How Does Label Noise Affect the Learning Curves of Graph Neural Networks?

Ivan Markov[1]

Supervisors: Elvin Isufi[1]**, Mohamed Jebali**[1]**, Chengen Liu**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Ivan Markov
Final project course: CSE3000 Research Project
Thesis committee: Elvin Isufi[1], Mohamed Jebali[1], Chengen Liu[1], Tom Viering[1]

Abstract

Graph Neural Networks (GNNs) achieve strong performance on node classification tasks, but their effectiveness often depends on the quality of the supervision, and real-world labels are often noisy. Learning curves - which describe how test performance scales with the number of labelled training nodes - have been extensively studied in classical machine learning, but their behaviour under realistic annotation noise in GNNs remains poorly explored. We present a systematic empirical study of how three label noise protocols - symmetric random flipping, feature-dependent asymmetric flipping, and structure-dependent flipping - affect the learning curve shape of ChebNet across four benchmark graphs spanning homophilic and heterophilic structure, at noise rates $\eta \in \{0.1, 0.3, 0.5\}$. The central finding is that noise does not simply shift the learning curve downward: above a moderate noise rate it reduces the effective slope, so the gap between clean and noisy performance widens as the label budget grows. Feature-dependent asymmetric noise is consistently the most harmful protocol across all datasets and budgets for $\eta \geq 0.3$, while structure-dependent noise is the least harmful on homophilic graphs. On graphs where the model already operates near its performance limit, noise type has little practical effect. These findings suggest that beyond a moderate noise rate, cleaning existing labels yields greater returns than acquiring more noisy ones, and that the nature of annotation error interacts with graph structure in ways that single-budget evaluations cannot detect.

1 Introduction

Graph Neural Networks have established themselves as one of the most powerful models for machine learning on graph-structured data, achieving strong results on tasks ranging from drug discovery to fraud detection [Zhou *et al.*, 2020]. The task of semi-supervised node classification is particularly prevalent: a model is trained on a labelled subset of nodes and must generalise to capture the signal of the rest of the graph. A natural and practically important question is how the performance of such a model scales with the size of the labelled training set n_{n} - the *learning curve*. Learning curves have been extensively studied in classical machine learning [Viering and Loog, 2023], where they exhibit well-characterised shapes including monotone improvement, diminishing returns, and, under certain conditions, non-monotone phenomena such as peaking and dipping. However, their behaviour in GNNs remains poorly understood, particularly under realistic conditions that deviate from the clean-label assumption made by most benchmarks. In practice, node labels are rarely clean: citation networks contain papers with wrong categories; web graphs are annotated by automated pipelines prone to systematic errors; crowd-sourced labels carry inherent disagreement [Frénay and Verleysen, 2014]. Label noise is therefore not an edge case but the norm of real-world GNN deployment, and understanding how it distorts the learning curve has direct consequences on the choices behind data collection.

Two substantial directions have been investigated regarding the issue. On one side, label noise in machine learning has been well studied: corrupted labels degrade generalisation and increase sample complexity, meaning substantially more data is required to achieve the same generalisation as under clean labels [Natarajan *et al.*, 2013; Frénay and Verleysen, 2014]. Noise-robust training strategies specific to GNNs have also been proposed [Dai *et al.*, 2021; Qian *et al.*, 2023]. On the other side, learning curves in deep learning have been shown to follow approximate power-law relationships with training set size [Hestness *et al.*, 2017; Rosenfeld *et al.*, 2020], and a comprehensive taxonomy of well-behaved and ill-behaved curve shapes has been established [Viering and Loog, 2023]. What neither line of work addresses is how these two phenomena interact with each other: whether label noise merely shifts the GNN learning curve downward, prematurely flattens it, or deforms its shape in ways that depend on the graph structure.

These two bodies of work surround our question without answering it. Noise-robust GNN methods operate at a fixed label budget; they do not ask how model performance *scales* with the budget as noise increases [Dai *et al.*, 2021; Qian *et al.*, 2023]. Learning curve studies in deep learning assume independent, identically distributed samples and clean labels. These conditions fail for GNNs, where representations in the same neighbourhood are aggregated and a single corrupted label distorts the representations of all nodes that aggregate over it. Standard GNN benchmarks, including the widely used 20-labels-per-class Cora split [Kipf and Welling, 2017], report accuracy at a single arbitrarily chosen budget, making it impossible to assess whether any performance advantage persists as the label set grows or shrinks. In this work we study how noise rate, noise structure, and graph homophily jointly determine the shape of the GNN learning curve.

We study the following main research question: *how does training label noise affect the learning curve shape of a GNN on node classification tasks?* Using ChebNet [Defferrard *et al.*, 2016] as a representative spectral GNN, we conduct a controlled experiment on four benchmark graphs: two homophilic citation networks, Cora [Kipf and Welling, 2017] ($h \approx 0.81$) and Pubmed [Yang *et al.*, 2016] ($h \approx 0.80$), and two heterophilic Wikipedia page-networks, Chameleon-filtered [Platonov *et al.*, 2023] ($h \approx 0.23$) and Squirrel-filtered [Platonov *et al.*, 2023] ($h \approx 0.20$), with h being the edge homophily ratio - the fraction of the edges in the graph that connect nodes with the same label. The noise rate $\eta \in \{0.1, 0.3, 0.5\}$ is investigated across three noise protocols: *symmetric* random label flipping; *feature-dependent asymmetric* flipping, where nodes furthest from their class centroid in feature space are more likely to be flipped to the label of the nearest other class centroid; and *structure-dependent* flipping, where nodes embedded in heterophilic neighbourhoods receive a higher probability of symmetric label corruption. This design separates the effect of noise severity from noise structure, and assesses whether homophilic and

heterophilic graphs respond differently.

The remainder of this paper is organised as follows. Section 2 establishes the mathematical foundations of spectral graph convolution, learning curve theory, and label noise models. Section 3 describes the noise protocols and the experiment procedure. Section 4 presents the experimental setup and the learning curves of all noise conditions. Section 5 interprets the findings, discusses validity, and offers recommendations for practitioners. Section 6 addresses reproducibility and responsible research. Section 7 concludes and outlines future directions.

