, RX)$ are identically distributed, hence $\mathbb{E}\|\hat{y} - y\|^2 = \mathbb{E}\|R\hat{y} - y\|^2 = \mathbb{E}\|\hat{y} - R^{-1}y\|^2$. The parallelogram inequality $\|a - y\|^2 + \|a - R^{-1}y\|^2 \geq \frac{1}{2}\|y - R^{-1}y\|^2$ at $a = \hat{y}$, averaged, gives $\mathbb{E}\|\hat{y} - y\|^2 \geq \frac{1}{4}\|y - R^{-1}y\|^2 = \|y_-\|^2$.

Theorem 2 (exact depth expressibility). With the zero root row, $U\mathbf{1} = \mathbf{1} - e_{\mathrm{root}} =: b$ and $(U^k x)(v)$ reads x at the k-th ancestor of v, or is 0 if v has none. Hence $(U^k b)(v) = 1$ exactly when $k < d(v)$, so $\sum_{k=0}^{H-1}(U^k b)(v) = d(v)$ and $U^H b = U^{H+1}\mathbf{1} = 0$. Substituting $b = U\mathbf{1}$ gives $d = \sum_{k=1}^{H} U^k \mathbf{1}$. (Equivalently $(I-U)d = b$, and $I - U$ is invertible because the strict U is nilpotent; with a root self-loop $I - U$ would be singular, which is why the strict convention is used.)

Theorem 3 (exact potential expressibility on a DAG). *(i) Nilpotency.* A nonzero entry $(U^k)_{vu}$ corresponds to a directed walk $v \to \cdots \to u$ of exactly k edges; since G is acyclic every directed walk is a simple path, hence has at most L edges. No walk of length $L{+}1$ exists, so $U^{L+1} = 0$; a longest path realizes a strictly positive entry of U^L (a product of positive $1/d^+$ factors with no cancellation), so the nilpotency index is exactly $L{+}1$. All eigenvalues of a nilpotent matrix are 0, so $1 \notin \mathrm{spec}(U)$ and $I - U$ is invertible; the identity $(I-U)\sum_{k=0}^{L} U^k = I - U^{L+1} = I$ gives the finite Neumann inverse. *(ii) The potential.* By (i) the system $(I-U)t = b$ has the unique solution $t = (I-U)^{-1}b = \sum_{k=0}^{L} U^k b$. Since $b = U\mathbf{1}$ and $U^{L+1} = 0$, the last term vanishes and $t = \sum_{k=1}^{L} U^k \mathbf{1} \in \mathcal{K}_L(U, \mathbf{1})$. Component-wise: for $v \in \mathcal{R}$ the row $U_{v\cdot} = 0$ and $b_v = 0$ give $t(v) = 0$; for $v \notin \mathcal{R}$, $b_v = 1$ and $(Ut)(v) = \mathrm{mean}_{u \in \mathrm{pa}(v)}\, t(u)$ give $t(v) = 1 + \mathrm{mean}_{u \in \mathrm{pa}(v)}\, t(u)$, the expected-absorption-time recursion of the walk that steps to a uniformly random parent and absorbs at $\mathcal{R}$. Invertibility uses acyclicity *only*, so (i)–(ii) hold for disconnected, multi-root, and multi-source DAGs; the single-root hypothesis enters solely in the tree reduction. *(iii) Tree reduction.* If every non-root node has one parent and $\mathcal{R} = \{\mathrm{root}\}$, each non-root row of U is a single 1 at the parent, so U is the upward walk of Theorem 2 and $b = \mathbf{1} - e_{\mathrm{root}}$. The recursion becomes $t(v) = 1 + t(\mathrm{pa}\, v)$ with $t(\mathrm{root}) = 0$, i.e. integer depth $t = d$. Here $L = H$, so $t = \sum_{k=1}^{H} U^k \mathbf{1}$, Theorem 2 verbatim, with no constant offset. *Label identification.* Off $\mathcal{R}$, $t(v) = 1 + \mathrm{mean}_{u \in \mathrm{pa}(v)}\, t(u)$ is an affine averaging recursion, so t is the unique discrete-harmonic-with-source potential, linear in the seed and exactly degree-L in U. Longest-path height $h(v) = 1 + \max_{u \in \mathrm{pa}(v)} h(u)$ and shortest-path depth $m(v) = 1 + \min_{u \in \mathrm{pa}(v)} m(u)$ are MAX/MIN recursions and do not average (on the asymmetric diamond $(I-U)h$ and $(I-U)m$ differ from b at the join node), and Theorem 3 does not cover them. Whether another polynomial represents them is a separate question: because U is nilpotent, every polynomial in U applied to the input lies in the finite space $\mathcal{K}_L(U, \mathbf{1})$, so it can be checked per graph by a degree-L fit (below). One has $m(v) = t(v) = h(v)$ when all rootward paths from v have equal length, and $m(v) < t(v) < h(v)$ otherwise (asymmetric diamond: $m = 2 < t = 2.5 < h = 3$); in every case $h(v) \leq L$. *Deployed walk.* The deployed $P = D^{-1}(I+A)$ adds a per-node self-loop, so $P\mathbf{1} = \mathbf{1}$: from the constant input P generates nothing, and it needs a non-constant input (in the experiments, the feature map of random features). Non-nilpotence does not by itself preclude exactness. On a single-parent tree $P = \frac{1}{2}(I + U_{\mathrm{a}})$, where U_{a} is the walk with an absorbing root, and $U^k_{\mathrm{a}} b = U^k b$ for the root-marking seed $b = \mathbf{1} - e_{\mathrm{root}}$, so $d = \sum_{k=0}^{H-1}(2P - I)^k b$ is an exact degree-$(H{-}1)$ polynomial in P applied to b. On multi-parent DAGs the self-loop weight depends on the out-degree, P is not a polynomial in U, and the fits of P reported below are empirical.

