Layer Embedding Deep Fusion Graph Neural Network

Supplementary Material

6. Attention-based Fusion Method

To further validate our claim that attention-based layer fusion inherently suffers from weight collapse, especially on heterophilic graphs, we conduct a detailed empirical analysis of attention weights (averaging all nodes) distributions across different propagation depths.

6.1. Experimental Setup

We evaluate attention-based multi-layer fusion on four representative datasets, using two backbone predictors: MLP (no message passing) and GCN (involving message passing). To demonstrate the advantage of the reconstructed topology over the original topology in weight collapse, the experimental results on the two topologies are compared side by side for each predictor.

In particular, we select two homophilic datasets (Cora and ACM) and two heterophilic datasets (Wisconsin and BlogCatalog), ensuring that the evaluation covers both structural regimes. This enables us to comprehensively examine attention weights distribution under different homophily settings.

6.2. Experiment Result and Analysis

The distribution of attention weights across the layers of the MLP predictor for two different topologies can be seen in Figure 6, and that of the GCN predictor for two different topologies can be seen in Figure 7.

More intuitively, the proportions of shallow (0-4) and deep (5-9) layer embeddings in final fused representation of the MLP predictor for two different topologies are compared, as shown in Figure 8, and that of the GCN predictor for two different topologies are compared in Figure 9.

The results clearly demonstrate that heterophilic graphs exhibit significantly stronger attention concentration in the shallow layer compared to homophilic graphs (e.g., the attention weight distribution of BlogCatalog dataset obtained by MLP using the original topology in Figure 6). Across all used heterophilic datasets, the attention weights assigned to the initial propagation layers dominate the final fused representation, confirming that attention-based convex fusion tends to collapse toward shallow semantics when deep embeddings are corrupted by structural noise.

When switching from the original topology to the reconstructed topology, we observe a consistent alleviation of shallow-layer collapse in heterophilic graphs (e.g., the attention weight distribution of BlogCatalog dataset obtained by MLP using the reconstructed topology in Figure 6). The reconstructed topology reduces cross-class noise and preserves the semantic validity of deeper embeddings, allowing attention to be more evenly distributed across layers. This verifies that the reconstructed topology plays the intended role of mitigating misaggregation in heterophilic settings.

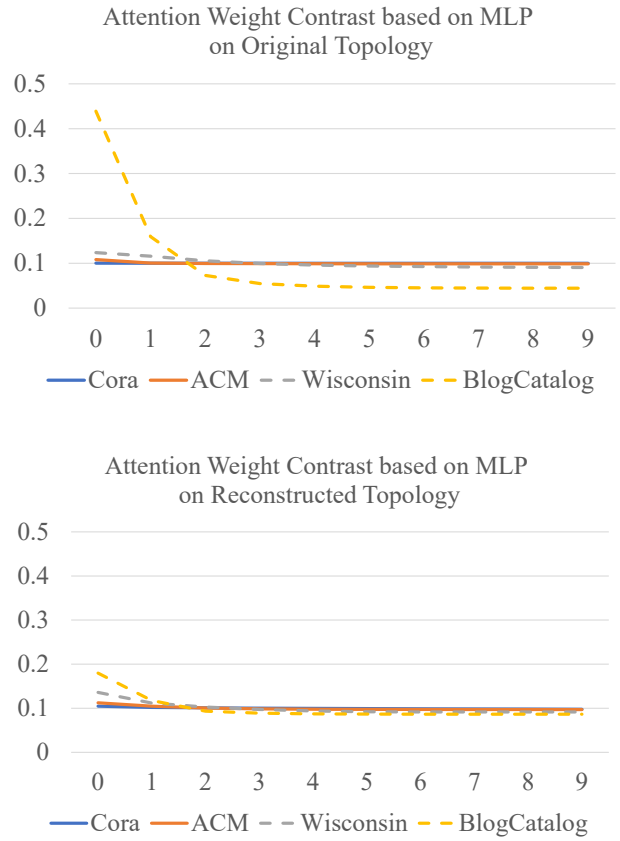

Figure 6. The distribution of attention weights across the layers of the MLP predictor for two different topologies. y-axis is attention wights, x-axis is the massage passing rounds.

