Beyond Random Masking: When Dropout Meets GCNs

Yuankai Luo, Xiao-Ming Wu, Hao Zhu

北京航空航天大學
BEIHANG UNIVERSITY

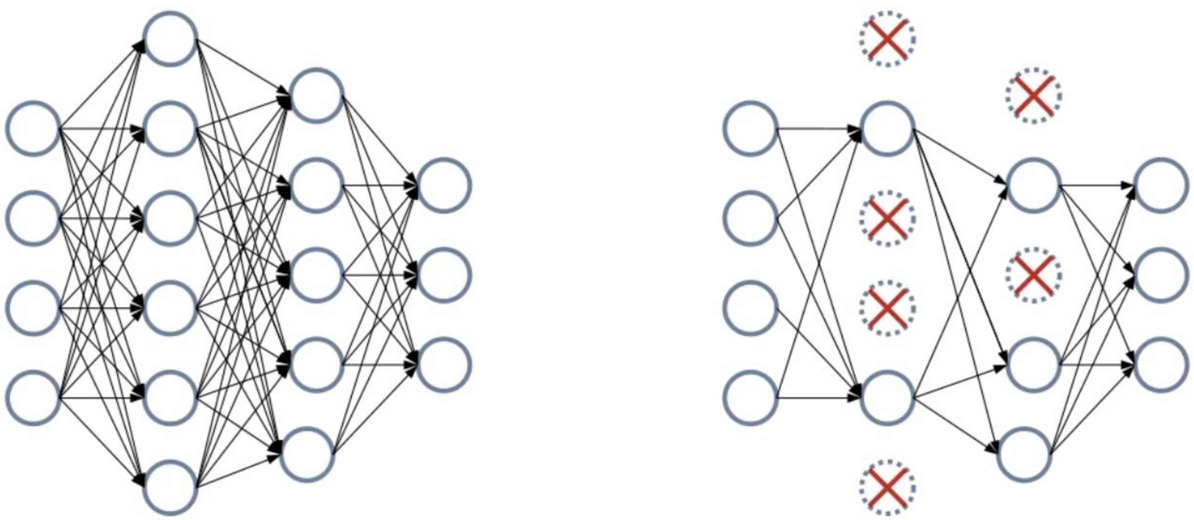
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

CSIRO
DATA 61

ICLR 2025

Motivation

- Dropout is widely used in deep learning.
- In GCNs, its mechanism and effect are still unclear. How significant is the role of it?
- Does it prevent co-adaptation like in MLPs?



Dropout in GCNs: Key Insights

- **Dimension-Specific Sub-Graphs**: Creates stochastic sub-graphs, enabling structural regularization unique to GCNs.
- **Degree-Dependent Effects**: Adaptive regularization based on node connectivity, emphasizing topological importance.
- **Over-smoothing Mitigation**: Primarily mitigates over-smoothing, extending beyond coadaptation, with nuanced effects.
- **Generalization Bounds**: Dependent on graph properties, diverging from traditional dropout theory.
- **Synergy with BatchNorm**: Dropout + BatchNorm achieves SOTA performance on many datasets.

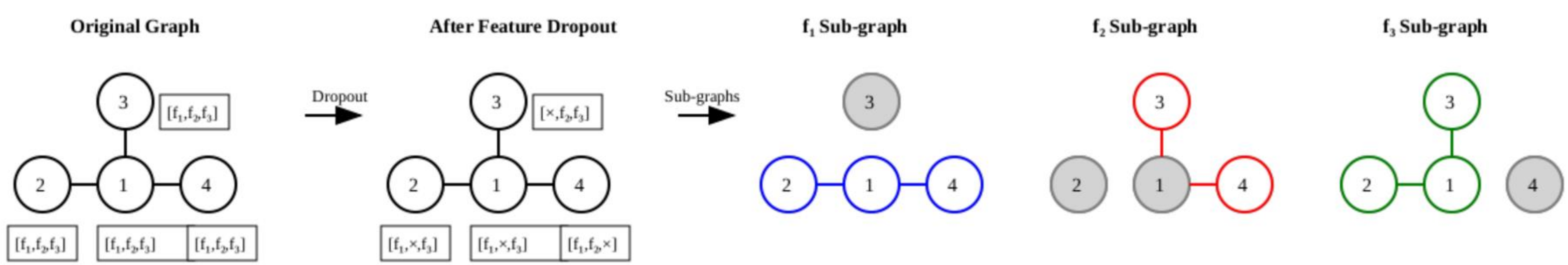
Theoretical Contributions

Dropout definitions:

$$\boldsymbol{H}^{(l)} = \frac{1}{1-p}\boldsymbol{M}^{(l)} \odot \sigma(\tilde{\boldsymbol{A}}\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)})$$