Numerical verification of Theorem 3. On five DAGs (a rooted binary tree, an asymmetric diamond, two random non-graded multi-parent DAGs of 90 and 112 nodes, and a WordNet-like hier-

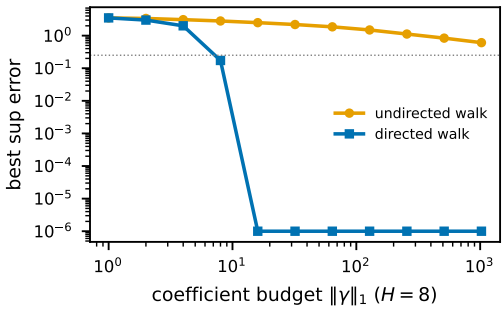


Figure 3: Best achievable sup error under a coefficient budget on the complete binary tree of height 8 (numerical linear programs on the exact level-chain reduction, degree up to $8H$); the dotted line marks error $\frac{1}{4}$.

archy) the strict averaging walk satisfies $\|U^{L+1}\|_F = 0$ and $\|(I-U)^{-1} - \sum_{k=0}^{L} U^k\|_F < 10^{-16}$, the potential $t = \sum_{k=0}^{L} U^k b$ matches $(I-U)^{-1}b$ to machine precision with degree-L Krylov fit $R^2 = 1$ (and lower degree underfitting), and the tree reduction matches integer depth exactly ($\max|\text{formula} - d| = 0$, $\|U^H(\mathbf{1} - e_{\text{root}})\| = 0$). The deployed self-looped P has $\rho(P) = 1$ and $\|P^{L+1}\|_F > 0$ (not nilpotent), with structural-seed Krylov fit $R^2 = 0.96$–1.0; longest-path height is fit at only $R^2 = 0.88$–0.90 by the degree-L Krylov space on the two non-graded DAGs, so on those graphs it is not a polynomial in U applied to the seed.

Theorem 4 (coefficient-budget separation). *(i)* Since the root is absorbing and b is the indicator of the non-root nodes, $(P^k b)(v) = \Pr_v[X_k \neq \text{root}] = \Pr[T_v > k]$, where X_k is the position after k steps of the lazy upward walk started at v and T_v is its absorption time. The walk moves to the parent with probability $\frac{1}{2}$ at every step, so T_v is a sum of $d(v)$ independent Geometric($\frac{1}{2}$) variables and $\sum_{k\geq 0}\Pr[T_v > k] = \mathbb{E}T_v = 2d(v)$; hence $\frac{1}{2}\sum_{k\geq 0} P^k b = d$ and the truncation error at v is $\frac{1}{2}\sum_{k\geq K}\Pr[T_v > k]$. For $k \geq K \geq 4H \geq 4d(v)$, $\Pr[T_v > k] = \Pr[\text{Bin}(k, \frac{1}{2}) \leq d(v) - 1] \leq \Pr[\text{Bin}(k, \frac{1}{2}) \leq k/4] \leq e^{-k/8}$ by Hoeffding's inequality, and $\frac{1}{2}\sum_{k\geq K} e^{-k/8} = \frac{1}{2}e^{-K/8}/(1 - e^{-1/8}) \leq 5e^{-K/8}$. All K coefficients equal $\frac{1}{2}$, so $\|\gamma\|_1 = K/2$, and $K = \max\{4H, 8\log(5/\varepsilon)\}$ gives error at most ε. *(ii)* P_{sym} is row-stochastic, so $q(P_{\text{sym}})b = q(1)\mathbf{1} - g$ with $g = \sum_k \gamma_k P_{\text{sym}}^k e_{\text{root}}$, i.e. $g(v) = \sum_k \gamma_k P_{\text{sym}}^k(v, \text{root})$. The walk is reversible with respect to $\pi(v) \propto \deg(v) + 1$ (the self-loop counts), so $P_{\text{sym}}^k(v, \text{root}) = \frac{\pi(\text{root})}{\pi(v)} P_{\text{sym}}^k(\text{root}, v)$. The automorphisms of the complete binary tree that fix the root act transitively on each level, so $P_{\text{sym}}^k(\text{root}, \cdot)$ is constant on the 2^j nodes of depth j and therefore at most 2^{-j} there. With $\deg + 1$ equal to 3 at the root, 4 at internal nodes and 2 at leaves, $|g(v)| \leq \frac{3}{2}\|\gamma\|_1 2^{-H}$ at a leaf and $|g(u)| \leq \frac{3}{4}\|\gamma\|_1 2^{-(H-1)}$ at its parent u (an internal non-root node because $H \geq 2$). A uniform error of at most $\frac{1}{4}$ forces the outputs at v and u, whose depths differ by one, to differ by at least $\frac{1}{2}$; the constants cancel, so $\frac{1}{2} \leq |g(v) - g(u)| \leq 3\|\gamma\|_1 2^{-H}$, i.e. $\|\gamma\|_1 \geq 2^H/6$.

Numerical check of Theorem 4. Both b and d are constant on levels and both walks preserve level functions, so the problem reduces exactly to $(H+1)$-state chains. For $H = 12$ and $K = 4H$ the directed filter has $\|\gamma\|_1 = 24$ and sup error 1.5×10^{-4}. For the undirected walk we solve numerically (floating-point LP solver, on the exact reduction) the program $\min \|\gamma\|_1$ subject to sup error at most $\frac{1}{4}$ (free intercept, degree up to $8H$): the optimum is 37.8, 6635 and 8.93×10^5 at $H = 4, 8, 12$, above the bound $2^H/6$ at every height from 2 to 12 and growing by a factor of about 3.3 per level, while the same program for the directed walk gives 3.54, 7.71 and 11.91 (Figures 1 and 3). Raising the degree cap from $4H$ to $8H$ leaves these optima unchanged.