Interestingly, on homophilic graphs, the reconstructed topology tends to make attention more biased toward shallow layers (e.g., the attention weight distribution of Cora dataset obtained by GCN using the reconstructed topology in Figure 9). This is because the additional edges increase local density and accelerate smoothing, making deeper embeddings less informative. Importantly, this is not detrimental to our model: the DTPS strategy in LEDF-GNN automatically adjusts the node-wise fusion weights α and β, suppressing the reconstructed topology when it becomes overly smoothing. This demonstrates that DTPS is essential for adapting to unknown and diverse homophily conditions, preventing negative propagation while retaining the benefits

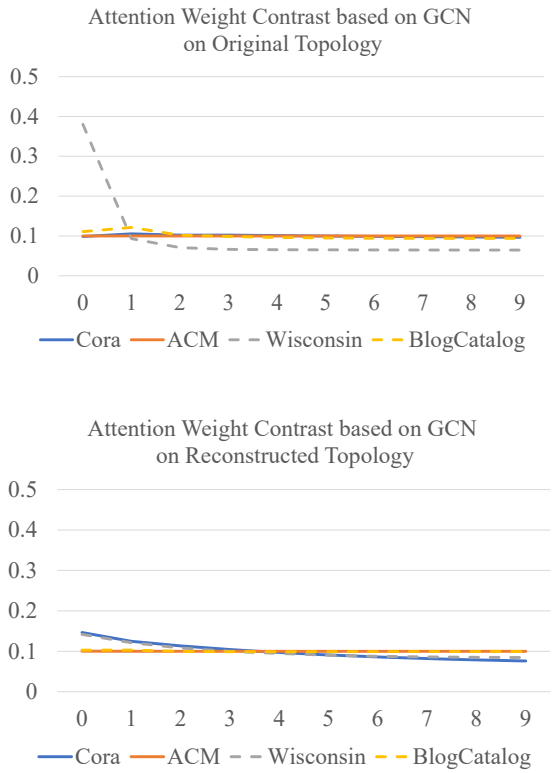

Figure 7. The distribution of attention weights across the layers of the GCN predictor for two different topologies. y-axis is attention wights, x-axis is the massage passing rounds.

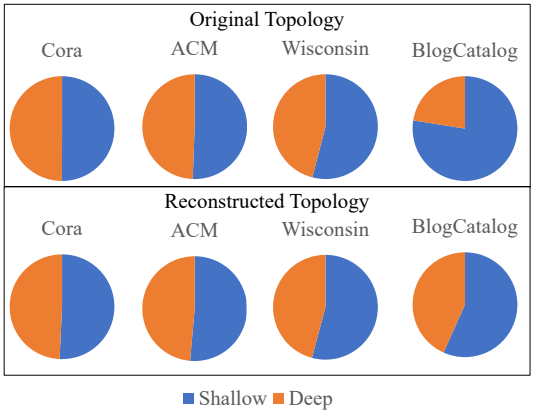

Figure 8. The proportion of attention weights in shallow and deep layers of the MLP predictor for two different topologies.

of reconstructed topology where needed.

7. The Details of Dataset

In total, fourteen datasets are used in this work, with each being fixedly divided. The details are provided in Table 3.

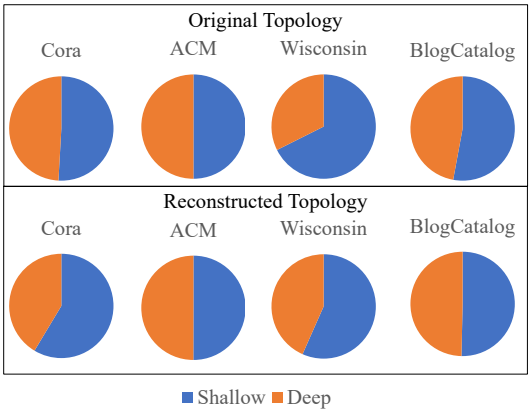

Figure 9. The proportion of attention weights in shallow and deep layers of the GCN predictor for two different topologies.