2 Background

Spectral graph convolution. Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be an undirected graph, with vertices $\mathcal{V}$, edges $\mathcal{E}$, adjacency matrix $\mathbf{A}$, degree matrix $\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})$, and node-feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The symmetrically normalised Laplacian,

$$\tilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{1}$$

encodes graph connectivity with eigenvalues in $[0, 2]$. ChebNet [Defferrard *et al.*, 2016] parameterises graph filters as degree-K Chebyshev polynomial expansions of the rescaled Laplacian $\hat{\mathbf{L}} = 2\tilde{\mathbf{L}}/\lambda_{\max} - \mathbf{I}$:

$$g_{\boldsymbol{\theta}}(\hat{\mathbf{L}}) = \sum_{k=0}^{K} \theta_k \, T_k(\hat{\mathbf{L}}), \tag{2}$$

where $T_k(x) = 2x\,T_{k-1}(x) - T_{k-2}(x)$, $T_0 = 1$, $T_1 = x$, making each filter exactly K-hop localised without requiring an eigendecomposition. For semi-supervised node classification, the model is trained on n_{n} labelled nodes and evaluated on a disjoint test set, using the full graph structure throughout. GCN [Kipf and Welling, 2017] is the first-order special case ($K = 1$) and performs well on homophilic benchmarks, but how performance scales with n_{n}, and how that scaling is affected by label noise, has not been systematically studied.

Learning curves. The learning curve $f(n_{\mathrm{n}})$ maps the number of labelled training examples to expected test error [Viering and Loog, 2023]. Curves can be *well-behaved* (monotone decreasing), *peaking* (a transient error increase before recovery), or *dipping* (persistent degradation with more data) [Viering and Loog, 2023]. Natarajan *et al.* [Natarajan *et al.*, 2013] show that under symmetric noise rate η, sample complexity scales as $(1 - 2\eta)^{-2}$ times the clean-label requirement: at $\eta = 0.3$ roughly six times as many examples are needed; at $\eta = 0.5$ the factor diverges. This predicts that noise raises the error floor *and* reduces the slope, deforming the curve's shape rather than just shifting its level.

Label noise models. Following Frénay and Verleysen [Frénay and Verleysen, 2014], label corruption is characterised by a transition matrix $\mathbf{T}$, where $T_{jk} = P(\tilde{y}_i = k \mid y_i = j)$. *Symmetric* noise corrupts labels uniformly at random: $T_{jj} = 1 - \eta$ and $T_{jk} = \eta/(C-1)$ for $j \neq k$. *Asymmetric* noise introduces structured confusions between specific classes, modelling systematic annotation errors. *Instance-dependent* noise makes corruption more likely for ambiguous or atypical nodes, with T_{jk} depending on the node's own features $\mathbf{x}_i$. In the GNN setting, all three types are amplified by neighbourhood aggregation: a single corrupted label distorts the representations of every node that aggregates over it [Dai *et al.*, 2021]. These archetypes directly motivate the three experimental noise protocols introduced in Section 3.

3 Method

We conduct a controlled experiment of how training label noise affects the shape of GNN learning curves on node classification. Using ChebNet [Defferrard *et al.*, 2016] as a representative spectral GNN, we measure test performance as a function of the labelled training budget n_{n} across four benchmark graphs - two homophilic citation networks (Cora and Pubmed) and two heterophilic Wikipedia page-networks (Chameleon-filtered and Squirrel-filtered). Three noise protocols are applied to the training labels at rates $\eta \in \{0.1, 0.3, 0.5\}$; a clean baseline ($\eta = 0$) is included for comparison. Validation and test labels remain clean under all conditions. All factors other than the noise protocol and noise rate are kept constant across conditions, ensuring that any observed change in learning-curve shape is attributable to label noise alone. The three protocols are defined in the remainder of this section and detailed experimental setup is presented in Section 4.

Noise injection protocols. Label noise is applied only to training labels; validation and test labels remain clean throughout. We study three protocols at rates $\eta \in \{0.1, 0.3, 0.5\}$, plus a clean baseline ($\eta = 0$). In each noisy run, exactly $\lfloor \eta \cdot n_{\mathrm{n}} + 0.5 \rfloor$ training nodes are selected for corruption. The random generator is seeded deterministically as $s_{\text{noise}} = s + 1009\,n_{\mathrm{n}} + 9176\,i_{\text{split}}$, making every corrupted label set exactly reproducible.

Feature-dependent asymmetric and structure-dependent noise occupy complementary positions in the space of annotation error - one targeting ambiguity in feature space, the other targeting ambiguity in graph structure - while symmetric noise provides an unbiased random baseline.

Symmetric noise. Each selected node's label is replaced uniformly at random from the $C - 1$ remaining classes,

$$P(\tilde{y}_i = c \mid y_i) = \begin{cases} 1 - \eta & \text{if } c = y_i, \\ \eta/(C-1) & \text{if } c \neq y_i, \end{cases} \tag{3}$$

with collision rejection sampling to guarantee a strict label change. This is the standard symmetric noise baseline of Section 2 [Natarajan *et al.*, 2013].

Feature-dependent asymmetric noise. Corruption targets the most ambiguous training nodes: those lying farthest from their true class in feature space. Let $\mathcal{T}_c = \{i \in \mathcal{L} : y_i = c\}$ be the set of labelled training nodes with true class c. The class centroid in feature space is

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{T}_c|} \sum_{i \in \mathcal{T}_c} \mathbf{x}_i, \tag{4}$$

and each training node i receives selection weight $w_i = \|\mathbf{x}_i - \boldsymbol{\mu}_{y_i}\|_2$, proportional to its distance from its own class centre. Exactly $\hat{N} = \lfloor \eta \cdot n_{\mathrm{n}} + 0.5 \rfloor$ nodes are drawn without replacement using these weights, and each selected node is