Numerical verification of the same-input separation. On a directed path ($n = 60$, σ the reflection) with the constant input, every symmetric operator we use commutes with R ($\|[R, \cdot]\|_{\max} \leq 10^{-14}$ for P_{sym}, $\hat{A}_{\text{sym}}$ and the symmetric-normalized adjacency), while $\|[R, U]\|_{\max} = 1$. For each of the three symmetric operators and every degree $K \in \{2, 4, 8, 16, 32, 59\}$, the least-squares residual of the depth label on $\text{span}\{\mathbf{1}, M\mathbf{1}, \ldots, M^K\mathbf{1}\}$ equals the lower bound $\|y_-\|^2 = 17995$ exactly, so the bound of Theorem 1(A) is attained and no degree helps. From the same input, $\sum_{k=1}^{H} U^k\mathbf{1}$

Table 3: Predictor robustness (15 directed datasets; probe scored on validation nodes; target: trained DIR-POLY minus SYM-POLY test gap). "rule" is the mean test accuracy of the per-split selection rule; "probe" is the test accuracy of the selected probe itself.

features	score	r [95% CI]	ρ	sign [CP 95%], p	LOFO r	rule	probe
random	$\hat{A}$	0.89 [0.69, 0.97]	0.86	14/15 [0.68, 1.00], 9.8×10^{-4}	0.84–0.94	65.6	58.8
random	raw walk	0.96 [0.81, 0.99]	0.90	14/15 [0.68, 1.00], 9.8×10^{-4}	0.93–0.97	65.4	61.3
dataset	$\hat{A}$	0.81 [0.60, 0.95]	0.82	12/15 [0.52, 0.96], 3.5×10^{-2}	0.77–0.87	77.1	67.8
dataset	raw walk	0.91 [0.60, 0.98]	0.83	11/15 [0.45, 0.92], 1.2×10^{-1}	0.76–0.96	76.4	69.7

matches depth with zero error and $U^{H+1}\mathbf{1} = 0$. On a random rooted tree ($n = 200$, height 11) the closed form is again exact, $P = \frac{1}{2}(I + U_{\mathrm{a}})$ holds, and $\sum_{k<H}(2P - I)^k(\mathbf{1} - e_{\mathrm{root}})$ reproduces depth exactly, while $P\mathbf{1} = \mathbf{1}$.

A.1 What reversibilization removes

A standard alternative to filtering the raw walk is to first make it self-adjoint in $\ell^2(\pi)$, replacing A_α by the reversibilized $\hat{A} = A_\alpha - \hat{S}$, which restores a real spectral interpretation and an orthogonal eigenbasis. Two facts are exact. First, reversibilization removes the skew part $\hat{S}$ by construction: it keeps the symmetric part of the stationary edge flows ΠA_α and removes their antisymmetric part. Second, $\hat{A}$ is row-stochastic, $\hat{A}\mathbf{1} = \frac{1}{2}(A_\alpha\mathbf{1} + \Pi^{-1}A_\alpha^\top\pi^\top) = \mathbf{1}$, so every polynomial in $\hat{A}$ maps the constant input to a constant and Theorems 2–3 do not transfer to it, whether or not π is uniform. How much label-relevant information the skew part carries is an empirical question that no result here settles. With random features, the reversibilized filter trails DIR-POLY by 15.2, 3.4 and 2.3 points on the three WordNet tasks and is 0.3 points above it on DSBM (Table 2). In this paper $\hat{A}$ serves as the REV baseline and inside the poly-gap predictor (§5). The deployed filter uses P.

B Predictor Robustness and the Seed-to-Input Bridge

Poly-gap predictor robustness. All predictor statistics use the probe scored on validation nodes and the trained DIR-POLY minus SYM-POLY test gap (5 seeded splits per dataset) on the 15 directed datasets. We report two feature settings. *Random features*: 16-dimensional Gaussian features on every dataset. *Dataset features*: the dataset's own node attributes on Cora_ml, Citeseer, Chameleon, Squirrel, Cornell, Texas and Wisconsin; Telegram's single attribute together with the logarithms of in- and out-degree; and random features on DSBM, WordNet and the food webs, which have no attributes. Table 3 reports, for both feature settings and both scores, Pearson r with a nonparametric bootstrap 95% interval over datasets (10^4 resamples), Spearman ρ (average ranks), sign agreement with its Clopper–Pearson interval and a two-sided exact binomial test, the range of r under leave-one-*family*-out, the mean test accuracy of the per-split selection rule, and the selected probe's own test accuracy (the trained filter adds 4–9 points over the probe that selected it). Replacing the matched teleportation $\alpha = 0.9$ by 0.999 on both sides, or by a mismatched pair (0.9 directed, 0.999 symmetric), changes r by at most 0.02 (random features: 0.89 and 0.89). The PyTorch Geometric Airports graphs are not included: they are undirected edge lists stored once per edge, so any direction is a storage artifact (reciprocity 0). cit-HepTh was assembled as the benchmark for comparisons with external baselines and was not on the predictor study's dataset list, so we report it separately and do not treat it as held-out validation: its validation poly-gap is +5.8 (score on $\hat{A}$) and +11.9 (raw walk) and DIR-POLY improves over SYM-POLY by 9.1 points.

Selection rule: cost and uncertainty. Table 4 compares the probe rule with choosing after a short training run and with training both models to completion and choosing on validation. We apply every rule within a split and score it by the chosen model's final test accuracy, and the cost is that of making the choice. The probe and the short-training rules also require training the selected model to completion (about 0.8 s here); full validation selection already includes that training, so the absolute savings are small on these graphs. The paired improvement of the rule (score on $\hat{A}$) over always using SYM-POLY is +2.5 [0.0, 6.3] with random features and +1.9 [−0.7, 5.8] with dataset features (hierarchical bootstrap over datasets and splits, 95%). Over always using DIR-POLY it is +4.3 [0.7, 8.9] and +3.1 [0.9, 5.6]. Restricted to the eight control graphs with dataset features, the rule is below the symmetric default, −1.0 [−2.6, −0.1], where the per-split oracle is only 0.4 points above that default.

