NEURAL INFORMATION PROCESSING SYSTEMS

LoSplit: Loss-Guided Dynamic Split for Training-Time Defense Against Graph Backdoor Attacks

Di Jin[1], Yuxiang Zhang[1], Binddao Feng[1]*, Xiaobao Wang[1] , Dongxiao He[1] , Zhen Wang[2]

The Thirty-Nineth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.

Code: https://github.com/zyx924768045/LoSplit

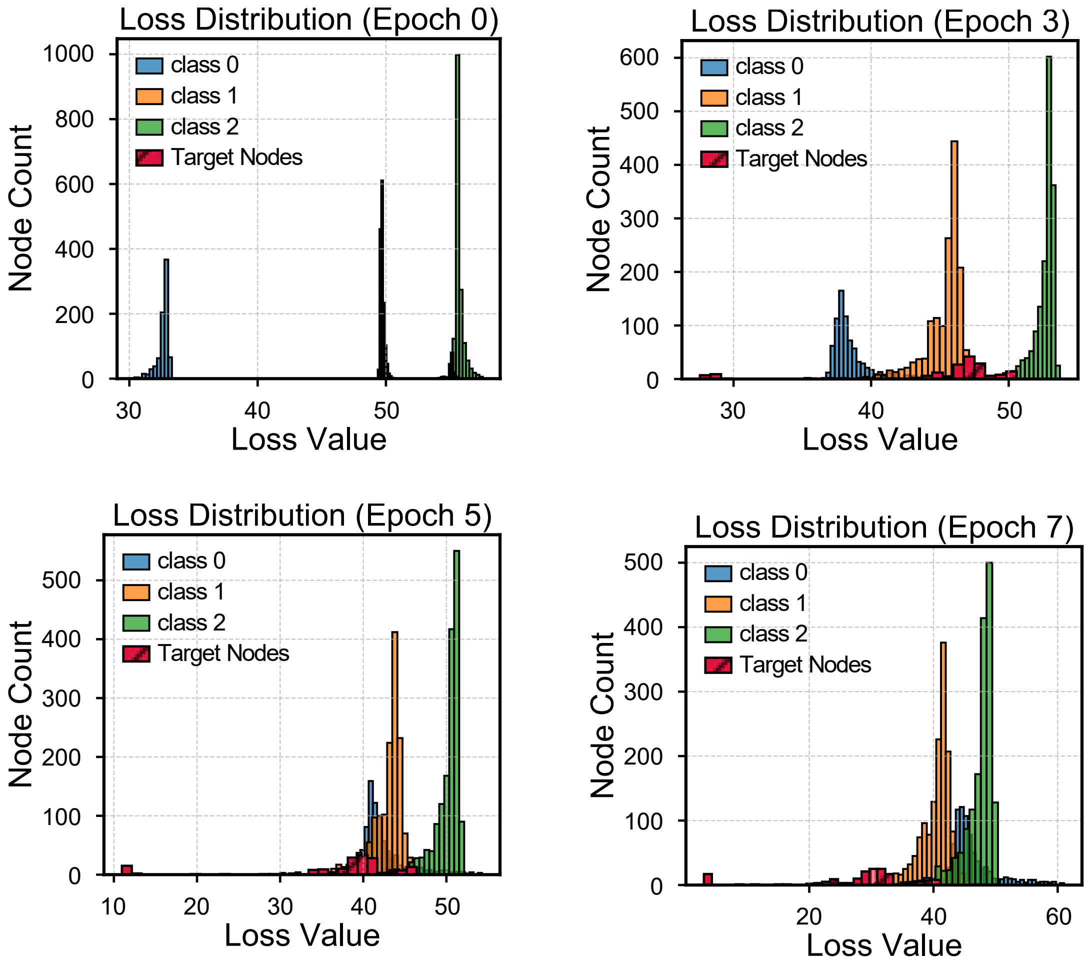
[1]Tianjin University, [2]Northwestern Polytechnical University, *Corresponding Author

Graph Backdoor Attack to GNNs

- Graph Neural Networks (GNNs) have strong performance in node classification
- GNNs are susceptible to backdoor attacks
- Adversaries insert triggers into training data to mislead the GNNs to malicious labels when trigger appears while maintain normal when there is no trigger, posing great risk to safety-critical applications.
- Previous defense strategies focus on detecting structural anomalies but fail against subtle feature-perturbing attacks, underscoring the need for more advanced defense.