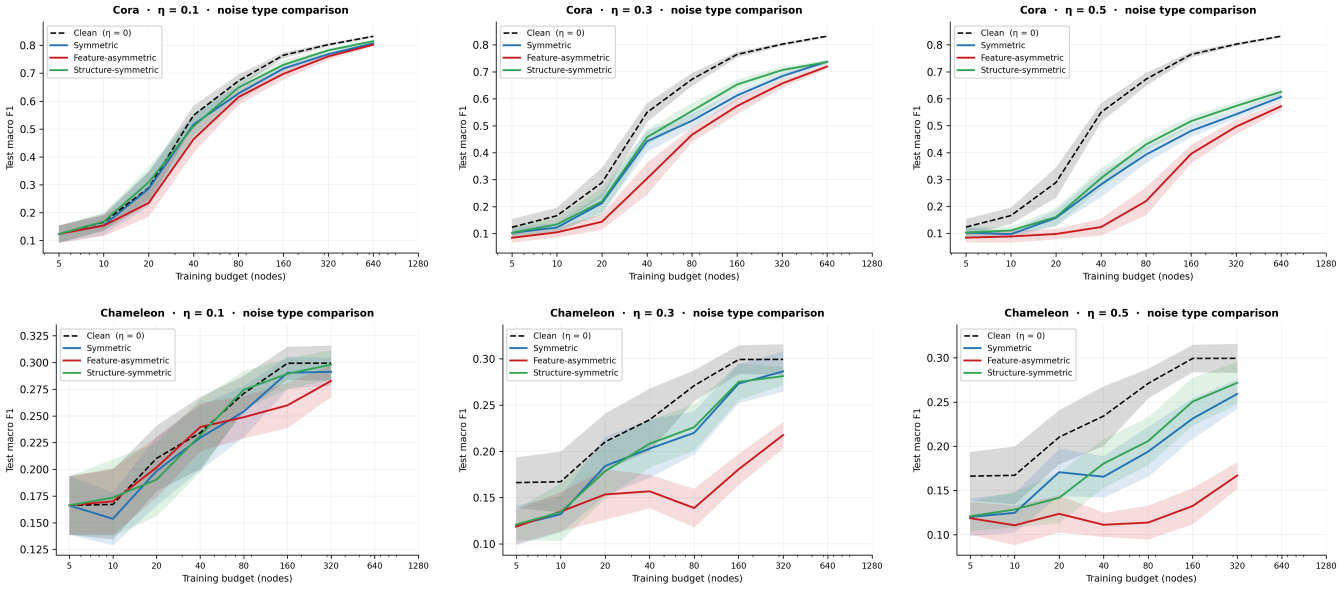

Figure 1: Test macro-F1 (higher is better) versus labelled training budget n_n for Cora (*top row*, $h \approx 0.81$) and Chameleon-filtered (*bottom row*, $h \approx 0.23$) at noise rates $\eta \in \{0.1, 0.3, 0.5\}$. Dashed black: clean baseline ($\eta = 0$); solid curves: symmetric (blue), feature-asymmetric (red), and structure-dependent (green) noise protocols. Shaded bands: 95% bootstrap confidence intervals over $K = 20$ independent runs. Test labels are clean under all conditions. The widening gap between the clean baseline and the noisy curves as n_n grows indicates slope reduction rather than a uniform downward shift.

flipped to the label of the nearest *other* class centroid:

$$\tilde{y}_i = \arg\min_{c \neq y_i} \|\mathbf{x}_i - \boldsymbol{\mu}_c\|_2. \tag{5}$$

This realises the instance-dependent noise model of Section 2 and models realistic crowd-sourced annotation, where non-expert labellers reliably agree on clear examples but consistently mislabel boundary-region instances - a paper that spans multiple research fields, a web page that fits multiple categories. When an error occurs, the nearest plausible alternative tends to be assigned rather than a random one.

Structure-dependent noise. Corruption targets training nodes embedded in heterophilic neighbourhoods. The local heterophily of node v,

$$h_v = \frac{1}{\deg(v)} \sum_{u \in \mathcal{N}(v)} \mathbf{1}[y_u \neq y_v], \tag{6}$$

measures the fraction of v's neighbours belonging to a different class. Each training node i receives selection weight $w_i = h_i$. Exactly $\hat{N}$ nodes are drawn without replacement using these weights, and each selected node is flipped symmetrically (Eq. (3)). The local heterophily h_v is computed over the *full* graph using the true labels of all nodes, including those in the validation and test sets. This is a deliberate design choice necessary due to the sparsity of homophilic citation graphs: Cora and Pubmed have average degrees of only 3.8 and 4.4 respectively, so the expected number of labelled neighbours per node at budget n_{n} is $\overline{d} \cdot n_{\text{n}}/n$, where $\overline{d}$ is the mean degree and n the total node count. At the largest budget in our grid ($n_{\text{n}} = 640$) for Cora and ($n_{\text{n}} = 1280$) for PubMed, this expectation reaches only $3.8 \times 640/2708 \approx 0.9$ for Cora and $4.4 \times 1280/19717 \approx 0.3$ for Pubmed; at smaller budgets it is effectively zero. Restricting neighbourhood labels to the labelled training pool would therefore drive $h_v = 0$ for the vast majority of training nodes, collapsing w_i to zero, eliminating any differentiation between nodes, and degenerating the protocol into symmetric noise, making it uninformative for the study. Using full graph labels ensures that h_v reflects the true structural position of each node. The true labels of validation and test nodes enter *only* the weight computation in Eq. (6); they are never exposed to the model during training, and both evaluation sets remain clean throughout.

This noise protocol models annotation errors that arise when labellers use neighbourhood context rather than node content to assign labels - a paper classified by the fields it cites, a web page assigned a category by the domains it links to. When a node sits in a heterophilic neighbourhood, the context surrounding it is contradictory, and no single alternative class is certainly more plausible than another, making the error direction effectively random.

4 Results

Experimental Setup

Datasets and split policy. We experiment on four benchmark graphs spanning two levels of homophily: two homophilic citation networks (Cora, Pubmed) and two heterophilic Wikipedia page-networks (Chameleon-filtered, Squirrel-filtered). The edge homophily ratio,

$$h = \frac{|\{(u, v) \in \mathcal{E} : y_u = y_v\}|}{|\mathcal{E}|}, \tag{7}$$

measures the fraction of edges that connect same-class nodes [Zhu *et al.*, 2020]. High h favours standard message-passing: neighbourhood aggregation smooths features within classes and sharpens between-class boundaries. Low h introduces an adversarial aggregation effect in which neighbours predominantly carry multi-class signal, preventing meaningful information aggregation, suppressing accuracy and inducing non-monotone learning-curve behaviour even under clean labels.

Cora [Kipf and Welling, 2017] is a citation network ($n = 2{,}708$, $|\mathcal{E}| = 5{,}278$, $d = 1{,}433$, $C = 7$, $h \approx 0.81$). We use the standard PyTorch Geometric split with 500 validation and 1,000 test nodes; the remaining 1,208 nodes form the training pool, giving a maximum label budget of $n_{\mathrm{n}} = 640$.

Pubmed [Yang *et al.*, 2016] is a citation network ($n = 19{,}717$, $|\mathcal{E}| = 44{,}334$, $d = 500$, $C = 3$, $h \approx 0.80$). We use the standard PyTorch Geometric split with 500 validation and 1,000 test nodes; the remaining 18,217 nodes form the training pool, giving a maximum label budget of $n_{\mathrm{n}} = 1280$.