8. Parameter Setup

Table 4 introduces the parameter settings for the semi-supervised node classification experiments, including the k parameter of Top_k in LSC, as well as the propagation depths $Q1$ and $Q2$ used in the LEDF operator. Note that for both the ablation and validation experiments, the parameter settings are identical to those used in the semi-supervised node classification experiment.

9. Visualization on Semi-supervised Node Classification

Figure 10 illustrates the qualitative comparison on the Cora dataset, showcasing the visual differences among the backbones (MLP and GCN) themselves, backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup). Figure 11 illustrates the qualitative comparison on the ACM dataset, showcasing the visual differences among the backbones (MLP and GCN) themselves, backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup). Figure 12 illustrates the qualitative comparison on the BlogCatalog dataset, showcasing the visual differences among the backbones (MLP and GCN) themselves, backbones with baselines (LEDF-GNN, BORF and ComFy). In particular, due to the complexity of AGMixup, we cannot get the result of AGMixup classification on the BlogCatalog dataset. Figure 13 illustrates the qualitative comparison on the Wisconsin dataset, showcasing the visual differences among the backbones (MLP and GCN) themselves, backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup).

10. LSC $vs.$ Cosine Similarity

Table 5 shows that the more intuitive LSC can obtain better homophily than cosine similarity. Bit operations based LSC needs lower memory overhead than floating-point co-

Dataset	Nodes	Edges	Features	Class	Homo.	Train/Valid/Test
Cora	2708	10556	1433	7	0.81	140/500/1000
CiteSeer	3327	9104	3703	6	0.74	120/500/1000
PubMed	19717	88648	500	3	0.80	60/500/1000
ACM	3025	26256	1870	3	0.82	60/600/1200
Coauthor CS	18333	163788	6805	15	0.81	300/500/1000
Arxiv2023	46198	78548	300	40	0.65	27718/9239/9239
MNIST	2000	26588	30	10	0.87	200/300/1500
Arxiv	169343	2315598	128	40	0.65	90941/29799/48603
BlogCatalog	5196	343486	8189	6	0.40	120/500/1000
Texas	183	325	1703	5	0.11	10/50/100
Wisconsin	251	515	1703	5	0.21	10/50/100
Cornell	183	298	1703	5	0.30	10/50/100
Chameleon	2277	36101	2325	5	0.23	114/500/1000
Squirrel	5201	217073	2089	5	0.22	260/500/1000

Table 3. The details of fourteen datasets used in this paper. 'Homo.' refers to the homophily ratio, which is the proportion of homophily edges relative to the total number of edges.

Dataset	k	MLP		GCN		GAT		GIN		APPNP	
		Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2
Cora	2	10	3	15	11	3	5	3	10	7	10
CiteSeer	3	10	9	6	20	8	1	3	10	2	7
PubMed	2	2	7	7	10	5	10	6	6	3	9
ACM	7	1	9	7	18	7	9	7	7	1	10
Coauthor CS	10	3	10	10	10	1	3	2	4	10	1
Arxiv2023	10	7	8	3	9	5	2	10	4	16	14
MNIST	10	7	10	6	9	9	7	4	9	5	5
Arxiv	10	9	6	4	6	10	4	6	6	5	8
BlogCatalog	20	3	8	16	14	1	10	5	9	10	9
Texas	20	3	8	8	16	8	20	6	8	7	4
Wisconsin	20	2	10	11	15	4	5	4	20	1	4
Cornell	20	9	7	18	2	9	8	16	10	1	17
chameleon	40	7	3	10	16	19	2	3	7	20	13
Squirrel	40	2	10	10	18	7	9	5	9	10	3

Table 4. The parameter settings of LEDF-GNN are unified across all backbones and used consistently in the semi-supervised node classification, ablation, and validation experiments.

sine similarity within comparable time.