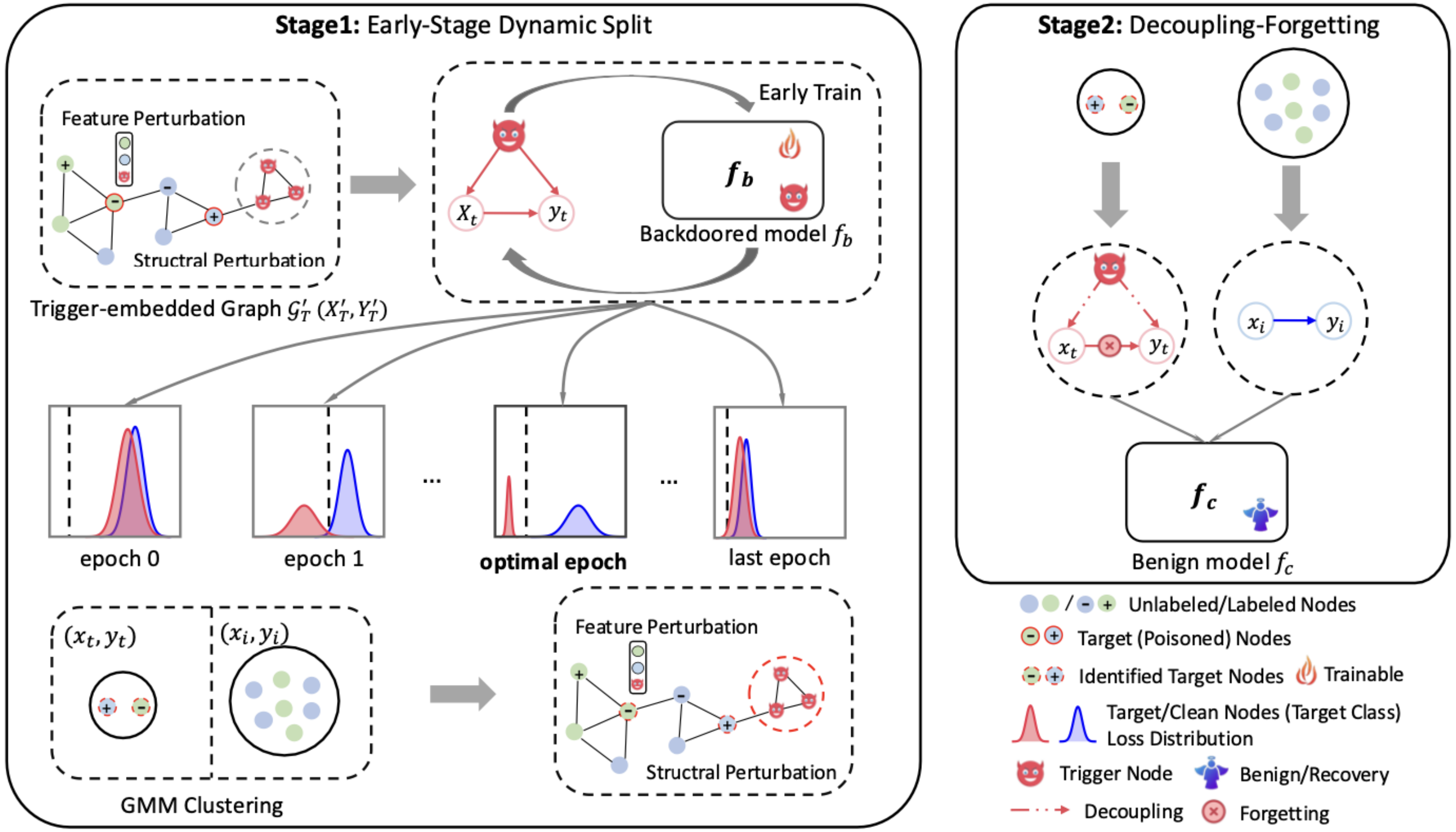
Class-wide loss drift

- Both structural and feature-based backdoor attacks show early convergence of target nodes due to shortcut learning.
- In graphs, message passing causes an unstable class-wide loss drift, making approaches in images ineffective in graphs.
- **Challenge:** How to precisely identify target nodes even in the presence of unstable class-wide loss drift?



Di Jin, Yuxiang Zhang et al., LoSplit: Loss-Guided Dynamic Split for Training-Time Defense Against Graph Backdoor Attacks. In NeurIPS 2025.

