Appendix A. Mathematical Details

In this part, we provide the proofs for the Propositions in the main paper.

A.1. Proof of Proposition 1.

Let $G = (V, E)$ be a graph with each node v_i having two neighbors, denoted v_i^1 and v_i^2. Further, let all node features be initialized such that both neighbors start with the same state $x_i^1 = x_i^2$. In f, all nodes x_i^1 are weighted by p, and all nodes x_i^2 are weighted by ϵ. In g, weights are exchanged. Thus, results for f, g are the same and $C(f, g) = 0$. However, the influence difference is large, precisely for $p > 3\epsilon$, it is

$$\mathrm{ID}(f, g) = \frac{p - \epsilon}{0.25 \cdot (p + \epsilon)} > 2\,. \tag{1}$$

□

A.2. Proof of Proposition 2.

Our proof closely follows the proof for random noise by Srinivas et al. We use the Taylor-approximation of T and S around the point $(\mathbf{X}, \mathbf{A})$ and use our assumption about zero mean for each entry of $\mathbf{A}_{\text{drop}}$.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}_{\text{drop}}}\left[\sum_{v,i=1}^{N,C}\left(T(\mathbf{X}, \mathbf{A} + \mathbf{A}_{\text{drop}})_{vi} - S(\mathbf{X}, \mathbf{A} + \mathbf{A}_{\text{drop}})_{vi}\right)^2\right] \\
&= \mathbb{E}_{\mathbf{A}_{\text{drop}}}\Bigg[\sum_{v,i=1}^{N,C}\Big(T(\mathbf{X}, \mathbf{A})_{vi} + vec(\nabla_{\mathbf{A}} T(\mathbf{X}, \mathbf{A}))^T vec(\mathbf{A}_{\text{drop}})) + \mathcal{O}(vec(\mathbf{A}_{\text{drop}})) \odot vec(\mathbf{A}_{\text{drop}}))\Big) \\
&- S(\mathbf{X}, \mathbf{A})_{vi} + vec(\nabla_{\mathbf{A}} S(\mathbf{X}, \mathbf{A}))^T vec(\mathbf{A}_{\text{drop}})) + \mathcal{O}(vec(\mathbf{A}_{\text{drop}})) \odot vec(\mathbf{A}_{\text{drop}})))\Big)^2\Bigg] \\
&= \mathbb{E}_{\mathbf{A}_{\text{drop}}}\left[\sum_{v,i=1}^{N,C}\left(T(\mathbf{X}, \mathbf{A})_{vi} - S(\mathbf{X}, \mathbf{A})_{vi}\right)^2\right] \\
&+ \mathbb{E}_{\mathbf{A}_{\text{drop}}}\left[\left(vec(\nabla_{\mathbf{A}} T(\mathbf{X}, \mathbf{A})_{vi})^T vec(\mathbf{A}_{\text{drop}})) - vec(\nabla_{\mathbf{A}} S(\mathbf{X}, \mathbf{A}))_{vi})^T vec(\mathbf{A}_{\text{drop}}))\right)^2\right] \\
&+ \mathbb{E}_{\mathbf{A}_{\text{drop}}}\left[\sum_{v,u=1}^{N,N}\mathcal{O}((\mathbf{A}_{\text{drop}})_{vu}^2)\right] \\
&= \sum_{v,i=1}^{N,C}\left(T(\mathbf{X})_{vi} - S(\mathbf{X})_{vi}\right)^2 \\
&+ \mathbb{E}_{\mathbf{A}_{\text{drop}}}\left[\left(vec(\nabla_{\mathbf{A}} T(\mathbf{X}, \mathbf{A})_{vi} - \nabla_{\mathbf{A}} S(\mathbf{X}, \mathbf{A})_{vi})^T vec(\mathbf{A}_{\text{drop}})\right)^2 + \sum_{v,u=1}^{N,N}\mathcal{O}((\mathbf{A}_{\text{drop}})_{vu}^2)\right]
\end{aligned}$$

All terms linear in $\mathbf{A}_{\text{drop}}$ have expectation zero, as $\mathbb{E}\left[(\mathbf{A}_{\text{drop}})_{uv}\right] = 0$ for all $u, v \in [1, \ldots, N]$.

Table 1: Mean and standard deviation of our proposed metrics over five runs with random parameter initializations.

Dataset	Acc./F1-score (%)	C (%)	ID (%)	corr($\mathbf{id}, \mathbf{s}$)	corr($\mathbf{h}, \mathbf{s}$)
Citeseer	54.3 ± 2.7	32.2 ± 3.9	47.8 ± 4.9	-0.03 ± 0.07	-0.08 ± 0.03
Photo	56.5 ± 12.2	59.0 ± 8.5	55.6 ± 11.5	0.00 ± 0.06	-0.08 ± 0.06
WikiCS	71.1 ± 1.0	29.7 ± 2.6	21.7 ± 5.7	-0.06 ± 0.02	-0.12 ± 0.01
Computers	47.0 ± 13.1	71.4 ± 7.2	64.6 ± 23.2	0.00 ± 0.09	-0.03 ± 0.03
Physics	90.0 ± 2.1	13.6 ± 3.9	29.1 ± 10.4	-0.01 ± 0.08	-0.22 ± 0.06
PPI	72.0 ± 0.1	10.7 ± 0.1	19.2 ± 0.5	-	-

Appendix B. Additional Experiments using the GCN

In this section, we provide an evaluation of the experiments shown in the main paper replacing the GAT layers with GCN layers. The experimental setup remains the same with the teacher being the same high-capacity GAT model. The motivation for employing a simpler student model stems from it being computationally more memory and runtime efficient during inference. However, due its inferior expressivity, the GCN may not be able to match influence. The extent of influence differences is generally unclear as the GCN uses fixed edge weights only based on the node degrees.