11. Complexity Analysis

The time and space complexity of our method beyond the backbone model are $O(cm + n)$ and $O(cn)$, respectively, where n is node number, m is edge number and c is class number. Accordingly, the runtime and memory overhead reported in Tables 6 also exhibit that our model only needs lightweight overhead beyond the backbone.

12. Hyperparameter Analysis

We conduct sensitivity analysis on two key hyperparameters, depth Q and reconstruction hyperparameter K, by using the heterophilic Wisconsin dataset. Figure 14 shows the robustness of our method across different settings.

13. Comparison with Heterophliy-focused and Deep-GNN Baseline

The comparison results are reported in Table 7. JKNet-Mean focus on layer fusion, GCNII and NDLS are deep-GNNs, H_2GCN is the representative heterophily-focused method. The datasets of Wisconsin and Chameleon are heterophilic.

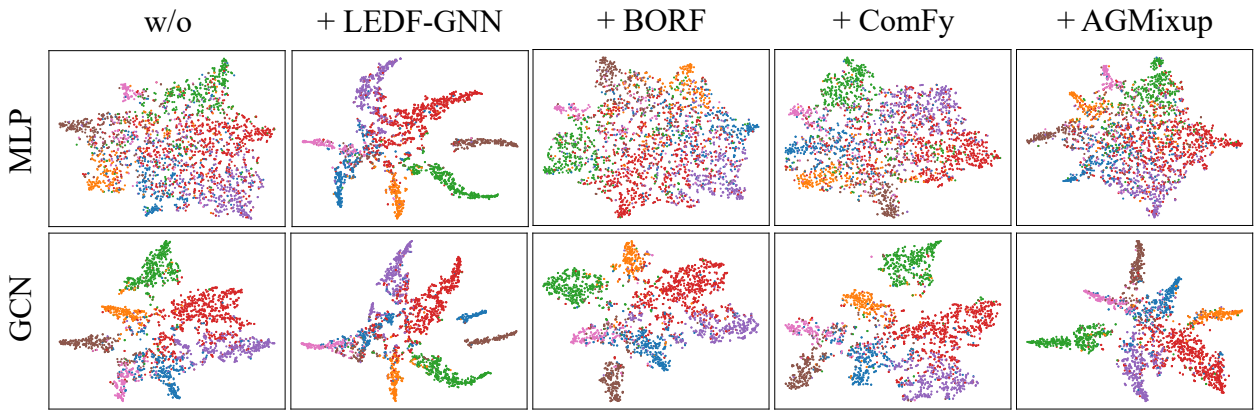

Figure 10. The visualization comparison of the backbones (MLP and GCN) and the backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup) on the **Cora** dataset.

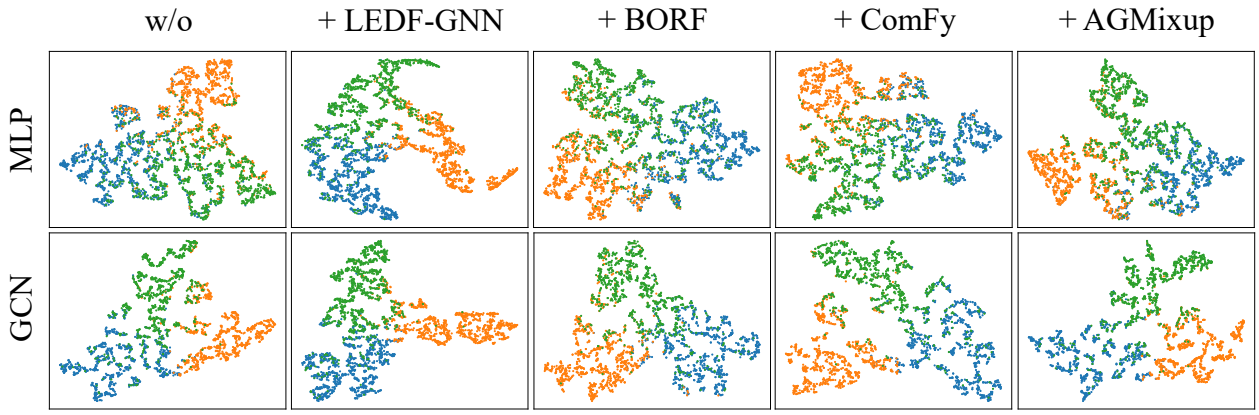

Figure 11. The visualization comparison of the backbones (MLP and GCN) and the backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup) on the **ACM** dataset.