Chameleon-filtered [Platonov *et al.*, 2023] is a Wikipedia page-network ($n = 890$, $|\mathcal{E}| = 8{,}854$, $d = 2{,}325$, $C = 5$, $h \approx 0.23$). We use the filtered variant of Platonov *et al.*, which removes duplicate nodes that artificially inflate accuracy [Platonov *et al.*, 2023]. The 10 pre-defined splits provide between 268-310 validation and between 159-194 test nodes, leaving at least 409 training nodes and a maximum budget of $n_{\mathrm{n}} = 320$.

Squirrel-filtered [Platonov *et al.*, 2023] is a Wikipedia page-network ($n = 2223, |\mathcal{E}| = 46{,}998, d = 2089, C = 5, h \approx 0.20$). We use the filtered variant of Platonov *et al.*. The 10 pre-defined splits provide between 691-739 validation and 416-483 test nodes, leaving 1047-1085 training nodes and a maximum budget of $n_{\mathrm{n}} = 640$.

Across all datasets, the labelled training budget n_{n} is varied over the geometric grid $\{5, 10, 20, 40, 80, 160, 320, 640, 1280\}$, truncated at the first value that exceeds the training pool. At each budget, n_{n} nodes are drawn uniformly at random from the pool, with a fresh draw per run.

Model and hyperparameters. All experiments use ChebNet (Eq. (2)), chosen as a well-established graph neural network architecture so that the study focuses on the effect of label noise rather than differences between model architectures. The model uses the following fixed configuration: 2 convolutional layers, Chebyshev order $K = 3$, hidden width 64, ReLU activations, dropout 0.5, Adam optimiser [Kingma and Ba, 2015] with learning rate 10^{-2} and weight decay 5×10^{-4}, and 200 training epochs with early stopping with patience 50, monitoring the validation cross-entropy metric. Hyperparameters were chosen to follow commonly used settings for semi-supervised node classification (e.g. [Kipf and Welling, 2017]) and were kept fixed across all four datasets, noise protocols, noise rates, and label budgets, ensuring that any observed change in the shape of the learning-curve is to be attributed to label noise rather than incidental effects of model-selection.

Statistical analysis. Each experimental condition (dataset $\times$ protocol $\times$ η $\times$ n_{n}) is replicated across $K = 20$ independent runs. For Cora and Pubmed, which use a single standard split, we replicate across 20 seeds. For Chameleon-filtered and Squirrel-filtered, which provide 10 pre-defined splits, we use 2 seeds per split ($K = 20$); the variance reported for these datasets therefore reflects both sampling randomness and split-to-split variation in the train/val/test partitions. We report **macro-averaged F1** (macro F1) as the primary performance metric and **test cross-entropy** as a secondary diagnostic, both computed as means across the K runs and accompanied by 95% bootstrap confidence intervals. When confidence intervals for two conditions do not overlap, the difference is treated as practically significant.

Macro F1 is the unweighted average of the per-class F1 score across all C classes:

$$F1_c = \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \qquad \mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} F1_c, \tag{8}$$

where TP_c, FP_c, and FN_c are the true positives, false positives, and false negatives for class c. Averaging uniformly across classes ensures that each class contributes equally regardless of its frequency, making macro F1 robust to the class-size imbalances present in our benchmark datasets. Unlike accuracy, which a model can inflate by predicting predominantly the majority class, macro F1 penalises poor performance on any individual class equally, giving a more reliable assessment of performance across the complete label space.

Test cross-entropy (CE) is the mean negative log-likelihood of the true class over all test nodes:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{V}_{\mathrm{test}}|}\sum_{i \in \mathcal{V}_{\mathrm{test}}} \log p_{i,\,y_i}, \tag{9}$$

where p_{i,y_i} is the model's predicted probability assigned to the true class of node i. Cross-entropy is chosen as a secondary diagnostic because it evaluates not only the classification decision of a model but also the prediction distribution - the confidence behind the assigned label. Under label noise, the model may retain the correct classification for a test node while assigning reduced confidence to the true class. F1 is blind to this degradation, whereas cross-entropy registers it directly. This makes cross-entropy more sensitive to how label noise distorts the model's learned distribution.

Numerical Results

We report test macro-F1 as the primary metric throughout; test cross-entropy is reported where it reveals curve-shape properties not visible in F1. In all figures, the dashed black curve is the clean baseline ($\eta = 0$); solid curves show symmetric (blue), feature-asymmetric (red), and structure-dependent (green) noise protocols. Shaded bands are 95% bootstrap confidence intervals over $K = 20$ runs; non-overlapping intervals are treated as practically significant.

Cora ($h \approx 0.81$). Figure 1 (top row) shows the Cora macro-F1 learning curves. All three noisy protocols improve monotonically but with a gap relative to clean label performance that widens as n_{n} grows. Slope reduction is most clearly visible in the cross-entropy curves of Figure 2. At $\eta = 0.1$ the effect is budget-dependent: the curves are indistinguishable