Table 4: Selection rules: mean test accuracy of the chosen model over 15 datasets and 5 splits, and mean wall-clock cost of making the choice (seconds per split, one GPU); the probe and short-training rows exclude the final training of the selected model, which full validation selection already includes.

	random features		dataset features	
rule	acc.	cost (s)	acc.	cost (s)
always DIR-POLY	61.2	–	74.0	–
always SYM-POLY	63.1	–	75.2	–
probe rule (score on $\hat{A}$)	65.6	0.18	77.1	0.22
probe rule (raw walk)	65.4	0.18	76.4	0.22
short training, 10 epochs	64.3	0.03	76.3	0.03
short training, 50 epochs	64.3	0.16	76.0	0.16
short training, 200 epochs	65.2	0.62	76.8	0.64
train both fully, choose on validation	65.7	1.56	77.2	1.57
oracle (per split)	66.7	–	78.2	–

Table 5: Within-group association between the validation poly-gap and the measured DIR-POLY minus SYM-POLY test gap (Pearson r, Spearman ρ with average ranks). Groups are small; the two food webs are omitted.

		random features				dataset features	
		score on $\hat{A}$		raw walk		score on $\hat{A}$	
group	n	r	ρ	r	ρ	r	ρ
hierarchy and ordering	5	0.78	0.10	0.96	0.30	0.68	0.00
controls	8	0.94	0.55	0.96	0.76	0.22	0.31
controls without Telegram	7	0.46	0.32	0.65	0.64	0.32	0.43

Within-group association and a family rule. The 15 datasets fall into a few families, so an overall correlation can reflect differences between families. Table 5 reports the association within groups. Among the five hierarchy and ordering datasets the score on $\hat{A}$ has $r = 0.78$ but $\rho = 0.10$. Among the eight controls $r = 0.94$ falls to 0.46 without Telegram, and with dataset features it is 0.22. As a post-hoc diagnostic we also score the rule "directed for WordNet and DSBM, symmetric elsewhere" in the same way as the probe rule. It reaches a mean test accuracy of 65.6 with random features (probe rule 65.6) and 77.7 with dataset features (probe rule 77.1). The rule agrees with the sign of the measured gap on 14 of 15 datasets. With random features it makes the same dataset-level decision as the probe on all 15 (the same decision on 71 of 75 splits). The selection results therefore do not show that the probe adds information beyond recognizing the dataset family. Table 6 lists all 15 datasets.

Operator selection on a standard large graph. We also run the selection rule on arxiv-year (Lim et al., 2021): the ogbn-arxiv citation graph (Hu et al., 2020) (169,343 nodes, 1,166,243 edges) with its 128-dimensional features, five publication-year classes, and the five published $50/25/25$ splits. The operators are applied as sparse products with no dense $n \times n$ matrix, and the stationary-measure power iteration converges to an ℓ_1 residual below 10^{-15}. The validation poly-gap on $\hat{A}$ is positive on four of five splits (mean $+2.0$) and the raw-walk score on all five (mean $+2.4$). Trained DIR-POLY reaches $42.0_{\pm 0.3}$ and SYM-POLY $41.5_{\pm 0.7}$, so the two filters differ by 0.6 points here. The rule on $\hat{A}$ gives 42.0, equal to training both models and choosing on validation, and the one split where it chooses SYM-POLY is a tie (41.5 for both filters). The probe plus training the selected model takes 13.0 s per split, including the one-time stationary-distribution computation (0.16 s) amortized over the five splits, against 26.3 s for training both. The raw-walk probe alone, a linear readout on the Krylov features, reaches 43.1, above both trained filters. This experiment tests operator selection at scale with a fixed architecture. Both filters are well below models tuned for this benchmark.

From the structural seed to random features. Theorems 2–3 are stated for the strict walk U and the constant input, whereas DIR-POLY applies the self-looped walk P to random node features. We test this bridge directly: fixing the degree ($K = 10$, the deployed value) and varying the operator and the input, we fit depth with a ridge read-out on the orthonormalized block-Krylov basis in double precision. The read-out is fit on 50% of the nodes, its penalty is selected on a 25% validation part, and we report test R^2 on the remaining 25%, averaged over five splits (and five feature draws). Table 7 reports the result. From the constant input only the strict U expresses depth (0.98–0.99);

Table 6: Predictor study, per dataset: mean test accuracy (%) of SYM-POLY and DIR-POLY, their gap, and the mean validation poly-gap, over 5 seeded splits. The last two columns repeat the gap and the score on $\hat{A}$ with dataset features.

	random features					dataset features	
dataset	SYM-POLY	DIR-POLY	gap	score ($\hat{A}$)	score (raw)	gap	score ($\hat{A}$)
DSBM k=5	95.2	98.6	+3.3	+10.0	+6.5	+1.9	+11.1
DSBM k=8	97.9	98.7	+0.8	+9.6	+7.6	+1.3	+11.2
WordNet animal	73.8	99.5	+25.7	+17.1	+33.5	+25.7	+17.1
WordNet plant	95.2	99.8	+4.6	+3.1	+5.9	+4.6	+3.1
WordNet artifact	96.0	99.5	+3.6	+7.7	+13.3	+3.6	+7.7
Food web (dry)	51.3	37.4	−13.8	−18.4	−17.6	−13.8	−18.4
Food web (wet)	36.9	30.3	−6.7	−7.2	−11.2	−6.7	−7.2
Telegram	80.0	51.2	−28.8	−22.1	−27.2	−8.8	−2.6
Cora_ml	63.9	59.2	−4.7	−6.8	+0.2	−5.7	−4.8
Citeseer	20.2	16.9	−3.3	−2.8	−1.3	−7.3	−8.7
Chameleon	44.7	50.8	+6.1	−2.3	+2.0	−4.6	−2.3
Squirrel	32.6	32.5	−0.1	−3.9	−1.8	−5.1	−2.9
Cornell	45.9	40.0	−5.9	−2.4	−3.4	−0.5	+0.3
Texas	63.2	60.0	−3.2	−4.4	−0.7	−2.2	+3.1
Wisconsin	49.0	43.9	−5.1	−6.0	−6.8	+0.4	−8.0