Dimension-specific stochastic sub-graphs:

$$\mathcal{E}_t^{(l,j)} = \{(u,v) \in \mathcal{E} \mid M_{uj}^{(l,t)} \neq 0 \text{ and } M_{vj}^{(l,t)} \neq 0\}.$$



Theorem 1 (Sub-graph Diversity). *The expected number of distinct sub-graphs per iteration is:*

$$\mathbb{E}[|\mathcal{E}_t^{(l,j)} \mid j = 1, \ldots, d_l|] = d_l(1-(1-p)^{2|\mathcal{E}|}),$$

Theoretical Contributions

Theorem 3 (Degree-Dependent Dropout Effect). *The expected effective degree and its variance are given by:*

$$\mathbb{E}[deg_{i,t}^{eff}] = (1-p)^2 deg_i \quad and \quad Var[deg_{i,t}^{eff}] = deg_i(1-p)^2(1-(1-p)^2), \tag{8}$$

where deg_i is the original degree of node i and p is the dropout probability.

Corollary 4 (Relative Stability of High-Degree Nodes). *The coefficient of variation of the effective degree, defined as $CV[deg_{i,t}^{eff}] = \sqrt{Var[deg_{i,t}^{eff}]/\mathbb{E}[deg_{i,t}^{eff}]}$, decreases with increasing node degree:*

$$CV[deg_{i,t}^{eff}] = \frac{\sqrt{1-(1-p)^2}}{\sqrt{deg_i}(1-p)}.$$

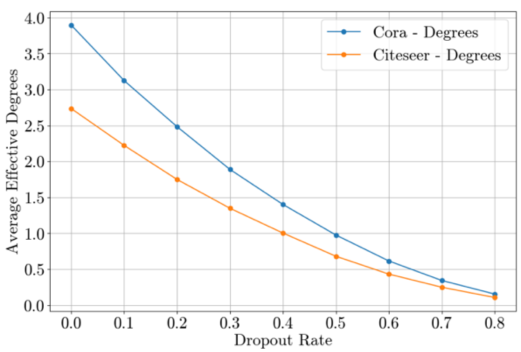


Figure 9: Effective degree.

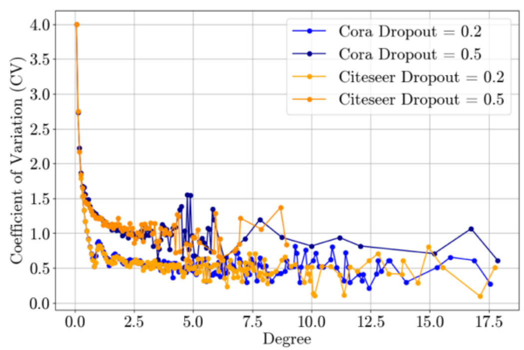


Figure 10: Effective CV vs degree.

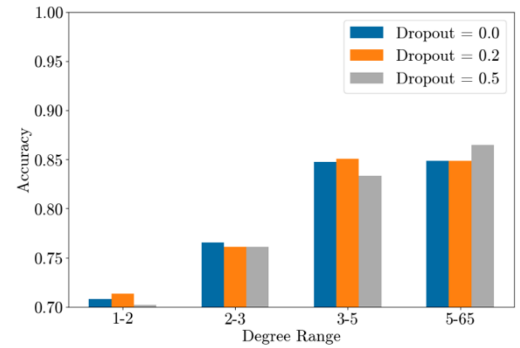


Figure 11: Accuracy on Cora.