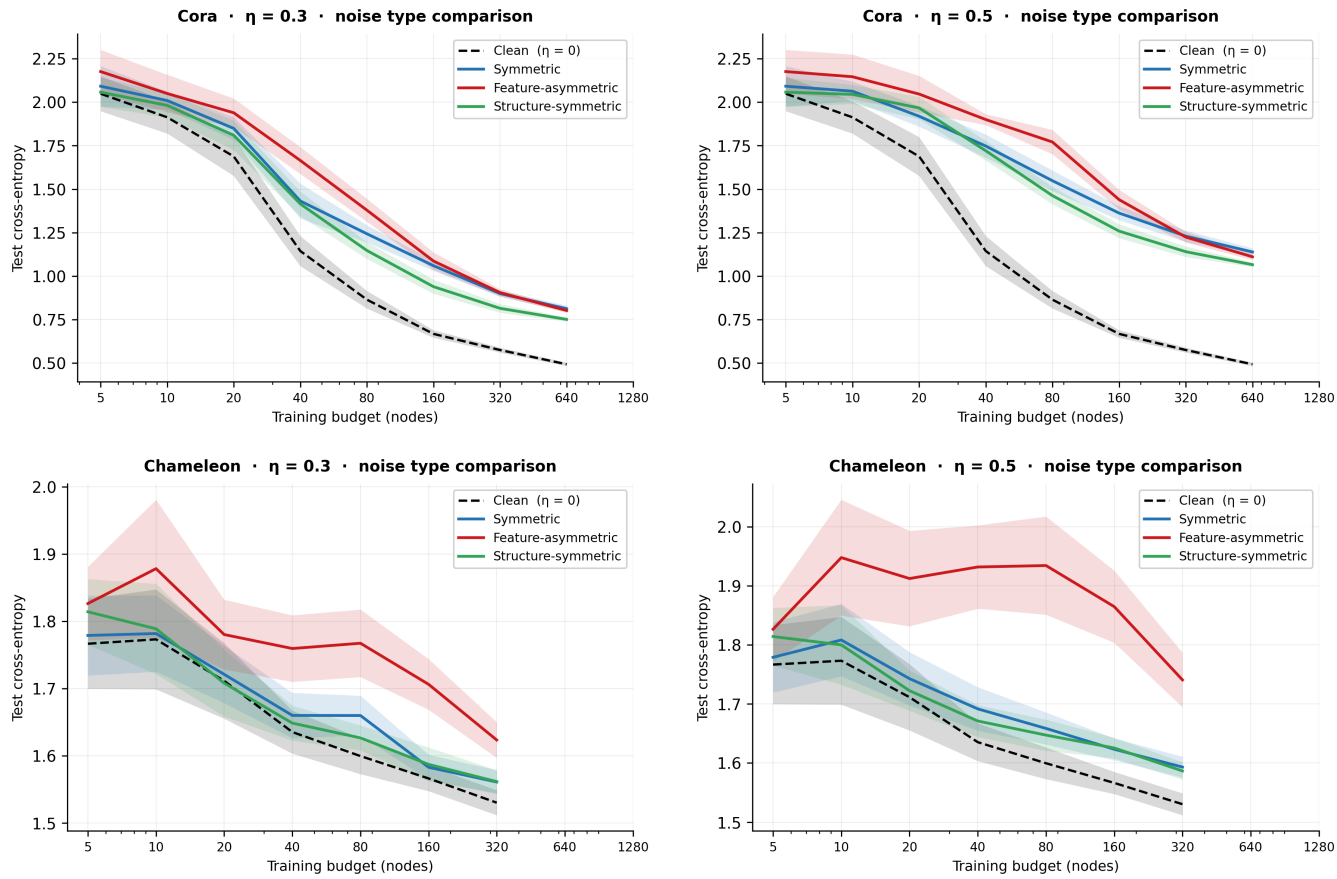

Figure 2: Test cross-entropy (lower is better) versus labelled training budget n_n for Cora (*top row*) and Chameleon-filtered (*bottom row*) at $\eta \in \{0.3, 0.5\}$. Dashed black: clean baseline ($\eta = 0$); solid curves: symmetric (blue), feature-asymmetric (red), and structure-dependent (green) noise protocols. Shaded bands: 95% bootstrap confidence intervals over $K = 20$ independent runs. Test labels are clean under all conditions. *Top row*: the gap between clean and noisy cross-entropy widens with n_n, demonstrating slope reduction rather than a uniform level shift. *Bottom right*: feature-asymmetric noise at $\eta = 0.5$ exhibits non-monotone cross-entropy, peaking at $n_n = 10$ and remaining elevated through $n_n = 80$.

from clean until $n_{\mathrm{n}} = 40$ for feature-asymmetric and until $n_{\mathrm{n}} = 160$ for the other two protocols. Across $\eta \in \{0.3, 0.5\}$, feature-asymmetric is consistently the most harmful at every budget and structure-dependent the least, with the ranking most pronounced at small n_{n} and narrowing - but not closing - at maximum budget (Table 1). At $\eta = 0.5$, the feature-asymmetric gap below clean reaches 0.26 F1 at $n_{\mathrm{n}} = 640$; at $\eta = 0.3$, symmetric and structure-dependent are statistically tied throughout (Table 2).

Chameleon-filtered ($h \approx 0.23$). Figure 1 (bottom row) shows Chameleon macro-F1 learning curves. Performance is the lowest in the study; the clean baseline is well-behaved and broadly monotone - no oscillation is present. At $\eta = 0.1$, all noisy conditions remain within the clean confidence interval at every budget. At $\eta \geq 0.3$, feature-asymmetric is the worst protocol by a clear and growing margin; the protocol ranking is otherwise consistent with Cora (Table 1). The cross-entropy curves (Figure 2, bottom row) reveal a non-monotone learning curve: at $\eta = 0.5$, feature-asymmetric CE rises to a peak at $n_{\mathrm{n}} = 10$, remains elevated through $n_{\mathrm{n}} = 80$, then descends. Every other condition decreases monotonically throughout.

Pubmed ($h \approx 0.80$). Figure 3 (top row) shows Pubmed results. The pattern closely replicates Cora: feature-asymmetric is the most harmful at every noise level, structure-dependent the least harmful, and the gap to clean label performance widens with budget. The slope-reduction effect persists to the maximum budget ($n_{\mathrm{n}} = 1280$) and is not absorbed by the larger training pool, confirming it as a noise-rate effect rather than a sample-size effect (Table 1).

Squirrel-filtered ($h \approx 0.20$). Figure 3 (bottom row) shows Squirrel-filtered results. The clean learning curve is near-flat and the model's performance ceiling is the lowest in this study, with gains across the entire budget range totalling less than 0.09 F1. Noise effects are markedly compressed: all three protocols remain close together and close to clean at every budget, with no protocol achieving practical significance at $\eta = 0.3$; structure-dependent noise performance never separates from clean at any noise rate (Table 2).