Table 7: Seed-to-input ablation: test R^2 of the depth label for a validation-selected ridge read-out on the degree-10 block-Krylov basis (double precision; $50/25/25$ split; mean over five splits and, for random features, five draws; the synthetic tree has 600 nodes). It complements Table 10, which uses an unregularized least-squares read-out, a $50/50$ split and single-vector inputs on smaller synthetic graphs. From the constant input only the strict walk U expresses depth. From random features the deployed walk P recovers most of it and the symmetric walk does not.

operator	input	syn. tree	WN animal	WN plant	WN artifact
U (strict)	constant $\mathbf{1}$	0.98	0.98	0.99	0.99
U (strict)	root seed $\mathbf{1} - \mathbf{1}_{\mathcal{R}}$	0.99	0.98	0.99	0.99
U (strict)	random features (d=16)	0.49	0.61	0.94	0.95
P (deployed)	constant $\mathbf{1}$	−0.02	0.00	−0.01	0.00
P (deployed)	root seed $\mathbf{1} - \mathbf{1}_{\mathcal{R}}$	0.99	0.98	1.00	1.00
P (deployed)	random features (d=16)	0.49	0.90	0.86	0.92
sym. walk	constant $\mathbf{1}$	−0.02	0.00	−0.01	0.00
sym. walk	root seed $\mathbf{1} - \mathbf{1}_{\mathcal{R}}$	0.34	0.49	0.47	0.35
sym. walk	random features (d=16)	−0.48	0.19	−0.02	−0.04

P and the symmetric walk fix the constant ($R^2 \approx 0$). From the deployed 16-dimensional random features, P recovers depth at $R^2 = 0.49$–0.92 and U at 0.49–0.95, while the symmetric walk reaches at most 0.19. An unregularized read-out on the random-feature basis is numerically meaningless (the training block is ill-conditioned), which is why the probe is regularized.

B.1 Simpler predictors and the two scores

The score on $\hat{A}$ compares two self-adjoint operators with the same teleportation. The raw-walk variant uses the operators that DIR-POLY and SYM-POLY filter with. Against the trained comparison the raw-walk score has the higher correlation in both feature settings, while the score on $\hat{A}$ selects better with dataset features (Table 3), so neither dominates.

Simpler explanations. We compare the poly-gap with four alternative predictors on the same 15 datasets. The first two are fit on training nodes and scored on validation nodes like the poly-gap. *(i) Static π-alignment* (how much better the label is predicted from $\log \pi$ and the in/out degrees than from the total degree): $r = -0.07$. *(ii) Label propagation* on $\hat{A}$ versus $\hat{A}_{\mathrm{sym}}$: $r = -0.50$, anti-correlated because labels for which direction helps place dissimilar labels on adjacent nodes. *(iii) Edge reciprocity*: $r = 0.03$. *(iv) Cyclicity* (share of edges inside a non-trivial strongly connected component): $r = -0.42$. With dataset-feature targets the four values are -0.05, -0.26, 0.01 and -0.45.

C Training Details and Extended Results

All models share one transductive node-classification setup with random node features, so the only learnable signal is graph structure (results with dataset features are reported separately in App. B). We report test accuracy over 5 seeds. The per-seed split and model initialization are fixed by the seed, and the WordNet subtrees are fixed by running with `PYTHONHASHSEED=0` (NLTK returns hyponyms in hash order); each result file records the graph's node and edge counts and an edge-list hash.

Datasets. WordNet *animal/plant/artifact* are 1500-node hypernym subtrees with a depth-quantile label (4 requested bins; ties in integer depth leave 3 classes for *animal* and 2 for *plant* and *artifact*), 16-dimensional random features, and a $50/20/30$ train/val/test split. DSBM has 5 ordered blocks of 150 nodes (750 total): each ordered within-block pair is sampled with probability 0.04 and inserted in both directions (undirected within-block density 0.078), and each pair from block c to block $c{+}1$ is connected one way with probability 0.04, 16-dimensional features, and a $20/20/60$ split. citHepTh is a 4000-node dense citation core with a 4-class publication-era label and 16-dimensional random features.

Polynomial filters (SYM-POLY, REV, DIR-POLY, BIDIR-POLY). An MLP feature map (hidden width 64, dropout 0.5) is followed by a degree-$K{=}10$ polynomial of the operator and a linear softmax readout. We train with Adam for up to 1000 epochs and early-stop on validation accuracy with patience 200. The learning rate is 0.01 for the MLP and readout and 0.05 for the polynomial coefficients (γ, and δ for BIDIR-POLY), with weight decay 5×10^{-4} (0 on the coefficients). Teleportation $\alpha{=}0.9$ enters only the π/reversibilization construction. Applying it inside DIR-POLY changes accuracy by at most 0.3 points on the four hierarchy and ordering tasks. BIDIR-POLY adds a reverse-walk channel at about $2K$ sparse matrix products.

