How Does Labelled-Node Sampling Shape GNN Learning Curves?

A controlled study of uniform vs. stratified label sampling on homophilic and heterophilic graphs

Radu-Andrei Zidaru | r.a.zidaru@student.tudelft.nl | Supervisors: Elvin Isufi, Chengen Liu, Mohamed Jebali | EEMCS, TU Delft • CSE3000 Research Project, Q4 2025/2026

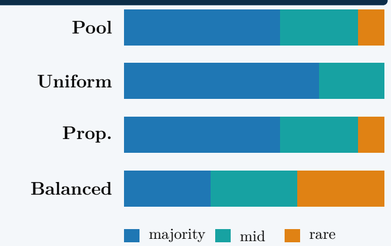
1. The Question

GNN papers report a *single* learning curve: accuracy vs. the number of labelled nodes n_n. At small n_n that curve hides which *nodes* were labelled.

- Does the random labelled set move accuracy?
- Does **stratified** sampling beat **uniform**?
- Does it differ for **homophilic** vs. **heterophilic** graphs?

We treat the labelled set as the *only* variable and measure its effect on the curve directly.

2. Three Ways to Pick n_n Labels



Uniform can drop a rare class; **proportional** keeps the data's true class prior; **balanced** forces an equal prior. The two stratified samplers differ *only* in the prior they impose.

3. Datasets

Dataset	Nodes	C	h
Cora	2,708	7	0.81
PubMed	19,717	3	0.80
Chameleon-filt.	890	5	0.24
Squirrel-filt.	2,223	5	0.22

Top two **homophilic** (h high: neighbours share labels); bottom two **heterophilic** (h low). C = classes, h = edge homophily.

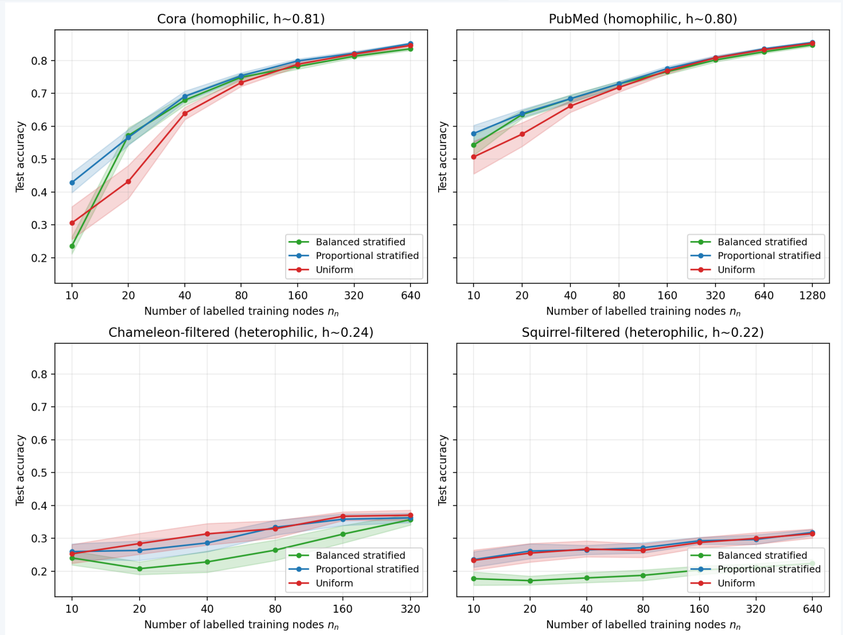
4. Setup & Model

Controlled design. Fix the model, hyperparameters, and the datasets' *published* splits; vary *only* the labelled set. K=20 repeats per budget; $n_n \in \{10, \ldots, 1280\}$.

Model. ChebNet (2 layers, order k=3, hidden 64) with Adam and early stopping, the canonical Kipf–Welling GCN recipe.

Why ChebNet? The representative spectral GNN whose polynomial filter underpins recent work (ChebNetII, BernNet, GPR-GNN); its simplicity keeps the effect attributable to the *labelled set*.