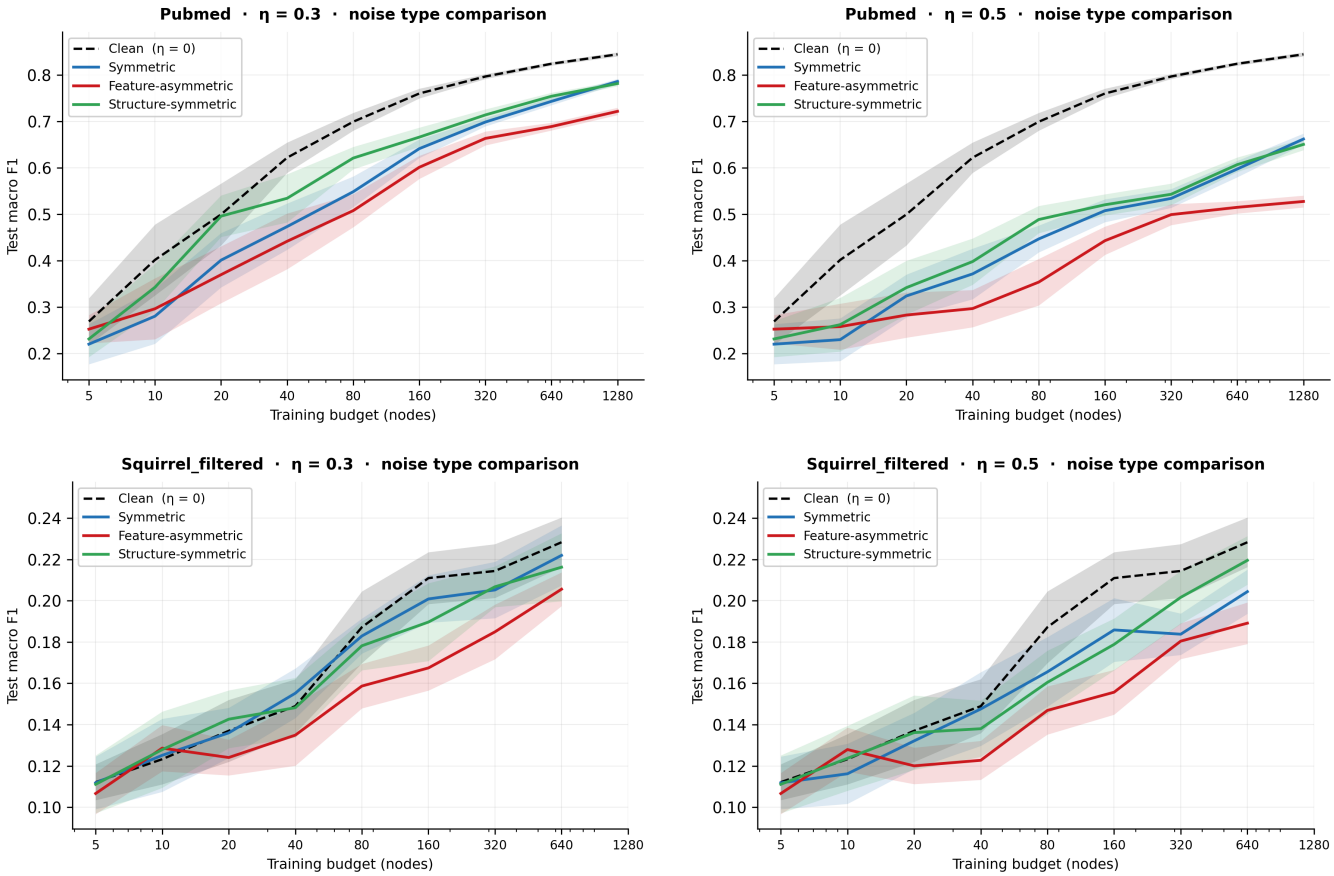

Figure 3: Test macro-F1 (higher is better) versus labelled training budget n_n for Pubmed (*top row*, $h \approx 0.80$) and Squirrel-filtered (*bottom row*, $h \approx 0.20$) at $\eta \in \{0.3, 0.5\}$. Dashed black: clean baseline ($\eta = 0$); solid curves: symmetric (blue), feature-asymmetric (red), and structure-dependent (green) noise protocols. Shaded bands: 95% bootstrap confidence intervals over $K = 20$ independent runs. Test labels are clean under all conditions. Pubmed replicates the Cora protocol ranking from Figure 1; Squirrel shows near-uniform noise response with all conditions within a 0.031 macro-F1 band at $\eta = 0.5$.

Table 1: Test macro-F1 (higher is better) at maximum label budget ($n_{n,\max}$: 640 for Cora and Squirrel, 1280 for Pubmed, 320 for Chameleon) for all datasets and noise conditions. **Bold**: lowest macro-F1 among noisy protocols at that noise rate - indicating the most damaging protocol. At $\eta = 0.1$, differences among noisy protocols on heterophilic datasets are generally within confidence intervals and no consistent protocol ranking should be inferred.

		$\eta = 0.1$			$\eta = 0.3$			$\eta = 0.5$		
Dataset	Clean	Sym	Feat	Struct	Sym	Feat	Struct	Sym	Feat	Struct
Cora	0.832	0.807	**0.802**	0.815	0.738	**0.720**	0.736	0.607	**0.572**	0.626
Pubmed	0.844	0.825	**0.815**	0.828	0.786	**0.722**	0.782	0.662	**0.528**	0.651
Chameleon	0.299	0.291	**0.283**	0.298	0.286	**0.218**	0.281	0.259	**0.167**	0.272
Squirrel	0.228	**0.220**	0.224	0.226	0.222	**0.206**	0.216	0.204	**0.189**	0.220

Cross-dataset comparison. Two findings are consistent across all four datasets. Feature-asymmetric noise is the worst protocol at every dataset and budget for $\eta \geq 0.3$ - the finding most robust to dataset variation (Tables 1 and 2). Low noise ($\eta = 0.1$) has no practically significant effect at small budget on any dataset. On the homophilic pair it becomes detectable at large budget, while on both heterophilic datasets it remains almost indistinguishable from clean throughout. Beyond these shared patterns, the four datasets form two groups: Cora and Pubmed show consistent slope reduction with a stable protocol ranking. Chameleon shows a clear and widening ranking with non-monotone cross-entropy curves; Squirrel shows near-uniform noise response with effects compressed near a performance ceiling the model cannot exceed under any labelling condition.

Table 2: Test macro-F1 (higher is better) at $n_n = 40$ (small-budget regime) for all four datasets across all noise conditions. The protocol ranking gap is widest at this budget level. **Bold**: lowest macro-F1 among noisy protocols at that noise rate - indicating the most damaging protocol. For Squirrel at $\eta = 0.1$, all values are near chance level and no consistent ordering relative to clean performance exists; no protocol ranking should be inferred.

		$\eta = 0.1$			$\eta = 0.3$			$\eta = 0.5$		
Dataset	Clean	Sym	Feat	Struct	Sym	Feat	Struct	Sym	Feat	Struct
Cora	0.549	0.516	**0.463**	0.510	0.442	**0.304**	0.457	0.282	**0.123**	0.305
Pubmed	0.622	0.584	**0.575**	0.595	0.474	**0.442**	0.535	0.372	**0.297**	0.398
Chameleon	0.234	**0.229**	0.240	0.232	0.203	**0.157**	0.208	0.166	**0.111**	0.181
Squirrel	0.149	0.160	**0.144**	0.146	0.155	**0.135**	0.148	0.148	**0.123**	0.138

5 Discussion

The results across four datasets establish that label noise does not act on GNN learning curves uniformly. Its effect is modelled by three interacting factors: noise structure (which nodes are corrupted and how), noise rate, and graph structure (homophilic or heterophilic). Depending on how these align, noise can reduce the learning-curve slope, trigger a non-monotone cross-entropy path, or leave the curve almost entirely unchanged. Each of these outcomes has a distinct interpretation.