Theoretical Contributions

Theorem 5 (Dropout and Feature Energy). *For a GCN with dropout probability p, the expected feature energy at layer l is bounded by:*

$$\mathbb{E}[E(\boldsymbol{H}^{(l)})] \leq \frac{deg_{\max}}{|\mathcal{E}|}\left(\frac{1}{1-p}\right)^{l} \|\tilde{\boldsymbol{A}}\|_2^{2l} \prod_{i=1}^{l} \|\boldsymbol{W}^{(i)}\|_2^2 \|\boldsymbol{X}\|_F^2 \tag{9}$$

where $E(\boldsymbol{X})$ is the energy of the input features and $\boldsymbol{W}^{(i)}$ are the weight matrices (The complete proof is in the Appendix.A.2).

$$\boxed{E(\boldsymbol{H}^{(l)}) = \frac{1}{2|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \|\boldsymbol{h}_i^{(l)} - \boldsymbol{h}_j^{(l)}\|_2^2}$$

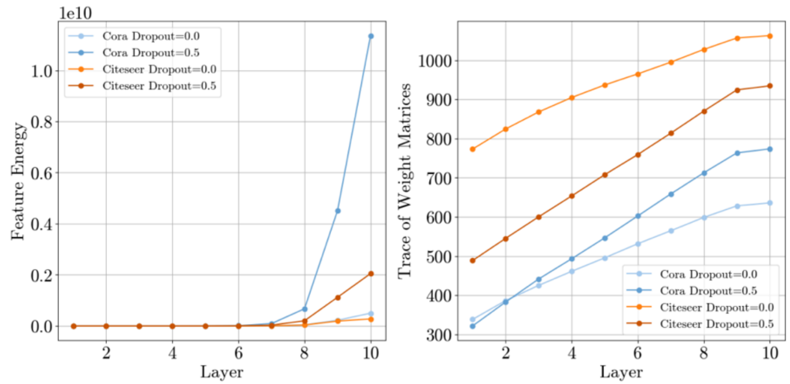


Figure 2: Feature energy vs dropout rates.

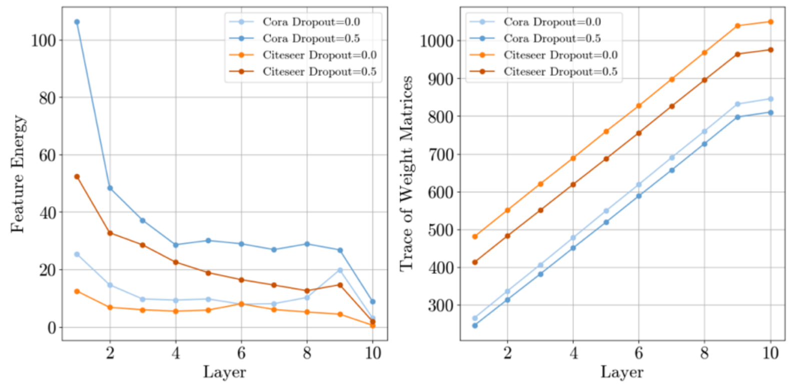


Figure 3: BN feature energy vs dropout rates.

Theoretical Contributions

Theorem 5 (Dropout and Feature Energy). *For a GCN with dropout probability p, the expected feature energy at layer l is bounded by:*

$$\mathbb{E}[E(\boldsymbol{H}^{(l)})] \leq \frac{deg_{\max}}{|\mathcal{E}|}(\frac{1}{1-p})^l \|\tilde{\boldsymbol{A}}\|_2^{2l} \prod_{i=1}^{l} \|\boldsymbol{W}^{(i)}\|_2^2 \|\boldsymbol{X}\|_F^2 \tag{9}$$

where $E(\boldsymbol{X})$ is the energy of the input features and $\boldsymbol{W}^{(i)}$ are the weight matrices (The complete proof is in the Appendix.A.2).