5. Learning Curves: the Headline



Mean of K=20 draws; bands are 95% bootstrap CIs; n_n on a log axis. **Top (homophilic)**: stratified (blue/green) beat uniform (red) at small n_n, then curves merge. **Bottom (heterophilic)**: proportional tracks uniform; balanced (green) sits below both.

6. Accuracy at $n_n = 20$

Dataset	Uniform	Prop.	Balanced
Cora (homo.)	0.43	0.57	**0.57**
PubMed (homo.)	0.58	**0.64**	0.64
Chameleon (hetero.)	**0.28**	0.26	**0.21**
Squirrel (hetero.)	0.26	**0.26**	**0.17**

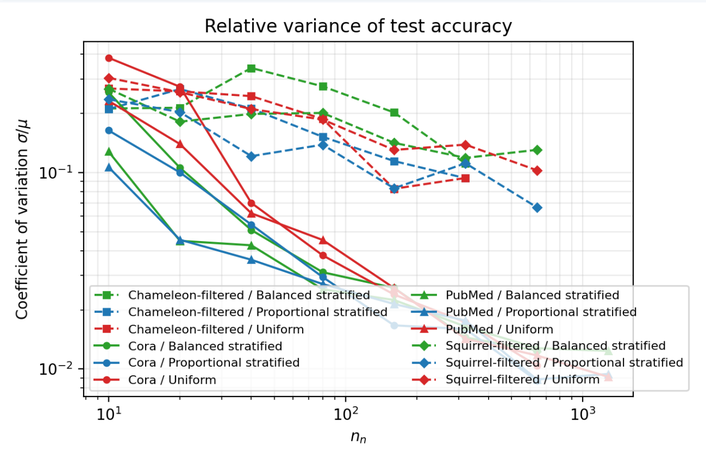
- **Homophilic:** stratified beats uniform by up to **+14 pp** and **halves** the variance.
- **Heterophilic:** proportional = uniform (no gain); balanced **actively hurts**.

7. But the Metric Matters: Macro-F_1 at $n_n = 20$

Dataset	Uniform	Prop.	Balanced
Cora (homo.)	0.31	0.47	**0.54**
PubMed (homo.)	0.51	0.60	**0.63**
Chameleon (hetero.)	0.20	**0.21**	0.16
Squirrel (hetero.)	0.14	**0.17**	0.14

Under macro-F_1 (every class weighted equally) balanced **wins** on the homophilic graphs. **The verdict on balanced sampling is metric-specific:** harmful for accuracy on skewed classes, helpful when rare classes matter.

8. Variance Tells the Same Story



Relative spread σ/μ vs. n_n. On homophilic graphs (solid) uniform is far noisier at small budgets; stratification removes it. On heterophilic graphs (dashed) there is no clean ordering.

9. Why the Regimes Differ

Homophilic. A labelled node is an *anchor* whose signal spreads to its same-class neighbourhood. Covering every class matters, so an unlucky uniform draw that misses a class is costly. Stratification then fixes it, raising the mean *and* cutting variance.

Heterophilic. Neighbours differ in class, so message passing carries little class signal and coverage stops mattering, leaving proportional = uniform. Balanced additionally *distorts the class prior* away from the (skewed) test set, hurting most on the most imbalanced graph (Squirrel). This is also why the harm reverses under macro-F_1, which no longer rewards matching that prior.

10. Takeaway & Next Steps

Report repeated draws with confidence intervals, never a single uniform curve. **Proportional stratified is the safe default:** it never significantly underperformed uniform on any dataset. **Avoid balanced** on skewed classes when accuracy is the target.

Limitations. Homophily covaries with size, features and class count, so it is not causally isolated; two graphs per regime.

Next: more architectures (GCN, GAT), more graphs per regime, a systematic multi-metric study.

Refs: Kipf & Welling 2017; Defferrard et al. 2016; He et al. 2022 (ChebNetII); Platonov et al. 2023; Benjamini & Hochberg 1995.