Property-Driven Comparison of GNNs on Multi-Label Graphs

Victor Paiu Supervisors: Megha Khosla and Elena Congeduti
Delft University of Technology

Introduction

Node classification on graphs appears in social, biological, and recommendation settings. Most GNN work focuses on multi-class graphs, where each node has one label. In multi-label graphs, two connected nodes may share all, some, or none of their labels, so a single scalar like homophily can hide important neighborhood patterns.

Understanding which dataset properties matter can guide model choice and explain why results differ across datasets. This motivates the research question:

How do structural, feature, and label properties of multi-label graphs influence the performance of GCN and H2GCN?

Models and Metrics

- **GCN**: repeated graph convolution on the normalized adjacency. It works well when direct neighbors carry reliable label signal.
- **H2GCN**: separates ego and neighborhood representations and uses higher-order neighborhoods. It is designed for heterophilic graph settings.
- **Metric**: macro average precision for multi-label classification.
- **Notation**: h - Jaccard homophily, $|C|$ - number of classes, $\bar{\ell}$ - mean labels per node, LI - label informativeness, $|F|$ - feature dimension.

Methodology

Synthetic graphs vary one property at a time, while real-world datasets anchor the findings.

1. **Generate** features + labels (hyperspheres) and edges (SDA) at a target h.
2. **Measure** all properties on the realised graph.
3. **Train** GCN and H2GCN over 3 seeds, report macro AP and F_1.
4. **Pool** all 97 graphs into a Ridge regression that weighs properties jointly.

Real-World Datasets

Dataset	h	density	clustering	MeanIR	unlabeled	GCN AP	H2GCN AP
BlogCat	0.10	1.25%	0.46	15.4	0.0%	0.037	0.039
OGB-Proteins	0.15	0.44%	0.28	6.4	40.3%	0.054	0.036
PCG	0.17	0.72%	0.34	3.6	0.0%	0.210	0.192
Yelp	0.22	0.003%	0.09	17.4	0.1%	0.131	0.226
HumLoc	0.42	0.38%	0.13	15.3	0.0%	0.252	0.172
EukLoc	0.46	0.05%	0.14	45.0	0.0%	0.152	0.134
DBLP	0.76	0.02%	0.61	1.6	0.0%	0.893	0.858

Table 1: Properties and macro AP for the seven real-world multi-label graphs (sorted by h).

Results: Homophily Sweep

We train both models on 9 SDA graphs at $h \in \{0.2, 0.3, \ldots, 1.0\}$, with 3 random seeds per graph.

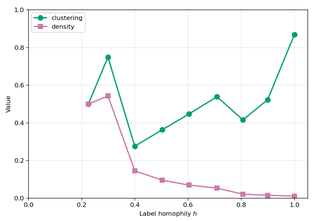

Figure 1: Density and clustering vs h.

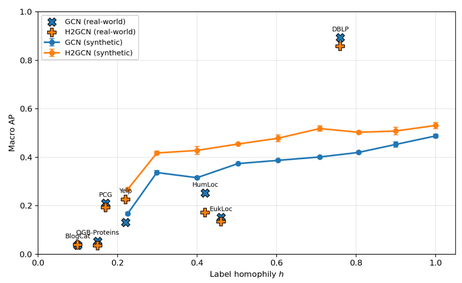

Figure 2: Macro AP vs h with real-world overlay.

Both models improve monotonically with h. The H2GCN–GCN gap stays in a narrow +0.04 to +0.12 band across the sweep. Density spans 50× and is anti-correlated with h (SDA structurally couples them); clustering is non-monotonic and only weakly tied to AP.

Results: Label Imbalance and Unlabeled Nodes

Label imbalance ratio:

$$\text{MeanIR} = \frac{1}{|C|} \sum_{k=1}^{|C|} \frac{\max_j \text{count}(j)}{\text{count}(k)},$$

the mean over labels of how many times rarer each label is than the most common one (MeanIR = 1 is perfect balance).

Condition	MeanIR	unlabeled	GCN AP	H2GCN AP
balanced	1.25	1.6%	0.598	0.705
mild skew	2.21	1.3%	0.553	0.672
strong skew	2.95	1.2%	0.520	0.650
+ 20% unlabeled	2.93	20.9%	0.497	0.516
+ 40% unlabeled	3.00	40.6%	0.441	0.453

Table 2: Macro AP across MeanIR and unlabeled-fraction conditions at $h \approx 0.4$.