Primarily mitigates over-smoothing

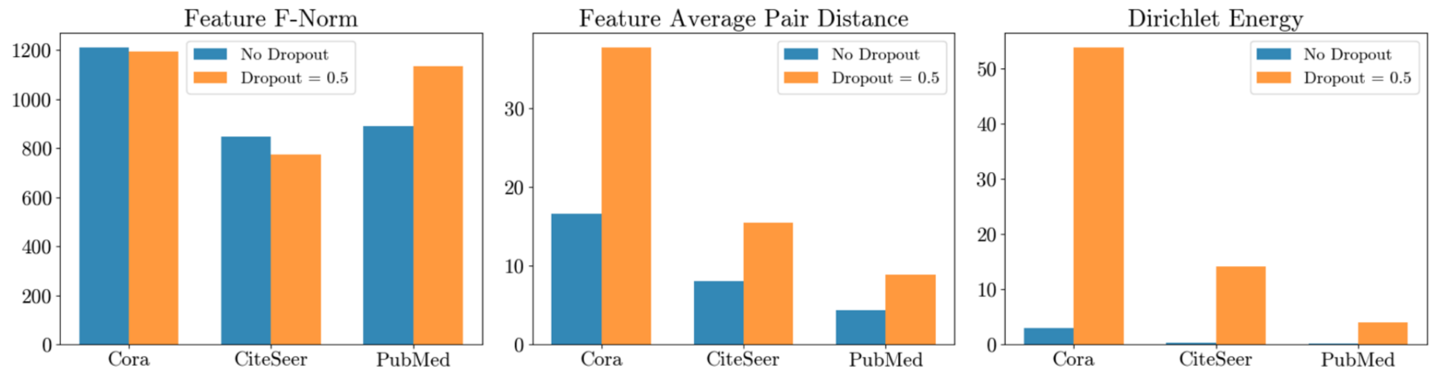


Figure 4: Effect of dropout on feature F-norm, average pair distance, and Dirichlet energy.

Theoretical Contributions

Theorem 6 (Generalization Bound for L-Layer GCN with Dropout)**.** *For an L-layer GCN F with dropout probability p_l at layer l and L_σ-Lipschitz activation function σ, with probability at least $1-\delta$ over the training examples, the following generalization bound holds:*

$$\mathbb{E}_D[L(F(X))] - \mathbb{E}_S[L(F(X))] \leq O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \sum_{l=1}^{L} L_{loss} \cdot L_l \cdot \sqrt{\frac{p_l}{(1-p_l)\boxed{\chi_f(G)}}} \|\sigma(\tilde{\boldsymbol{A}}\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)})\|_F, \tag{10}$$

where $\mathbb{E}_D$ is the expectation over the data distribution, $\mathbb{E}_S$ is the expectation over the training samples, L is the loss function with Lipschitz constant L_{loss}, $L_l = \prod_{i=l}^{L}(L_\sigma \|\boldsymbol{W}^{(i)}\|_2 \cdot \|\tilde{\boldsymbol{A}}\|_2)$ is the Lipschitz constant from layer l to output, $\|\boldsymbol{W}^{(i)}\|_2$ is the spectral norm of the weight matrix at layer i, $\|\tilde{\boldsymbol{A}}\|_2$ is the spectral norm of the normalized adjacency matrix, and $\chi_f(G)$ is the fractional chromatic number of the dependency graph G induced by the message passing structure.

Dropout & Batch Normalization

- Dropout provides stochastic sparsity.
- BN preserves minimum feature energy.
- Together, they ensure both diversity and stability in GCNs.

Experiments (Node-Level)

- 10 datasets: Cora, PubMed, WikiCS, ogbn-arxiv, ogbn-products, etc.
- GCN + Dropout + BN outperforms SOTA GNNs
- Dropout boosts Dirichlet energy, preventing oversmoothing

Table 2: Node classification results (%). The baseline results are taken from Deng et al. (2024); Wu et al. (2023). The top 1st, 2nd and 3rd results are highlighted. "dp" denotes dropout.