The results for our metrics for GCN models are shown in Table 1. The influence difference is still very noticeable across all datasets, albeit less pronounced. It is again not correlated to the stability of a node prediction. The correlation to the entropy of the neighboring node labels is larger in all cases, though it is rather weak.

The effects on Knowledge Distillation are presented in Table 2. Here, the results demonstrate a higher degree of variance. For the Photo dataset, accuracy is improved by 16.0% and for the Computers dataset by 14.6%. These results indicate a large potential in guiding less expressive models toward desired solutions. However, DD is not always as effective as the accuracy is slightly behind the best other method for three datasets. Results regarding prediction churn are presented in Table 3. Again, DD achieves large improvements for some datasets but is ineffective for others. A similar influence may not lead to optimal results for models with different expressive power.

Table 2: comparison of the performance on the node classification tasks. For PPI, the F1-score is reported, and in all other cases, accuracy is reported. The best results are indicated in bold, the second-best are underlined.

Accuracy/F1-score	Computers	Physics	WikiCS	Photo	Citeseer	PPI
Teacher	80.8	91.2	79.7	85.4	68.8	98.8
Student	47.0 ± 13.1	90.0 ± 2.1	71.1 ± 1.0	56.5 ± 12.2	54.3 ± 2.7	$\underline{72.0} \pm 0.1$
Student+DropEdge	48.0 ± 12.6	$\underline{90.8} \pm 1.1$	71.6 ± 1.1	58.8 ± 10.5	$\mathbf{57.5} \pm 2.7$	70.4 ± 0.1
KD	47.6 ± 13.8	90.2 ± 2.0	73.5 ± 0.8	61.2 ± 7.9	$\underline{57.0} \pm 2.1$	$\mathbf{72.1} \pm 0.3$
KD+DropEdge	46.0 ± 9.2	$\mathbf{91.0} \pm 1.4$	73.7 ± 0.5	60.3 ± 8.2	56.6 ± 1.6	70.5 ± 0.2
G-CRD	$\underline{49.9} \pm 12.4$	89.6 ± 1.9	73.5 ± 0.6	62.1 ± 13.5	54.3 ± 2.4	70.9 ± 0.2
G-CRD+DropEdge	49.1 ± 10.4	90.8 ± 1.4	$\underline{73.8} \pm 0.6$	$\underline{62.5} \pm 14.3$	56.2 ± 2.2	68.0 ± 0.4
DropDistillation	$\mathbf{63.7} \pm 4.5$	88.0 ± 1.2	$\mathbf{74.3} \pm 0.7$	$\mathbf{78.5} \pm 6.5$	$\underline{57.0} \pm 2.8$	71.9 ± 0.2

Table 3: Average model churn C(S, T) between the teacher and each student. The models are the same as in Table 2. Lower scores are better.

Churn C	Computers	Physics	WikiCS	Photo	Citeseer	PPI
Student	57.0 ± 14.2	10.8 ± 3.3	26.2 ± 1.4	44.8 ± 10.9	41.5 ± 1.8	$\underline{15.7} \pm 0.1$
Student+DropEdge	59.6 ± 12.9	$\underline{9.7} \pm 3.3$	24.8 ± 0.5	42.9 ± 10.3	$\mathbf{37.2} \pm 3.6$	16.3 ± 0.1
KD	57.7 ± 12.2	12.9 ± 5.3	19.6 ± 0.4	41.5 ± 7.7	60.1 ± 3.2	$\underline{15.6} \pm 0.1$
KD+DropEdge	59.8 ± 10.4	11.8 ± 2.7	$\underline{19.0} \pm 0.5$	41.3 ± 7.7	$\underline{57.3} \pm 8.0$	16.3 ± 0.1
G-CRD	$\underline{54.8} \pm 11.6$	10.8 ± 1.4	22.1 ± 0.6	$\underline{40.6} \pm 11.5$	44.3 ± 2.2	16.3 ± 0.1
G-CRD+DropEdge	55.4 ± 7.7	$\underline{9.8} \pm 1.7$	22.1 ± 0.9	41.0 ± 14.2	$\mathbf{37.1} \pm 4.7$	17.8 ± 0.2
DropDistillation	$\mathbf{39.4} \pm 9.1$	15.1 ± 3.8	$\mathbf{18.4} \pm 0.5$	$\mathbf{22.4} \pm 5.6$	51.7 ± 2.3	15.8 ± 0.1