Methodology

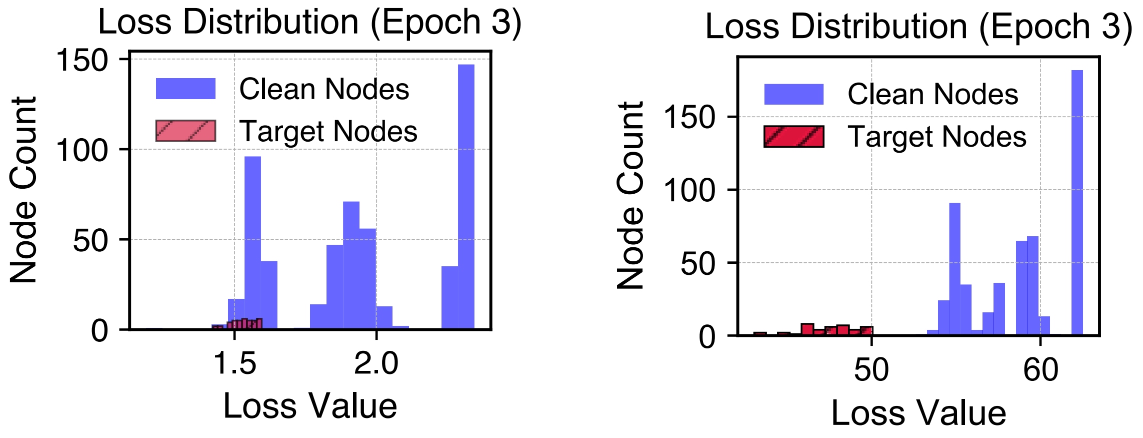


- **Early-Stage Dynamic Split:** we exploit the distinct early loss dynamics under RCE loss to split target nodes and clean nodes.
- **Decoupling–Forgetting:** the identified target nodes are decoupled and forgotten from the malicious labels to mitigate the backdoor/shortcut effect.

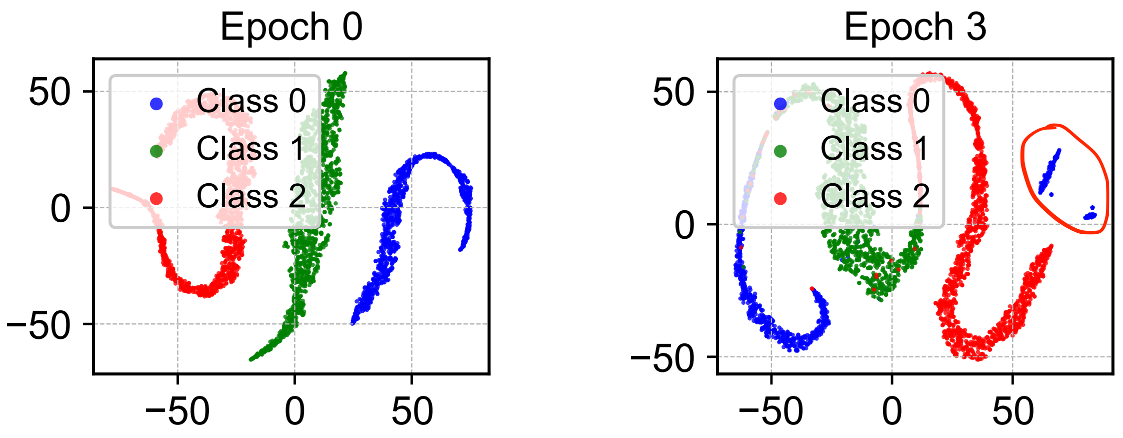
Di Jin, Yuxiang Zhang et al., LoSplit: Loss-Guided Dynamic Split for Training-Time Defense Against Graph Backdoor Attacks. In NeurIPS 2025.