	Cora	CiteSeer	PubMed	Computer	Photo	CS	Physics	WikiCS	ogbn-arxiv	ogbn-products
# nodes	2,708	3,327	19,717	13,752	7,650	18,333	34,493	11,701	169,343	2,449,029
# edges	5,278	4,732	44,324	245,861	119,081	81,894	247,962	216,123	1,166,243	61,859,140
Metric	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑	Accuracy↑
GCNII	85.19 ± 0.26	73.20 ± 0.83	80.32 ± 0.44	91.04 ± 0.41	94.30 ± 0.20	92.22 ± 0.14	95.97 ± 0.11	78.68 ± 0.55	72.74 ± 0.31	79.42 ± 0.36
GPRGNN	83.17 ± 0.78	71.86 ± 0.67	79.75 ± 0.38	89.32 ± 0.29	94.49 ± 0.14	95.13 ± 0.09	96.85 ± 0.08	78.12 ± 0.23	71.10 ± 0.12	79.76 ± 0.59
APPNP	83.32 ± 0.55	71.78 ± 0.46	80.14 ± 0.22	90.18 ± 0.17	94.32 ± 0.14	94.49 ± 0.07	96.54 ± 0.07	78.87 ± 0.11	72.34 ± 0.24	78.84 ± 0.09
tGNN	82.97 ± 0.68	71.74 ± 0.49	80.67 ± 0.34	83.40 ± 1.33	89.92 ± 0.72	92.85 ± 0.48	96.24 ± 0.24	71.49 ± 1.05	72.88 ± 0.26	81.79 ± 0.54
GraphGPS	82.84 ± 1.03	72.73 ± 1.23	79.94 ± 0.26	91.19 ± 0.54	95.06 ± 0.13	93.93 ± 0.12	97.12 ± 0.19	78.66 ± 0.49	70.97 ± 0.41	OOM
NAGphormer	82.12 ± 1.18	71.47 ± 1.30	79.73 ± 0.28	91.22 ± 0.14	95.49 ± 0.11	95.75 ± 0.09	97.34 ± 0.03	77.16 ± 0.72	70.13 ± 0.55	73.55 ± 0.21
Exphormer	82.77 ± 1.38	71.63 ± 1.19	79.46 ± 0.35	91.47 ± 0.17	95.35 ± 0.22	94.93 ± 0.01	96.89 ± 0.09	78.54 ± 0.49	72.44 ± 0.28	OOM
GOAT	83.18 ± 1.27	71.99 ± 1.26	79.13 ± 0.38	90.96 ± 0.90	92.96 ± 1.48	94.21 ± 0.38	96.24 ± 0.24	77.00 ± 0.77	72.41 ± 0.40	82.00 ± 0.43
NodeFormer	82.20 ± 0.90	72.50 ± 1.10	79.90 ± 1.00	86.98 ± 0.62	93.46 ± 0.35	95.64 ± 0.22	96.45 ± 0.28	74.73 ± 0.94	59.90 ± 0.42	73.96 ± 0.30
SGFormer	84.50 ± 0.80	72.60 ± 0.20	80.30 ± 0.60	92.42 ± 0.66	95.58 ± 0.36	95.71 ± 0.24	96.75 ± 0.26	80.05 ± 0.46	72.63 ± 0.13	81.54 ± 0.43
Polynormer	83.25 ± 0.93	72.31 ± 0.78	79.24 ± 0.43	93.68 ± 0.21	96.46 ± 0.26	95.53 ± 0.16	97.27 ± 0.08	80.10 ± 0.67	73.46 ± 0.16	83.82 ± 0.11
GCN	85.22 ± 0.66	73.24 ± 0.63	81.08 ± 1.16	93.15 ± 0.34	95.03 ± 0.24	94.41 ± 0.13	97.07 ± 0.04	80.14 ± 0.52	73.13 ± 0.27	81.87 ± 0.41
Dirichlet energy	74.671	9.934	4.452	8.020	3.765	20.241	8.966	6.109	8.021	7.771
GCN w/o dp	83.18 ± 1.22	70.48 ± 0.45	79.40 ± 1.02	90.60 ± 0.84	94.10 ± 0.15	94.30 ± 0.22	96.92 ± 0.05	77.61 ± 1.34	72.05 ± 0.23	77.50 ± 0.37
Dirichlet energy	2.951	0.170	0.247	0.592	1.793	3.980	0.318	1.592	1.231	1.745
GCN w/o BN	84.97 ± 0.73	72.97 ± 0.86	80.94 ± 0.87	92.39 ± 0.18	94.38 ± 0.13	93.46 ± 0.24	96.76 ± 0.06	79.00 ± 0.48	71.93 ± 0.18	79.37 ± 0.42
SAGE	84.14 ± 0.63	71.62 ± 0.29	77.86 ± 0.79	92.65 ± 0.21	95.71 ± 0.20	95.90 ± 0.09	97.20 ± 0.10	80.29 ± 0.97	72.72 ± 0.13	82.69 ± 0.28
SAGE w/o dp	83.06 ± 0.80	69.68 ± 0.82	76.40 ± 1.48	90.17 ± 0.60	94.90 ± 0.17	95.80 ± 0.08	97.06 ± 0.06	78.84 ± 1.17	71.37 ± 0.31	79.82 ± 0.22
SAGE w/o BN	83.89 ± 0.67	71.39 ± 0.75	77.26 ± 1.02	92.54 ± 0.24	95.51 ± 0.23	94.87 ± 0.15	97.03 ± 0.03	79.50 ± 0.93	71.52 ± 0.17	80.91 ± 0.35
GAT	83.92 ± 1.29	72.00 ± 0.91	80.48 ± 0.99	93.47 ± 0.27	95.53 ± 0.16	94.49 ± 0.17	96.73 ± 0.10	80.21 ± 0.68	72.83 ± 0.19	80.05 ± 0.34
GAT w/o dp	82.58 ± 1.47	71.08 ± 0.42	79.28 ± 0.58	92.94 ± 0.30	93.88 ± 0.16	94.30 ± 0.14	96.42 ± 0.08	78.67 ± 0.40	71.52 ± 0.41	77.87 ± 0.25
GAT w/o BN	83.76 ± 1.32	71.82 ± 0.83	80.43 ± 1.03	92.16 ± 0.26	95.05 ± 0.49	93.33 ± 0.26	96.57 ± 0.20	79.49 ± 0.62	71.68 ± 0.36	78.21 ± 0.32

