NDSS 2025

DShield: Defending against Backdoor Attacks on Graph Neural Networks via Discrepancy Learning

Hao Yu[1], Chuan Ma[2], Xinhang Wan[1], Jun Wang[1], Tao Xiang[2], Meng Shen[3], and Xinwang Liu[1]
National University of Defense Technology
Chongqing University
Beijing Institute of Technology

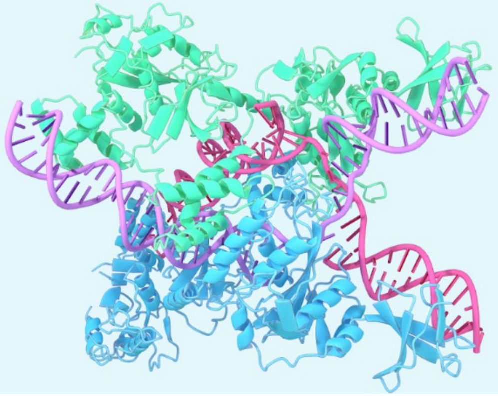
Speaker: **Zhongming Wang**

Feb. 27, 2025@San Diego, USA

Graph Neural Networks

- Many types of data can be represented as **graph structures**, e.g., social networks and protein molecules
- GNNs **propagate** information across graph structures, effectively capturing complex relationships between nodes and edges
- Tasks: node classification, graph classification, link prediction