MagNet. We use the `torch-geometric-signed-directed` implementation with hidden width 64, dropout 0.5, and symmetric normalization. We consider three configurations of Chebyshev order and depth: the standard $(K, \text{layers}) = (2, 2)$, and $(5, 2)$ and $(2, 5)$, whose propagation reach $K \times \text{layers} = 10$ matches the degree-10 polynomial filters. The charge q ranges over $\{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$. We train with Adam (learning rate 0.01, weight decay 5×10^{-4}) with the same budget as the polynomial filters (1000 epochs, patience 200), on CPU, since the complex magnetic-Laplacian kernel is unavailable on GPU. One configuration and charge per dataset is chosen by mean validation accuracy.

Dir-GNN. We use the PyTorch Geometric `DirGNNConv` wrapper around GCN convolutions (separate in- and out-neighbor aggregation, mixing weight 0.5) with hidden width 64, dropout 0.5, and the optimizer and budget of the polynomial filters. We choose the depth once per dataset from $\{2, 4\}$ by mean validation accuracy, and four layers are selected on every dataset.

Poly-gap predictor. The score fits a multinomial logistic-regression readout (Adam, 300 steps) on the degree-$K{=}8$ Krylov basis $[X, MX, \ldots, M^K X]$, with no GNN training, for the directed operator $\hat{A}$ and the symmetric operator $\hat{A}_{\text{sym}}$, both with $\alpha = 0.9$ (Section 5); Krylov features are standardized and the readout uses learning rate 0.02 and weight decay 10^{-4}. The raw-walk variant uses the same readout on P and P_{sym} (after standardization, teleporting the raw walk leaves the probe unchanged).

Reproducibility. Every training run is seeded and the WordNet subtrees are fixed. The reported numbers come from result files written by the scripts, which save per-seed values. The code, the result files and the seeds that fix the splits are at the link given in the reproducibility statement.

C.1 Method details: bidirectional extension and algorithm

Horner evaluation. Equation (8) is evaluated as

$$\begin{aligned} Y_K &= \gamma_K H, \\ Y_k &= \gamma_k H + P\, Y_{k+1} \quad (k = K{-}1, \ldots, 0), \\ \text{DIR-POLY}(X) &= Y_0, \end{aligned} \tag{9}$$

which costs exactly K sparse matrix products against P and never materializes a power P^k.

DIR-POLY forward pass
Input: features X; row-stochastic walk P; coeffs $\gamma_{0:K}$
1. $Z \leftarrow \mathrm{MLP}(X)$
2. $Y \leftarrow \gamma_K\, Z$
3. **for** $k = K-1$ down to 0 **do**
   $Y \leftarrow \gamma_k\, Z + P\, Y$ // one sparse matrix product
4. **end for**
Return: node-wise readout of Y

Figure 4: DIR-POLY is evaluated by Horner's rule in exactly K sparse products against P. No power P^k is materialized and no $n \times n$ operator is eigendecomposed.

Bidirectional extension. The polynomial in Equation (8) integrates only along the forward walk P, which moves probability toward sinks. This is sufficient when the label is a function of one walk direction (depth and topological order on a nearly acyclic hierarchy). Some labels, however, depend on *both* edge directions. The publication era of a citation node, for instance, is informed both by which older papers it cites (out-edges) and by which newer papers cite it (in-edges). For such targets we add a channel that propagates with the transpose $P^\top$ (in-neighbor aggregation). $P^\top$ performs unnormalized in-neighbor aggregation; it differs from the row-normalized walk on the reversed graph.

$$\text{BIDIR-POLY}(X) \;=\; \sum_{k=0}^{K} \big(\gamma_k\, P^k + \delta_k\, (P^\top)^k\big)\, Z, \tag{10}$$

with separate learnable coefficients γ (forward, out-citation walk) and δ (reverse, in-citation walk). Each channel is evaluated by the same Horner recurrence as Equation (9), so BIDIR-POLY stays real and eigendecomposition-free at about $2K$ sparse matrix products, unlike the complex magnetic Laplacian of MagNet (Zhang et al., 2021). The two forward/reverse aggregations make BIDIR-POLY a polynomial analogue of Dir-GNN (Rossi et al., 2024), which separates in- and out-neighbor aggregation. Forward-only DIR-POLY is the simpler filter, and on nearly acyclic hierarchies it already matches BIDIR-POLY. The reverse channel is useful only when the label depends on both directions.

C.2 SYMMETRIC BASELINE CONSTRUCTION

The symmetric operators. Both symmetric operators are built on the symmetrized edge set $E \cup E^\top$. SYM-POLY filters P_{sym}, the self-looped row-normalized random walk of the undirected graph: the same construction as the directed P, with no teleportation. It is reversible with respect to the degree distribution, hence self-adjoint in its own $\ell^2(\pi)$. The predictor's $\hat{A}_{\mathrm{sym}}$ applies the teleport-and-reversibilize pipeline to the undirected graph with the same $\alpha = 0.9$ as the directed side. Uniform teleportation can break detailed balance even on an undirected graph (the skew part is then small but nonzero for $\alpha < 1$), which is why we keep the reversibilization step there. At $\alpha = 1$ the two operators coincide. DIR-POLY and SYM-POLY therefore share the architecture, the parameter count and the operator construction, and differ only in whether the edge set is symmetrized. The teleportation ablation covers the directed side.

C.3 SEMANTIC-LABEL BENCHMARK: CITATION ERA

WordNet depth is a structural label. To show the effect on a genuinely semantic target we turn to cit-HepTh, where the label is the publication era (when a physics paper was written), recovered purely from citation direction (citations point backward in time) and never from node content (features are random). The 4000-node citation core is nearly acyclic (reciprocity 0.006; 99.9% of the edges that cross era bins point backward in time). Theorem 3 does not apply to it: the core is not exactly acyclic, and publication era is not the absorption-time potential. The benchmark tests whether a directed filter helps on a semantic label tied to edge direction. Table 8 reports the comparison on this semantic-label benchmark.