Why feature-asymmetric noise is most harmful. Feature-asymmetric noise corrupts the nodes lying farthest from their class centroid in feature space and flips their labels to the nearest other class centroid. This protocol introduces a doubly concentrated disruption: it selects boundary-region nodes - those whose feature representations are most ambiguous with respect to their true class, whose correct labels are therefore most informative for establishing the decision boundary between competing classes - and replaces each label with the nearest alternative centroid assignment rather than a uniformly random one. The corruption is maximally confusing on two levels simultaneously: it targets the highest-value, boundary-region supervision signal and replaces it with the most feature-consistent incorrect alternative, so each corrupted label is both maximally damaging and maximally difficult to distinguish from a correct one on the basis of node features alone. At small label budget, when every labelled node exerts a disproportionately large influence on the gradient, corrupting them removes the highest value training signal when it is most scarce; the large F1 penalties at $n_n = 40$ in Table 2 are consistent with this interpretation across all four datasets and noise rates.

The observed narrowing of the protocol ranking gap at large n_n, visible across all four datasets, is also consistent with this view. One possible explanation is that as the label pool grows, an increasing proportion of clean, unambiguous interior-region nodes progressively dilutes the influence of corrupted boundary nodes on the gradient, reducing the targeted penalty of the protocol - degrading it at least partially towards symmetric noise. This mechanism is not directly verified here and should be treated as a hypothesis.

Noise deforms the learning-curve shape. The widening gap between clean and noisy cross-entropy curves on Cora and Pubmed (Figure 2) suggests that label noise changes the *shape* of the learning curve, not just its level. A uniform downward shift would keep the separation between clean and noisy curves constant as n_{n} grows; instead, the clean curve continues to fall steeply while noisy curves flatten, and the distance between them grows with each additional label. This is consistent with the sample-complexity result of Natarajan *et al.* [Natarajan *et al.*, 2013]: under noise rate η, the number of examples needed to achieve a given error scales as $(1 - 2\eta)^{-2}$ times the clean requirement, which means that each noisy label is worth less than a clean one by that factor. At a certain budget the noisy curve continues to improve but at a progressively reduced rate relative to the clean one.

The budget-dependent effect of low noise. The finding that $\eta = 0.1$ is undetectable at small budget but separable from clean at large budget can be interpreted as a competition between two kinds of variability. At small n_{n}, the natural stochasticity of training - random node sampling, weight initialisation, and gradient variance - masks any systematic effect that low-rate noise can produce; the corruption signal is swamped by sampling noise. As budget grows and training variance falls, even a 10% corruption rate accumulates enough influence to emerge above the confidence band. Feature-asymmetric noise crosses this threshold earlier ($n_{\mathrm{n}} = 40$) than symmetric and structure-dependent ($n_{\mathrm{n}} = 160$) because its targeting concentrates the signal loss on high-value boundary nodes, making each corrupted label more damaging than the nominal rate suggests. On Chameleon and Squirrel the base signal is already too weak for mild corruption to produce any detectable effect at any budget - the noise is lost in the model's existing confusion.

Why structure-dependent noise is least harmful on homophilic graphs. Structure-dependent noise targets nodes in heterophilic neighbourhoods - those for which a large fraction of edges connect to nodes of a different class. On Cora ($h \approx 0.81$) and Pubmed ($h \approx 0.80$), such nodes are the structural outliers: the graph is dominated by within-class edges, and heterophilic nodes sit at class boundaries, weakly embedded in cohesive clusters. A possible explanation of why this protocol is the least damaging is that for nodes in heterophilic positions, neighbourhood aggregation already mixes features from multiple classes and degrades their representations before any label corruption is applied, such that corrupting their labels adds comparatively little further damage. On Chameleon and Squirrel, where heterophilic edges are the majority, structure-dependent noise reaches almost the

entire training set - degrading the protocol towards symmetric noise - and the advantage is lost almost entirely.

The Chameleon cross-entropy peaking. The non-monotone cross-entropy trajectory of feature-asymmetric noise at $\eta = 0.5$ on Chameleon (Figure 2, bottom row) is the most distinctive shape in the study. A possible explanation is that it emerges from the compounding of two mechanisms that reinforce each other on this specific dataset. The first is ChebNet's aggregation behaviour under heterophily: with $h \approx 0.23$, the K-hop neighbourhood of most nodes contains predominantly cross-class features, which suggests the structural signal may actively mislead the model toward incorrect predictions regardless of label quality. The second is feature-asymmetric noise's targeting strategy, which selects nodes that are ambiguous in feature space. On Chameleon, structural and feature ambiguity may compound: the nodes that are hardest to classify from features alone are often those whose structural context is also most confusing. At very small budget ($n_n = 5$), the model lacks sufficient supervision to learn any consistent pattern and predicts near-randomly. As budget grows to $n_n = 10$, one possible explanation for the observed CE increase is that there are now enough corrupted labels from doubly-ambiguous nodes to steer the model toward systematically incorrect decision boundaries: rather than optimising toward the true label structure, the model may optimise toward the class assignments prescribed by the corrupted labels, which for these nodes point toward the nearest other centroid, shifting the learned class boundaries away from the true ones and driving cross-entropy on the clean test set upward. This CE peak may persist through the mid-budget range because the corrupted labels, being concentrated on boundary-region nodes, could disproportionately shape the gradient even as the label pool grows. Only at large budget may a sufficient influx of clean, unambiguous labels dilute the corrupted signal and reverse the trend. This is consistent with the peaking phenomenon catalogued by Viering and Loog [Viering and Loog, 2023], and the effect does not appear on Cora, possibly because homophilic aggregation acts as an error-correcting mechanism: even with a corrupted node label, consistent neighbour labels and features may allow ChebNet to partially recover from and smooth over the corrupted supervision.