Experiments (Graph-Level)

- Datasets: MNIST, CIFAR10, Peptides-func/struct
- GCN + Dropout + BN outperforms SOTA GNNs

Table 3: Graph classification results on two peptide datasets from LRGB (Dwivedi et al., 2022).

Model	Peptides-func	Peptides-struct
# graphs	15,535	15,535
Avg. # nodes	150.9	150.9
Avg. # edges	307.3	307.3
Metric	AP ↑	MAE ↓
GT	0.6326 ± 0.0126	0.2529 ± 0.0016
SAN+RWSE	0.6439 ± 0.0075	0.2545 ± 0.0012
GraphGPS	0.6535 ± 0.0041	0.2500 ± 0.0012
MGT+WavePE	0.6817 ± 0.0064	0.2453 ± 0.0025
DRew	0.7150 ± 0.0044	0.2536 ± 0.0015
Exphormer	0.6527 ± 0.0043	0.2481 ± 0.0007
Graph-MLPMixer	0.6970 ± 0.0080	0.2475 ± 0.0015
GRIT	0.6988 ± 0.0082	0.2460 ± 0.0012
CKGCN	0.6952 ± 0.0068	0.2477 ± 0.0019
GRED	0.7085 ± 0.0027	0.2503 ± 0.0019
Graph Mamba	0.6972 ± 0.0100	0.2477 ± 0.0019
GCN	0.7015 ± 0.0021	**0.2437** ± 0.0012
Dirichlet energy	9.649	6.121
GCN w/o dp	0.6484 ± 0.0034	0.2541 ± 0.0026
Dirichlet energy	6.488	3.725

Table 4: Graph classification results on two image datasets from (Dwivedi et al., 2023).

Model	MNIST	CIFAR10
# graphs	70,000	60,000
Avg. # nodes	70.6	117.6
Avg. # edges	564.5	941.1
Metric	Accuracy ↑	Accuracy ↑
GT	90.831 ± 0.161	59.753 ± 0.293
SAN+RWSE	-	-
GraphGPS	98.051 ± 0.126	72.298 ± 0.356
MGT+WavePE	-	-
DRew	-	-
Exphormer	98.550 ± 0.039	74.696 ± 0.125
Graph-MLPMixer	97.422 ± 0.110	73.961 ± 0.330
GRIT	98.108 ± 0.111	76.468 ± 0.881
CKGCN	98.423 ± 0.155	72.785 ± 0.436
GRED	98.383 ± 0.012	76.853 ± 0.185
Graph Mamba	98.392 ± 0.183	74.563 ± 0.379
GatedGCN	**98.684** ± 0.137	**76.931** ± 0.367
Dirichlet energy	1.119	1.541
GatedGCN w/o dp	98.235 ± 0.136	71.384 ± 0.397
Dirichlet energy	0.987	0.845

Comparison with Drop Variants

- Compared methods: DropEdge, DropNode, DropMessage
- Standard dropout achieves stronger and more consistent performance

Table 1: Comparison of different dropout variants in GNNs. Each method is characterized by its masking operation $\boldsymbol{M}_d$, the resulting sub-graph formation $\mathcal{G}_t$, and expected effective degree $\mathbb{E}[deg_{i,t}^{\text{eff}}]$, where p is the dropout probability.