Early-Stage Loss Dynamics

- **CE Loss vs. RCE Loss**
  - Target nodes converge much faster than clean nodes in early loss behavior.
  - This is amplified under RCE loss, making targets nodes more easily distinguishable.



- **Loss Dynamics**
  - Nodes of the same class form compact loss clusters.
  - Nodes in the target class (malicious label) cluster splits into smaller sub-clusters mainly containing target nodes.



Target Nodes Identification via Early-Stage Dynamic Split

- **Malicious label Identification:**

$$y_t = \arg\max_{y_j \in \mathcal{Y}'_T} \mathrm{Var}\left(\{\ell_i^{(t)} \mid y_i = y_j\}\right)$$

- **Epoch-wise Loss normalization:**

$$\zeta_i^{(t)} = \frac{\ell_i^{(t)} - \mu}{\sigma + \epsilon}, \quad \forall v_i \in \mathcal{V}_{y_t}^{(t)},$$

- **Optimal Epoch Selection:**

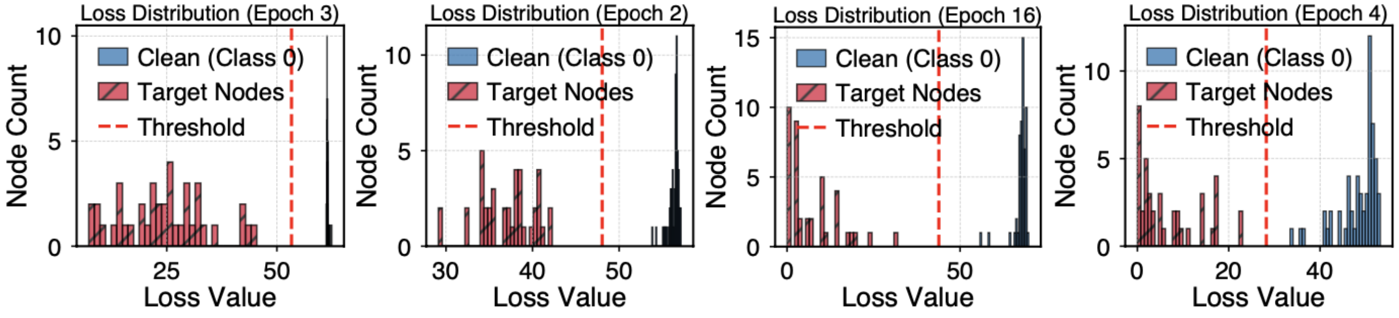
$$t^* = \arg\max_t \left(\mathbb{E}_{v_i \in \mathcal{C}_{\text{high}}^{(t)}}[\ell_i^{(t)}] - \mathbb{E}_{v_j \in \mathcal{C}_{\text{low}}^{(t)}}[\ell_j^{(t)}]\right).$$

- **Splitting Point:**

$$\tau^{(t)} = \max\left\{\zeta_i \mid v_i \in \mathcal{C}_{\text{low}}^{(t)}\right\} + \frac{\min\left\{\zeta_j \mid v_j \in \mathcal{C}_{\text{high}}^{(t)}\right\} - \max\left\{\zeta_i \mid v_i \in \mathcal{C}_{\text{low}}^{(t)}\right\}}{2}.$$