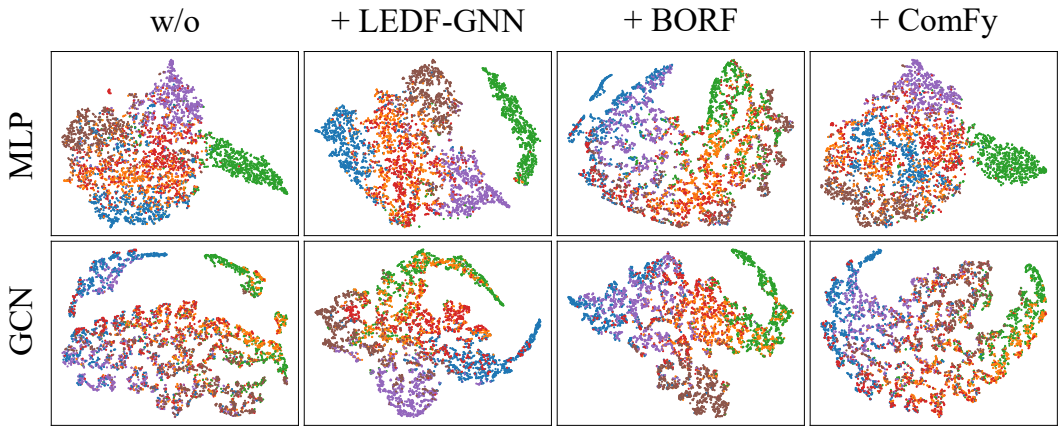

Figure 12. The visualization comparison of the backbones (MLP and GCN) and the backbones with baselines (LEDF-GNN, BORF and ComFy) on the **BlogCatalog** dataset.

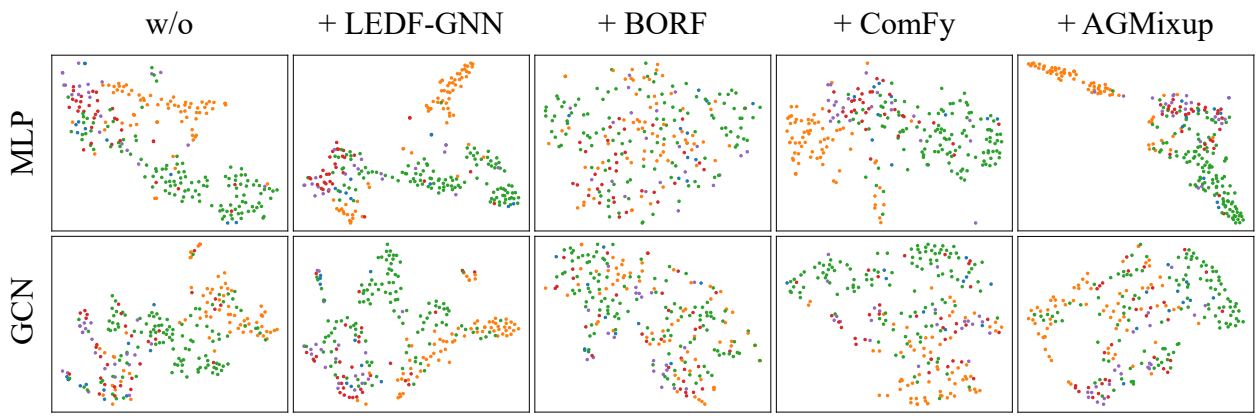

Figure 13. The visualization comparison of the backbones (MLP and GCN) and the backbones with baselines (LEDF-GNN, BORF, ComFy and AGMixup) on the **Wisconsin** dataset.