The forward-only DIR-POLY reaches 61.3, 9 points above the symmetric baseline but well below the external directed baselines: the forward walk integrates only toward sinks, while publication era depends on both edge directions (which older papers a node cites and which newer papers cite it). The bidirectional filter BIDIR-POLY adds a separate reverse-walk channel, $\text{BIDIR-POLY}(X) = \sum_{k=0}^{K} \gamma_k P^k Z + \sum_{k=0}^{K} \delta_k (P^\top)^k Z$, with learnable coefficients γ for the out-citation walk and δ for the in-citation walk. It costs about $2K$ sparse matmuls by Horner. On this benchmark BIDIR-POLY achieves 65.1 ± 1.5, compared with 66.9 ± 1.6 for MagNet in its standard two-layer form. At

Table 8: Semantic-label benchmark: cit-HepTh publication-era prediction on the 4000-node citation core (4 classes, random features). The core is nearly but not exactly acyclic. Mean ± sample standard deviation of test accuracy (%) over 5 seeded runs; MagNet's configuration and charge q and Dir-GNN's depth are selected once, by mean validation accuracy, and both use the polynomial filters' training budget. MagNet in its standard two-layer form is shown for reference.

method	accuracy
SYM-POLY	52.2 ± 1.2
DiGCN	57.7 ± 1.8
REV	58.3 ± 4.1
DIR-POLY	61.3 ± 1.6
BIDIR-POLY	65.1 ± 1.5
MAGNET (K=2, 2 layers)	66.9 ± 1.6
MAGNET (selected: K=2, 5 layers)	76.7 ± 1.4
Dir-GNN (4 layers)	$\mathbf{77.0 \pm 1.5}$

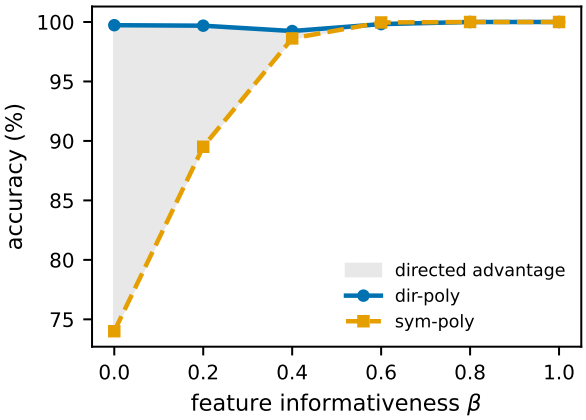


Figure 5: Feature-informativeness sweep on WordNet *animal* (5 seeded runs). As β rises, SYM-POLY approaches 100 while DIR-POLY stays near 100 throughout, and the directed advantage (shaded) falls to zero.

matched propagation reach MagNet reaches 76.7 ± 1.4 (order 2, five layers, q=0.1, selected on validation), against 77.0 ± 1.5 for a four-layer Dir-GNN. Both are about 12 points above BIDIR-POLY (means and sample standard deviations over five runs). BIDIR-POLY uses the same configuration as DIR-POLY (K=10, raw walk P); in a sensitivity sweep over $K \in \{10, 16, 24\}$ and teleportation $\alpha \in \{1, 0.95, 0.9\}$ this configuration also has the highest mean validation accuracy, and test means across the nine configurations range from 64.0 to 65.8. Architecturally BIDIR-POLY resembles a polynomial form of Dir-GNN (Rossi et al., 2024), which aggregates in- and out-neighbors separately. The symmetric impossibility theorem (§4) motivates moving to directed operators. Both DIR-POLY and BIDIR-POLY lie outside that symmetric obstruction.

C.4 OPERATOR VS. CAPACITY

DIR-POLY and SYM-POLY have identical parameter counts and differ only in the propagation operator. The gap between them reaches +25.7 points (*animal*, Table 2).

C.5 ORIENTATION AND NODE FEATURES ARE COMPLEMENTARY

We sweep a feature informativeness β on WordNet *animal*, mixing random features (β=0) with a label-informative encoding (β=1). As feature informativeness increases, the DIR-POLY minus SYM-POLY gap falls from +25.7 to 0 (Figure 5). DIR-POLY remains at 99.2 or above, while SYM-POLY rises from 74.0 to 100.

On **ogbn-arxiv** (a real 169,343-node citation graph with real 128-dimensional features and a publication-year label), DIR-POLY improves over SYM-POLY by +3.2 points with random features but only +0.3 with the real features (BIDIR-POLY +0.6), because the real features already encode most of the year signal. This experiment uses 500 epochs with patience 100 and a 50/25/25 split.

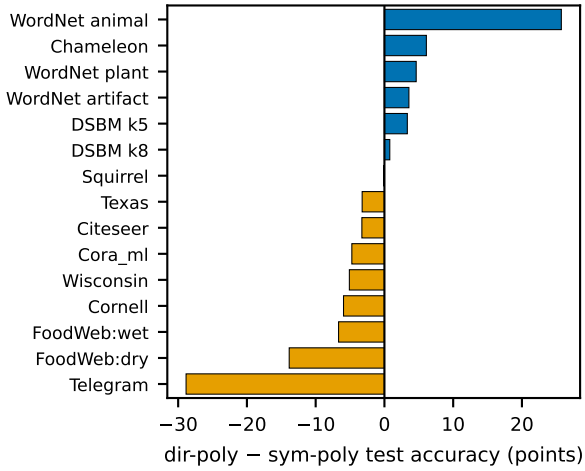


Figure 6: Per-dataset test-accuracy gap of DIR-POLY over SYM-POLY (points) on the 15 datasets of the predictor study, with random features and 5 seeded runs. Blue: the directed filter is better; orange: the symmetric one.