- **Target nodes and clean nodes candidates:**

$$\mathcal{V}_B^{S(t^*)} = \left\{v_i \in \mathcal{V}_{y_t}^{(t^*)} \mid \zeta_i^{(t^*)} \leq \tau^{(t^*)}\right\}, \quad \mathcal{V}_C^{S(t^*)} = \mathcal{V}_T \setminus \mathcal{V}_B^{S(t^*)}.$$



GTA UGBA DPGBA SPEAR

Backdoor Recovery via Decoupling-Forgetting

NEURAL INFORMATION PROCESSING SYSTEMS

$$\min_{\theta} \mathcal{L}_{\theta} = \underbrace{\gamma \sum_{v_i \in \mathcal{V}_B^{S(t^*)}} \mathcal{L}\big(f_\theta(v_i), \tilde{y}_i\big)}_{\text{Random Relabeling}} + \underbrace{(1-\gamma) \sum_{v_i \in \mathcal{V}_B^{S(t^*)}} -\mathcal{L}\big(f_\theta(v_i), y_t\big)}_{\text{Gradient Ascent}} + \underbrace{\sum_{v_j \in \mathcal{V}_C^{S(t^*)}} \mathcal{L}\big(f_\theta(v_j), y_j\big)}_{\text{Normal Training}},$$

- LoSplit removes backdoor effects using a Decoupling–Forgetting strategy combining **random label reassignment** and **gradient ascent**.
- Random relabeling breaks shortcut learning, while gradient ascent pushes target nodes away from malicious boundary.
- Clean nodes are trained normally to maintain model performance.

Strategy	GTA		UGBA		DPGBA		SPEAR	
	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑	ASR↓	CA↑
GCN (No Defense)	97.81	84.42	95.69	83.16	98.78	84.98	<u>92.90</u>	**85.13**
Node Removal	0.13	84.98	0.00	85.08	<u>95.30</u>	**85.39**	0.33	84.73
Feature Reinitialization	100.00	80.92	0.00	84.07	98.39	84.78	100.00	81.78
Restore Original Label	**0.00**	<u>84.37</u>	0.00	81.28	**1.15**	84.93	0.08	84.24
SCRUB [26]	**0.00**	<u>84.58</u>	0.00	82.90	97.75	84.63	0.00	84.47
LoSplit	0.06	**85.19**	**0.00**	**85.33**	1.92	84.93	**0.00**	**85.13**

Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS

Attack	Defense	Cora		CiteSeer		PubMed		Physics		Flickr		OGB-arXiv	
		ASR(%) ↓	CA(%) ↑	ASR(%) ↓	CA(%) ↑	ASR(%) ↓	CA(%) ↑	ASR(%) ↓	CA(%) ↑	ASR(%) ↓	CA(%) ↑	ASR(%) ↓	CA(%) ↑
GTA	GCN	98.52	82.96	99.40	73.80	97.62	84.53	100.00	96.23	100.00	42.39	94.70	63.12
	RobustGCN	100.00	81.85	99.70	73.49	97.87	85.19	100.00	94.98	99.89	40.44	99.83	60.16
	GNNGuard	38.38	75.19	12.31	62.95	21.35	81.33	80.94	96.35	0.24	43.75	0.88	63.42
	Prune	12.88	82.22	13.21	71.39	21.10	85.08	1.16	95.42	0.00	40.41	0.01	62.45
	OD	0.37	81.85	0.00	74.10	0.90	84.63	0.00	96.36	0.00	41.47	0.00	63.31
	ABL	4.80	78.52	1.50	73.19	1.77	83.71	100.00	96.25	0.00	40.80	0.00	63.92
	RIGBD	3.56	83.70	0.00	74.10	3.25	83.21	100.00	96.43	0.00	43.98	0.00	63.07
	LoSplit	0.00	84.81	0.00	75.60	0.06	85.29	0.56	96.43	0.00	44.19	0.00	65.74
UGBA	GCN	98.52	83.70	100.00	74.10	98.97	84.88	100.00	96.26	100.00	40.68	99.08	65.65
	RobustGCN	94.10	80.37	100.00	6.63	95.84	85.59	99.98	95.23	90.25	40.34	87.13	60.87
	GNNGuard	99.63	77.78	100.00	6.63	69.83	82.19	97.86	96.06	99.07	40.80	96.21	65.51
	Prune	98.52	78.52	96.70	72.89	88.29	85.08	95.73	95.16	90.23	40.45	93.99	64.46
	OD	12.92	83.70	0.00	75.30	83.98	84.88	0.00	96.20	0.00	40.25	10.13	65.32
	ABL	6.64	78.15	0.00	71.69	3.35	83.41	1.93	95.19	0.00	36.85	6.45	63.26
	RIGBD	7.11	83.70	0.00	73.49	2.54	82.65	0.56	96.38	0.00	40.58	0.00	66.06
	LoSplit	0.00	85.07	0.00	75.60	0.00	85.23	0.14	96.57	0.00	40.94	0.00	66.52
DPGBA	GCN	98.67	84.44	98.66	73.49	97.88	85.19	100.00	96.58	99.98	40.29	93.12	65.47
	RobustGCN	97.79	84.65	100.00	74.40	99.52	84.86	94.44	96.35	95.61	40.95	87.29	60.07
	GNNGuard	99.63	78.15	99.70	62.95	72.97	81.28	95.59	95.74	4.50	40.46	90.39	63.17
	Prune	22.88	79.63	11.41	72.89	40.92	84.53	1.61	96.23	0.00	40.62	0.12	62.76
	OD	96.31	81.85	97.90	74.10	84.89	81.13	94.52	96.25	98.56	40.59	94.21	65.06
	ABL	4.80	81.85	0.00	71.99	5.22	76.86	81.85	93.30	50.16	40.26	3.91	55.10
	RIGBD	2.22	84.07	0.30	74.40	4.92	84.37	0.98	96.27	0.00	40.78	11.83	63.43
	LoSplit	0.00	85.56	0.00	74.40	1.93	84.93	0.00	96.52	0.00	41.24	0.00	65.24
SPEAR	GCN	100.00	81.85	99.10	73.49	97.87	84.98	95.36	96.27	100.00	45.56	98.98	66.38
	RobustGCN	100.00	16.30	91.59	74.40	93.61	85.44	90.91	96.30	98.91	40.43	53.44	62.08
	GNNGuard	53.51	80.37	29.72	62.95	62.73	81.84	63.48	96.14	71.84	44.64	94.60	66.79
	Prune	100.00	84.07	100.00	72.29	98.83	85.19	96.78	96.15	100.00	40.52	99.83	65.77
	OD	100.00	80.00	100.00	76.50	94.48	85.29	53.92	96.22	41.59	41.48	66.31	66.31
	ABL	30.26	82.59	0.00	73.19	5.32	84.27	11.56	94.69	100.00	40.59	11.75	62.31
	RIGBD	97.78	83.70	90.27	72.29	88.98	84.68	88.03	96.35	100.00	44.24	97.09	66.72
	LoSplit	0.00	84.44	0.00	75.00	0.25	85.24	0.00	96.42	0.00	45.80	0.20	66.68