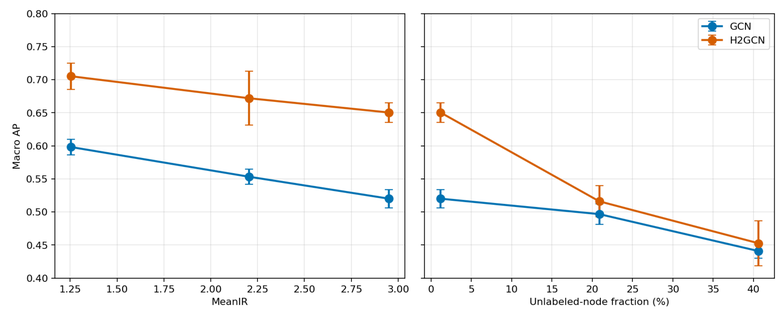

Figure 3: Macro AP across the two label-side axes. Left: skew preserves the H2GCN–GCN gap. Right: unlabeled supervision collapses it.

Skew preserves the H2GCN advantage (+0.11 → +0.13): the ego term protects rare-label signal that GCN's averaging dilutes. Unlabeled nodes collapse it (+0.13 → +0.01 at 40%): H2GCN's larger parameterisation needs more supervised neighbors.

Results: Pooled Ridge Regression

Pooled fit over 97 trained graphs (sweeps, gap-fillers, real-world): standardised Ridge coefficients with 95% bootstrap CIs for both macro AP and macro F_1.

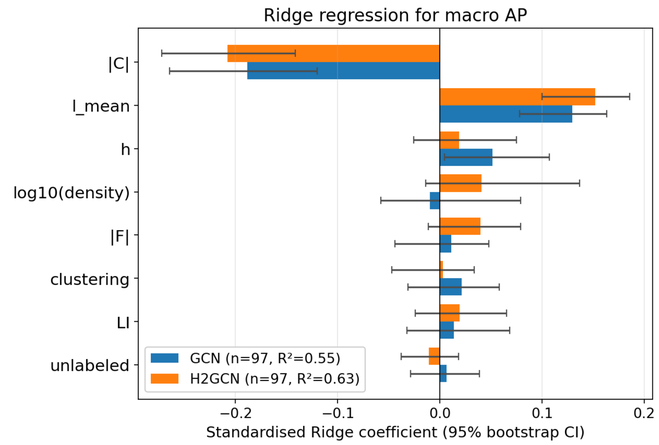

Figure 4: Ridge coefficients for macro AP.

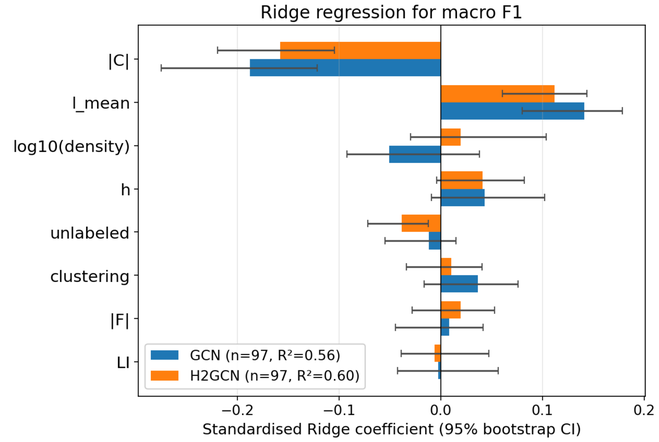

Figure 5: Ridge coefficients for macro F_1

$|C|$ and $\bar{\ell}$ dominate every cell; AP and F_1 agree closely, so the findings are robust to the metric. The model-side asymmetries match each architecture: GCN's averaging assumes neighbour agreement, so it pays for low h; H2GCN's richer parameterisation needs labels, so it pays for sparse supervision. Label-side properties weigh as much as graph-side ones.

Discussion and Conclusions

No single scalar property predicts GNN performance on multi-label graphs.

- Homophily explains most variance in a controlled sweep, but the H2GCN–GCN gap stays narrow (+0.04 to +0.12).
- Per-label imbalance preserves the gap; unlabeled supervision collapses it.
- Pooled Ridge ($n = 97$): $|C|$ and $\bar{\ell}$ dominate, h is significant for GCN but not H2GCN, unlabeled hurts H2GCN more than GCN.