Social Networks



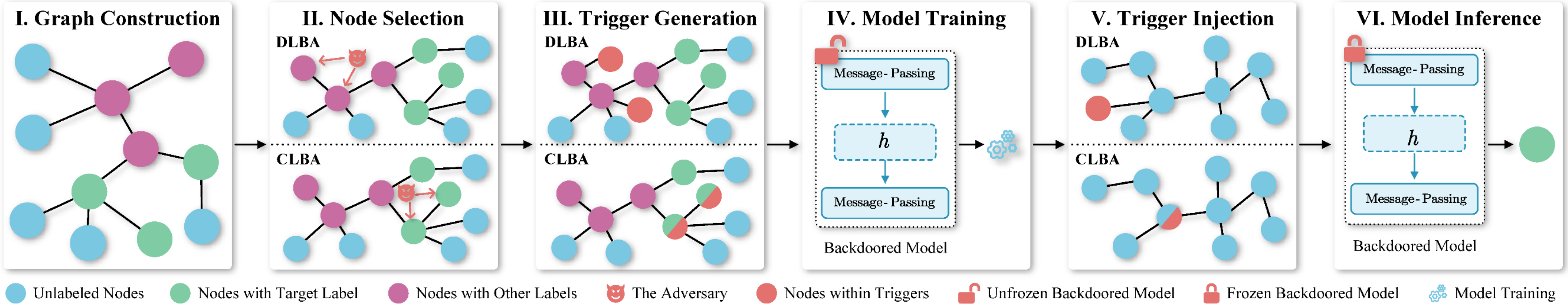
Drug Discovery

Smart City

Bank Risk Management

Backdoor Attacks on GNNs

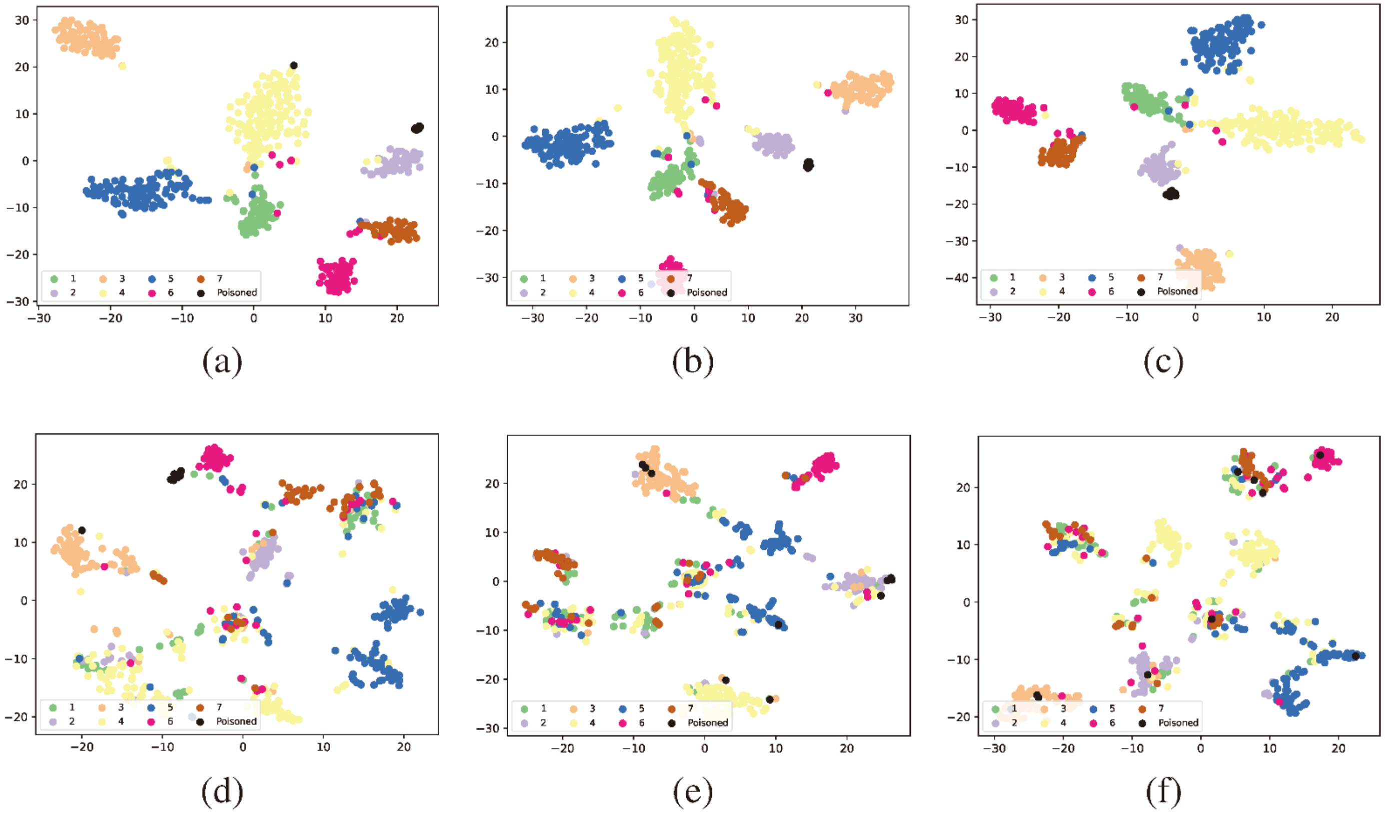
- Backdoor attacks
  - Inducing **undesirable** behaviors in **backdoored** models by injecting **triggers** into original graphs
- Two variants
  - Dirty-label backdoor attacks (**DLBAs**): Altering label information
  - Clean-label backdoor attacks (**CLBAs**): Altering attributes of normal nodes