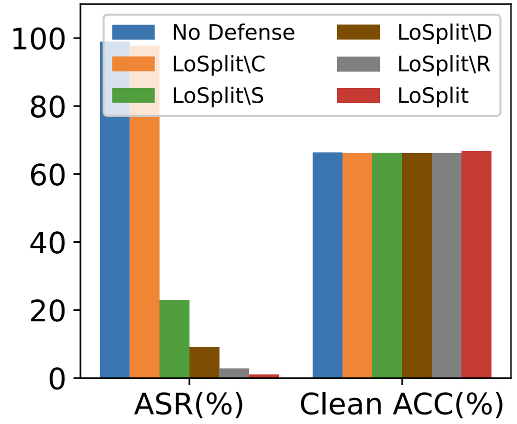
Comparison of Defense Performance

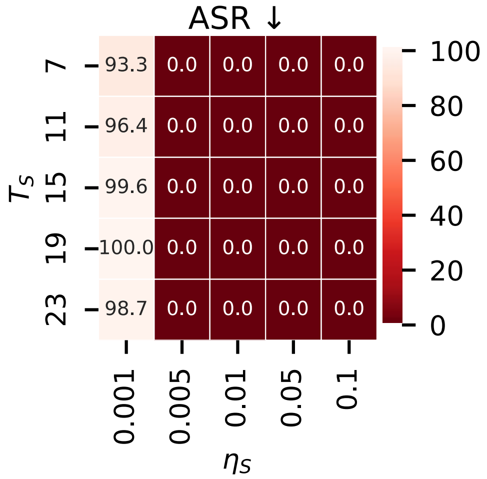
Attack	Defense	Cora			Citeseer			PubMed		
		Prec. ↑	Rec. ↑	FPR ↓	Prec. ↑	Rec. ↑	FPR ↓	Prec. ↑	Rec. ↑	FPR ↓
GTA	ABL	100.00	85.00	0.00	88.10	92.50	0.75	24.39	100.00	3.05
	RIGBD	85.00	85.00	1.11	95.00	95.00	0.30	93.75	93.75	0.25
	LoSplit	100.00	100.00	0.00	100.00	97.50	0.00	98.77	100.00	0.05
UGBA	ABL	100.00	27.50	0.00	100.00	35.00	0.00	100.00	51.25	0.00
	RIGBD	72.50	72.50	2.03	77.50	77.50	1.35	92.41	92.41	0.30
	LoSplit	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00
DPGBA	ABL	100.00	42.50	0.00	100.00	52.50	0.00	19.51	60.00	2.44
	RIGBD	81.08	75.00	1.29	97.06	82.50	0.15	92.06	72.50	0.25
	LoSplit	100.00	100.00	0.18	100.00	97.50	0.00	100.00	83.12	0.00
SPEAR	ABL	100.00	12.50	0.00	97.14	85.00	0.15	40.00	40.00	0.60
	RIGBD	83.33	12.50	0.18	93.75	37.50	0.10	80.77	51.22	0.13
	LoSplit	100.00	100.00	0.00	93.02	100.00	0.45	88.90	100.00	0.05

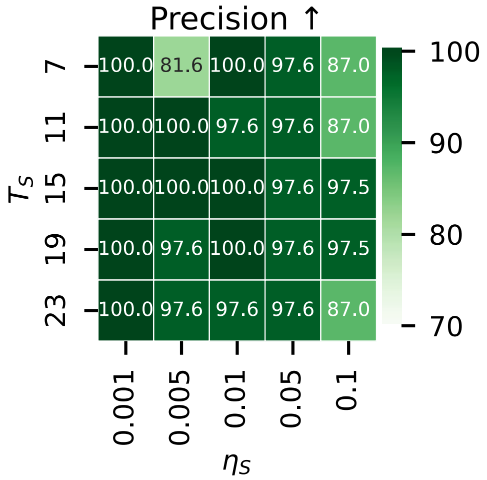
Attack	Defense	Physics			Flickr			OGB-arXiv		
		Prec. ↑	Rec. ↑	FPR ↓	Prec. ↑	Rec. ↑	FPR ↓	Prec. ↑	Rec. ↑	FPR ↓
GTA	ABL	12.50	10.00	0.92	21.67	97.50	0.78	8.14	70.00	0.92
	RIGBD	87.14	75.00	0.18	93.13	93.13	0.61	97.17	97.17	0.05
	LoSplit	96.97	100.00	0.07	96.36	99.37	0.03	99.65	99.82	0.01
UGBA	ABL	8.33	6.25	0.95	21.67	97.50	0.78	7.72	62.50	0.93
	RIGBD	92.41	80.00	0.14	99.37	98.12	0.06	98.76	98.76	0.02
	LoSplit	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00
DPGBA	ABL	10.71	7.50	0.88	0.00	0.00	1.01	9.26	85.00	0.53
	RIGBD	85.29	71.25	0.25	100.00	83.12	0.00	86.21	4.42	0.01
	LoSplit	96.97	100.00	0.07	100.00	88.12	0.00	100.00	97.50	0.00
SPEAR	ABL	11.76	9.38	0.90	0.00	0.00	1.01	96.80	58.94	0.03
	RIGBD	78.57	72.50	0.33	0.89	100.00	0.00	0.00	0.00	0.00
	LoSplit	92.12	90.50	0.56	100.00	100.00	0.00	99.45	96.81	0.01