Dataset	Empty(%)		Rewire Origin(%)		Runtime(second)		Memory(MB)	
	Cosine	Logic	Cosine	Logic	Cosine	Logic	Cosine	Logic
Cora	61.8	**63.2**	71.1	**75.0**	0.1057	**0.0931**	93.93	**51.12**
CiteSeer	64.0	**64.7**	67.0	**69.5**	**0.1489**	0.3464	186.76	**97.52**
Chameleon	24.2	**30.0**	23.7	**28.4**	**0.1012**	0.1082	89.10	**48.90**
Wisconsin	63.9	**66.6**	60.5	**62.6**	0.09816	**0.0035**	11.88	**10.01**

Table 5. Edge homophily ratio(%) of topologies generated from an empty graph and topologies modified from the original graph based on cosine similarity and logical similarity.

Dataset	Runtime(Second)				Memory(MB)			
	GCN	BORF	AGM.	Ours	GCN	BORF	AGM.	Ours
Cora	0.0051	0.0119	0.1535	0.0124	49.35	49.44	351.62	50.36
CiteSeer	0.0058	0.0114	0.1298	0.0178	84.74	84.90	302.70	86.75
Chameleon	0.0061	0.0118	0.1446	0.0147	108.87	109.56	9574.70	111.88
Wisconsin	0.0057	0.0141	0.0234	0.0133	22.08	22.09	66.85	22.68

Table 6. The runtime (Second) and Memory (MB) of backbone GCN, baselines and LEDF-GNN with backbone.

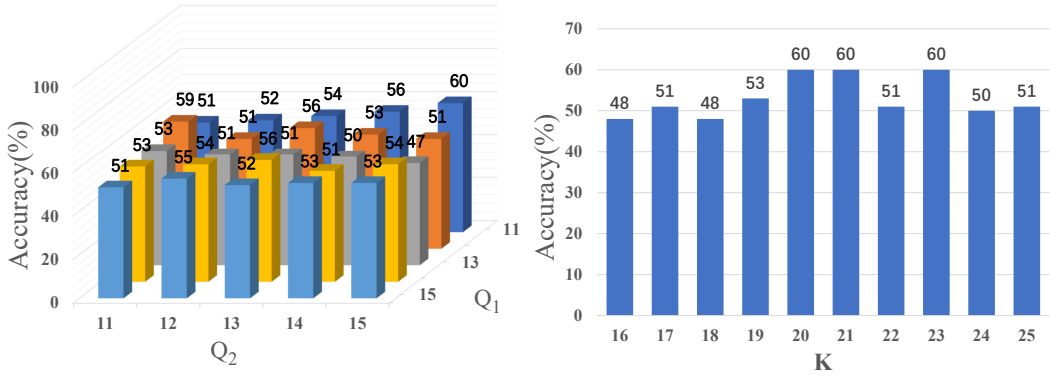

Figure 14. Sensitivity Analysis of two hyperparameters (Q and K) on heterophilic Wisconsin dataset, one changes the other is fixed.

Model	GCN	JKNet-Mean	GCNII	NDLS	H_2GCN	LEDF-GCN
Cora	82.4	67.0	**84.3**	83.5	75.5	<u>84.0</u>
CiteSeer	70.7	52.2	<u>73.1</u>	71.5	64.1	**74.8**
PubMed	79.2	75.2	<u>79.7</u>	79.2	77.2	**82.1**
ACM	82.1	76.5	<u>82.3</u>	82.2	73.7	**87.5**
BlogCatalog	42.0	49.8	20.4	<u>71.9</u>	49.0	**79.6**
Texas	55.0	54.0	55.0	<u>63.0</u>	55.0	**66.0**
Wisconsin	48.0	57.0	55.0	<u>58.0</u>	53.0	**60.0**
Cornell	<u>49.0</u>	**55.0**	47.0	**55.0**	**55.0**	**55.0**
Chameleon	49.6	40.7	38.9	41.9	<u>50.4</u>	**53.2**
Squirrel	<u>33.7</u>	32.7	27.3	24.6	26.3	**39.0**

Table 7. Node classification accuracy (%) of our model and baselines, **bold** is the best, <u>underline</u> is the second-best.