Method	Masking Operation	Sub-graph Formation	Expected Effective Degree
DropNode	$\boldsymbol{M}_d = \tilde{\boldsymbol{A}}((\boldsymbol{M}_{node} \odot \boldsymbol{H}^{(l-1)})\boldsymbol{W}^{(l)})_d$	$\mathcal{G}_t = (\mathcal{V} \setminus \mathcal{V}_{dropped}, \mathcal{E} \setminus \{(i,j) \mid i \in \mathcal{V}_{dropped}\})$	$deg_i \prod_{j \in \mathcal{N}(i)}(1-p)$
DropEdge	$\boldsymbol{M}_d = (\boldsymbol{M}_{edge} \odot \tilde{\boldsymbol{A}})(\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)})_d$	$\mathcal{G}_t = (\mathcal{V}, \mathcal{E} \setminus \mathcal{E}_{dropped})$	$(1-p)deg_i$
DropMessage	$\boldsymbol{M}_d = \tilde{\boldsymbol{A}}(\boldsymbol{M}_{msg_d} \odot (\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)}))_d$	$\mathcal{G}_t^d = (\mathcal{V}, \{(i,j) \in \mathcal{E} \mid \boldsymbol{M}_{msg_{d_{ij}}} \neq 0\})$	$(1-p)deg_i$
Dropout	$\boldsymbol{M}_d = \boldsymbol{M}_{feat_d} \odot \tilde{\boldsymbol{A}}(\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)})_d$	$\mathcal{G}_t^d = (\mathcal{V}, \{(i,j) \in \mathcal{E} \mid \boldsymbol{M}_{feat_{d_i}} \neq 0, \boldsymbol{M}_{feat_{d_j}} \neq 0\})$	$(1-p)^2 deg_i$

Table 5: Experimental results of different regularization methods on Cora, Citeseer, and PubMed.

	Cora (GCN)	**CiteSeer (GCN)**	**PubMed (GCN)**	**Cora (SAGE)**	**CiteSeer (SAGE)**	**PubMed (SAGE)**	**Cora (GAT)**	**CiteSeer (GAT)**	**PubMed (GAT)**
GNN	83.18 ± 1.22	70.48 ± 0.45	79.40 ± 1.02	83.06 ± 0.80	69.68 ± 0.82	76.40 ± 1.48	82.58 ± 1.47	71.08 ± 0.42	79.28 ± 0.58
GNN+Dropout	**85.22 ± 0.66**	**73.24 ± 0.63**	**81.08 ± 1.16**	**84.14 ± 0.63**	71.62 ± 0.29	77.86 ± 0.79	**83.92 ± 1.29**	**72.00 ± 0.91**	**80.48 ± 0.99**
GNN+DropEdge	84.88 ± 0.68	72.96 ± 0.38	80.42 ± 1.15	83.10 ± 0.51	71.72 ± 0.92	77.88 ± 1.31	83.44 ± 0.78	71.60 ± 1.14	79.82 ± 0.68
GNN+DropNode	84.92 ± 0.52	73.08 ± 0.39	80.60 ± 0.49	83.42 ± 0.58	**71.92 ± 0.65**	78.06 ± 1.09	83.80 ± 0.97	71.30 ± 0.87	79.50 ± 0.68
GNN+DropMessage	84.78 ± 0.58	73.12 ± 1.19	80.92 ± 0.88	83.18 ± 0.62	71.22 ± 1.34	**78.20 ± 0.80**	83.46 ± 1.06	71.38 ± 1.12	79.36 ± 1.22

Conclusion

- Dropout in GCNs: structural, degree-aware, anti-oversmoothing
- Use dropout + BN in GNNs to get SOTA results

Thank You

- Acknowledgments
- Code: https://github.com/LUOyk1999/dropout-theory