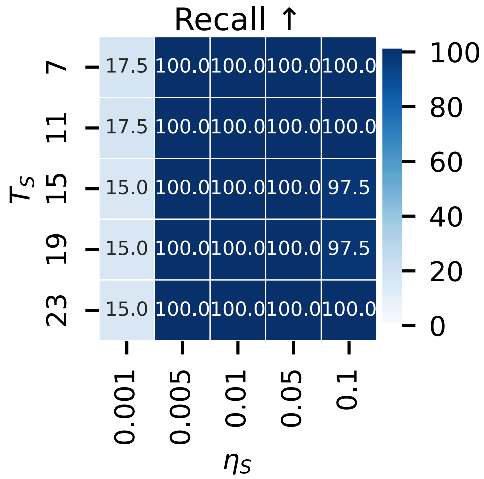
Comparison of Target Nodes Identification Ability

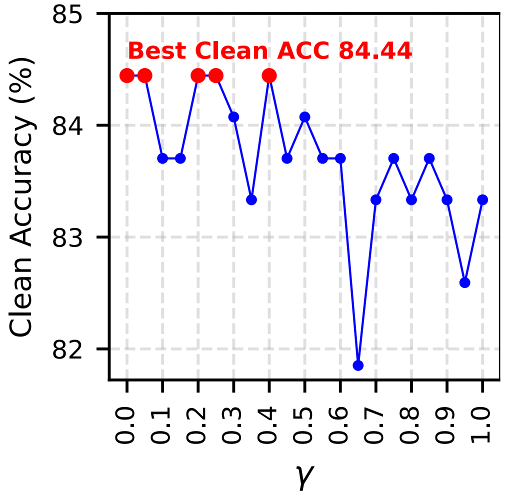
Ablation Study and Hyperparameter Analysis











- Ablation study shows that each component—RCE loss, dynamic split, and Decoupling-Forgetting—plays a crucial role, and removing any of them significantly weakens defense performace.

- Hyperparameter analysis reveals that moderate early-stage epochs (TS) and learning rates (ηS) yield optimal attack suppression and precise target identification.

- Overall, LoSplit maintains high robustness and clean accuracy across a broad hyperparameter range, outperforming SOTA method RIGBD.

Performance on Clean Graph

	Cora	Citeseer	PubMed	Physics	Flickr	OGB-arXiv
GCN (CA)	83.70	74.70	85.18	96.02	45.33	66.12
LoSplit (CA)	83.33	74.39	85.03	95.87	45.11	65.98
LoSplit (FPR)	0.18	1.05	0.48	0.07	0.92	0.36

- On clean graphs, LoSplit maintains almost the same accuracy compared to when there is no defense (GCN).
- The false positive rate is nearly zero, meaning clean nodes are not misclassified.
- This demonstrates LoSplit's utility and safety even when we don't know whether the graph is contaminated or not.

Code: https://github.com/zyx924768045/LoSplit