C.6 Baselines and control results

MagNet is swept over three configurations and six charges with the polynomial filters' training budget and reported at its validation-selected setting (Table 9). Depth matters for MagNet on these tasks: the five-layer form is selected everywhere and raises accuracy from 79.2 to 93.9 on *animal* and from 66.9 to 76.7 on cit-HepTh. Against it DIR-POLY is higher on the three WordNet tasks ($+5.6$, $+0.9$, $+1.0$) and lower on DSBM (-0.8), and it is within one point of a four-layer Dir-GNN. DiGCN (Tong et al., 2020) stays at the majority-class rate on the three WordNet tasks ($32.9/51.7/54.0$ against majority rates of $34.5/50.1/57.4$) and reaches 86.3 on DSBM, so it is not informative as a baseline on the hierarchies.

On the eight control graphs with random features, DIR-POLY minus SYM-POLY averages -5.7 points (Telegram -28.9, Cora_ml -4.7, Citeseer -3.3, Chameleon $+6.1$, Squirrel -0.1, Cornell -6.0, Texas -3.2, Wisconsin -5.1; $7/8$ negative). With dataset features the mean is -4.2 (Telegram -8.8, Cora_ml -5.7, Citeseer -7.3, Chameleon -4.6, Squirrel -5.1, Cornell -0.5, Texas -2.2, Wisconsin $+0.4$; $7/8$ negative).

C.7 Future work

One next step is extending the BIDIR-POLY comparison against MagNet and MSGNN beyond cit-HepTh to more semantic-label directed graphs, to test whether the cit-HepTh result generalizes. A second is extending the truncation argument from strictly acyclic to "nearly acyclic" regimes so the predictor degrades gracefully on partially cyclic graphs. More broadly, the poly-gap predictor suggests a model-selection procedure for directed graph learning, namely measuring orientation-expressibility before choosing the operator; testing whether it generalizes beyond the polynomial filters studied here is left to future work.

C.8 Broader related work

Our impossibility argument is at root equivariance: a symmetric filter commutes with automorphisms, so it cannot distinguish what the symmetry identifies. Where invariant and equivariant networks (Maron et al., 2019) use symmetries to *design* architectures, we use one to prove a limitation. GFSA (Choi et al., 2024) reads a transformer attention map as a learnable graph filter but carries no directed-orientation theory, so our separation and predictor are orthogonal to it.

Earlier directed designs include the motif-based MotifNet (Monti et al., 2018) and the directional aggregations of Beaini et al. (2021); more recent work extends spectral convolution to digraphs (Koke & Cremers, 2024) and brings directed structure into graph transformers (Geisler et al., 2023). The Hermitian and magnetic operators these methods use rest on spectral analyses of directed graphs (Chung, 2005; Fanuel et al., 2018; Furutani et al., 2020) and on signal processing over directed structures (Sandryhaila & Moura, 2013).

Table 9: Directed baselines with the polynomial filters' training budget. MagNet is run in three configurations (K, layers), each at its validation-selected charge q (grid 0–0.25, shown in parentheses); the configuration selected on validation is $(2, 5)$ on every dataset. Gap is DIR-POLY minus the selected MagNet. 5 seeded runs (fixed real-data graphs; DSBM regenerated per seed). Mean$_{\pm}$sample standard deviation of test accuracy (%).

	MAGNET (K, layers)					
dataset	$(2,2)$	$(5,2)$	$(2,5)$	Dir-GNN	DIR-POLY	gap
animal	$79.2_{\pm 1.1}$ (0.15)	$85.6_{\pm 2.7}$ (0.05)	$93.9_{\pm 0.9}$ (0.10)	$98.8_{\pm 1.0}$	$99.5_{\pm 0.6}$	$+5.6$
plant	$98.4_{\pm 0.6}$ (0.10)	$98.4_{\pm 1.1}$ (0.10)	$98.8_{\pm 0.8}$ (0.10)	$99.8_{\pm 0.2}$	$99.8_{\pm 0.2}$	$+0.9$
artifact	$97.3_{\pm 1.1}$ (0.15)	$97.6_{\pm 0.7}$ (0.20)	$98.5_{\pm 0.5}$ (0.15)	$99.7_{\pm 0.3}$	$99.5_{\pm 0.5}$	$+1.0$
DSBM	$97.3_{\pm 1.1}$ (0.25)	$97.8_{\pm 1.5}$ (0.25)	$99.7_{\pm 0.1}$ (0.25)	$99.4_{\pm 0.6}$	$98.9_{\pm 0.3}$	-0.8
cit-HepTh	$66.9_{\pm 1.6}$ (0.15)	$72.7_{\pm 1.0}$ (0.10)	$76.7_{\pm 1.4}$ (0.10)	$77.0_{\pm 1.5}$	$61.3_{\pm 1.6}$	-15.4

Table 10: Depth expressibility: held-out R^2 of the depth function fit on the degree-10 Krylov basis (double precision, Arnoldi-orthonormalized; mean over five random $50/50$ splits, fit on the training half). Constant input: the strict walk U versus the symmetric-normalized adjacency $\hat{A}_{\text{gcn}}$; the row-stochastic operators (the symmetric walk, P, A_α) produce only a constant predictor there. Root-marking seed $\mathbf{1} - \mathbf{1}_{\mathcal{R}}$: U, the deployed P, and the symmetric walk. On the path, degree 10 is below the longest path length (39), so no operator is exact yet, and with the root seed all three coincide.

	constant input		root-marking seed		
graph	U	$\hat{A}_{\text{gcn}}$	U	P	sym. walk
random tree (n=127)	1.00	0.06	1.00	1.00	0.45
directed path (n=40)	0.47	-1.86	0.48	0.48	0.48
WordNet animal	0.98	0.19	0.98	0.86	0.50
WordNet plant	0.99	0.18	0.99	1.00	0.46
WordNet artifact	0.99	0.31	0.99	1.00	0.43