Squirrel's noise immunity. Squirrel-filtered presents a simpler case than Chameleon. The clean learning curve is near-flat across the full budget range - ChebNet gains less than 0.09 F1 from $n_{\mathrm{n}} = 5$ to the maximum budget. This suggests that the model cannot extract meaningful decision signal from this graph regardless of how many labels it receives. When a model fails to learn useful representations even under clean supervision, corruption has nothing to degrade. There is no correct decision boundary to disrupt and no informative gradient to misdirect. In this capacity-limited regime, annotation noise type is irrelevant; the main constraint is the architecture's inability to handle a strongly heterophilic graph, not the quality of the labels.

Practical recommendations. Three recommendations follow from the combined findings. First, on homophilic graphs at $\eta \geq 0.3$, cleaning existing corrupted labels will provide greater performance gains than acquiring additional noisy ones. The slope reduction means each new noisy label is worth progressively less, and the ceiling imposed by the noise floor cannot be broken by scale alone. Second, when selecting which nodes to label - whether manually or via active learning - avoid strategies that systematically favour class-boundary or uncertain nodes; the feature-asymmetric noise protocol is precisely a model of what such strategies produce under imperfect annotation, and it is consistently the most damaging configuration, especially at small budget. Third, on highly heterophilic graphs where ChebNet already approaches its capacity limit, investment in label cleaning offers limited return; the priority in such settings should be architectural - models capable of selectively attending to relevant neighbours, such as graph attention networks [Veličković *et al.*, 2018], might benefit more from the same data than improved label quality alone.

These recommendations are specific to the experimental conditions studied here - a single spectral GNN on four benchmark graphs - and should be treated as motivated hypotheses rather than universal prescriptions.

Generalisability and threats to validity. Several constraints bound the scope of our findings. All experiments use a single architecture (ChebNet), and different models may respond differently. Attention-based models such as GAT [Veličković *et al.*, 2018] can in principle assign lower weights to heterophilic neighbours, so the mechanism that makes structure-dependent noise mild on homophilic graphs - Laplacian smoothing already degrading heterophilic node representations - may not hold for them. Both homophilic datasets are sparse citation networks with similar degree distributions; denser or directed graphs may produce different slope-reduction dynamics. Hyperparameters were fixed across all noise conditions, and tuning regularisation per noise rate could partially compensate for corrupted supervision and shift protocol rankings at high η. The clean validation and test assumption is a necessary experimental control but does not reflect settings where noise is pervasive throughout the graph. Finally, using full-graph true labels to compute structure-dependent selection weights is a design choice - restricting to the labelled pool collapses the protocol to symmetric noise on sparse graphs - but it cannot be applied in practice when true labels are genuinely unavailable.

6 Responsible Research

Data. All four datasets used in this study are publicly available benchmarks. Cora and Pubmed are accessible via PyTorch Geometric [Fey and Lenssen, 2019]. Chameleon-filtered and Squirrel-filtered are the cleaned variants released by Platonov *et al.* [Platonov *et al.*, 2023]. No proprietary, sensitive, or personally identifiable data was used at any stage of this study.

Reproducibility. All experiments are fully deterministic. Seeds for Cora and Pubmed runs are $s \in \{0, 1, \ldots, 19\}$. Seeds for Chameleon-filtered and Squirrel-filtered runs are $s \in \{0, 1\}$ within each of the 10 dataset splits. All random

operations - training-node sampling, model weight initialisation, and label corruption - are seeded deterministically from s (and from s_{noise} for corruption).

AI use. Claude (Anthropic) was used for writing assistance, including paraphrasing and structural suggestions, and for code review. All scientific claims, experimental design, analysis, and conclusions are authors' own.

Limitations. The main limitations are discussed in Section 5. Results are limited to a single architecture (ChebNet) on four benchmark graphs; hyperparameters were held fixed and not tuned per noise condition; and the clean validation and test label assumption is a controlled experimental choice that may not reflect real-world deployment.

7 Conclusion

We presented a systematic empirical study of how label noise affects the learning curve shape of ChebNet across four benchmark graphs spanning a wide range of homophily, using three noise protocols at three noise rates. The central finding is that label noise above $\eta = 0.3$ does not just shift the learning curve - it changes its shape. The gap between clean and noisy performance widens as the training budget grows, with each additional noisy label contributing progressively less than a clean label would.

Feature-asymmetric noise, which targets the most ambiguous nodes and assigns them the most plausibly wrong label, is the most harmful protocol on every dataset and budget level for $\eta \geq 0.3$. Its penalty is sharpest at small budget and narrows at large budget. Structure-dependent noise is the least harmful on homophilic graphs, possibly due to ChebNet's Laplacian smoothing already degrading the representations of the heterophilic nodes it targets. On Chameleon, the compounding of structural and feature-space ambiguity under high-rate feature-asymmetric noise produces a non-monotone cross-entropy trajectory - a concrete instance of the peaking phenomenon. Squirrel illustrates a somewhat different regime: when a model operates near its expressive capacity, noise type becomes secondary to architecture, and performance differences across protocols compress into a practically insignificant band.

For practitioners, the slope-reduction finding implies that on homophilic graphs at $\eta \geq 0.3$, cleaning existing labels is a better investment than acquiring more noisy ones, and annotation strategies that target boundary-region nodes should be avoided at small budget. Future work should test whether these findings extend to attention-based and noise-robust GNN architectures, examine how label-cleaning methods reshape the learning curve beyond simply improving fixed-budget performance, and probe the boundary between the slope-reduction regime observed on Cora and Pubmed and the capacity-limited regime observed on Squirrel across a broader and more structurally diverse set of graphs.

References

[Dai *et al.*, 2021] Enyan Dai, Charu Aggarwal, and Suhang Wang. NRGNN: Learning a label noise-resistant graph neural network on sparsely and noisily labeled graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 227–236, 2021.

[Defferrard *et al.*, 2016] 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)*, pages 3844–3852, 2016.

[Fey and Lenssen, 2019] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[Frénay and Verleysen, 2014] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[Hestness *et al.*, 2017] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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

[Natarajan *et al.*, 2013] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1196–1204, 2013.

[Platonov *et al.*, 2023] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily–heterophily dichotomy and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[Qian *et al.*, 2023] Siyi Qian, Haochao Ying, Renjun Hu, Jingbo Zhou, Jintai Chen, Danny Z. Chen, and Jian Wu. Robust training of graph neural networks via noise governance. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 607–615, 2023.

[Rosenfeld *et al.*, 2020] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations (ICLR)*, 2020.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Viering and Loog, 2023] Tom Viering and Marco Loog. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2023.

[Yang *et al.*, 2016] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning (ICML)*, 2016.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[Zhu *et al.*, 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.