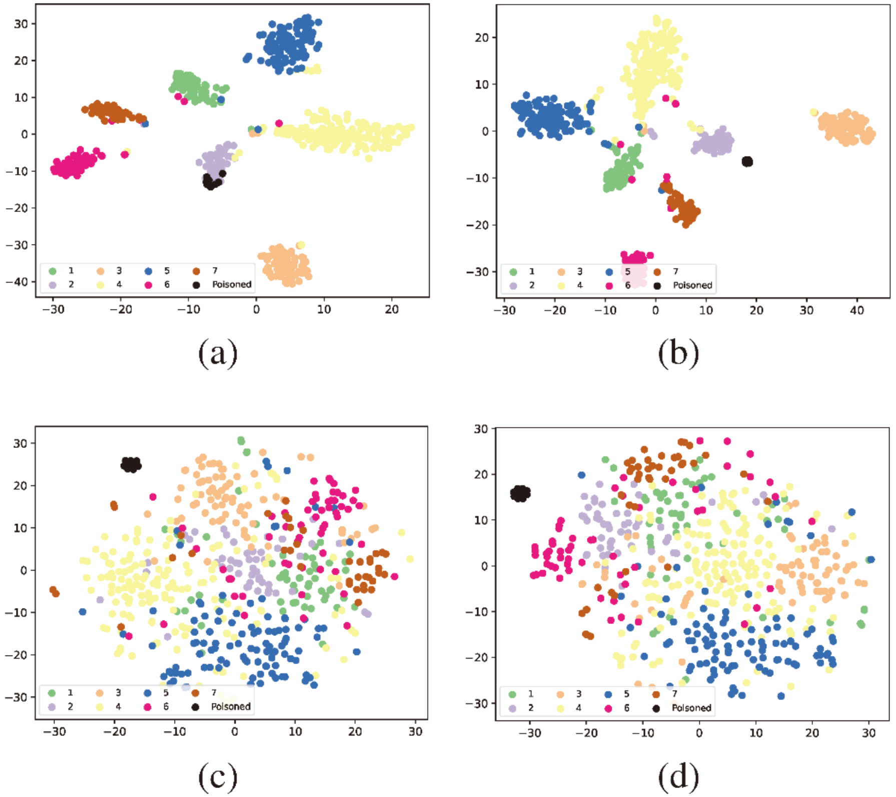
Motivation

➢ Two **phenomena**

- Semantic Drift of DLBAs： between node’s attributes and structures, and their labels
- Attribute over-emphasis of CLBAs: high similarities of important attributes



Semantic Drift

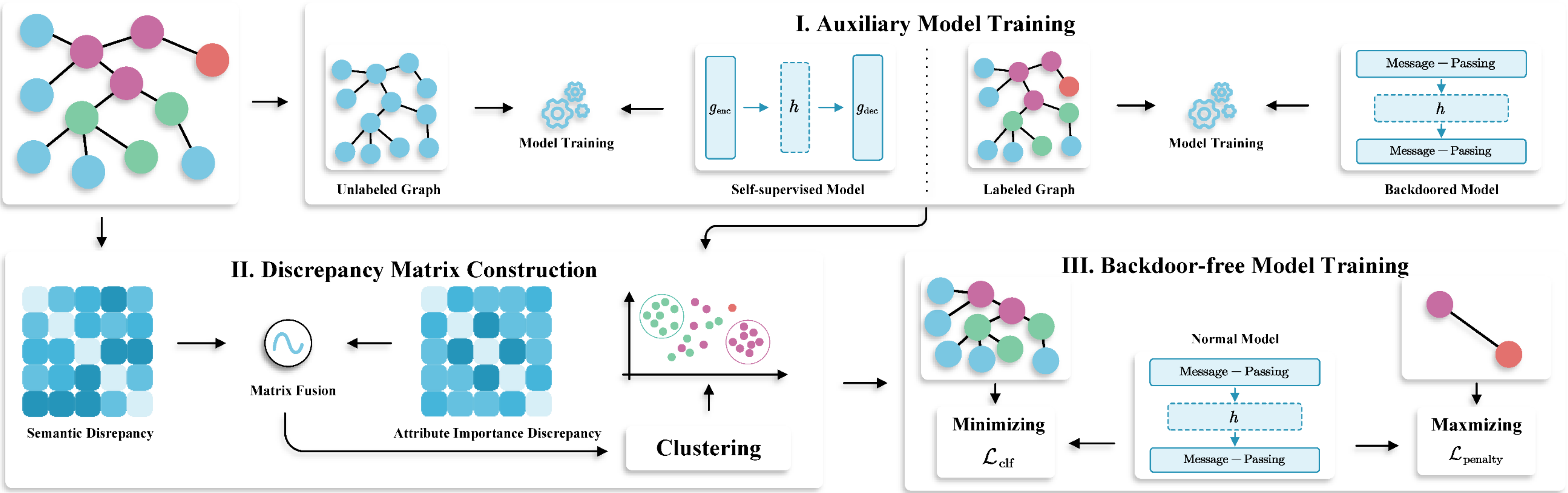


Attribute Over-emphasis

Proposed DShield

➢ Pipeline

- Auxiliary Model Training: Preparing to identify poisoned nodes in DLBAs
- Discrepancy Matrix Construction: Detecting poisoned nodes in both DLBAs and CLBAs
- Backdoor-free Model Training: Training a model without backdoors

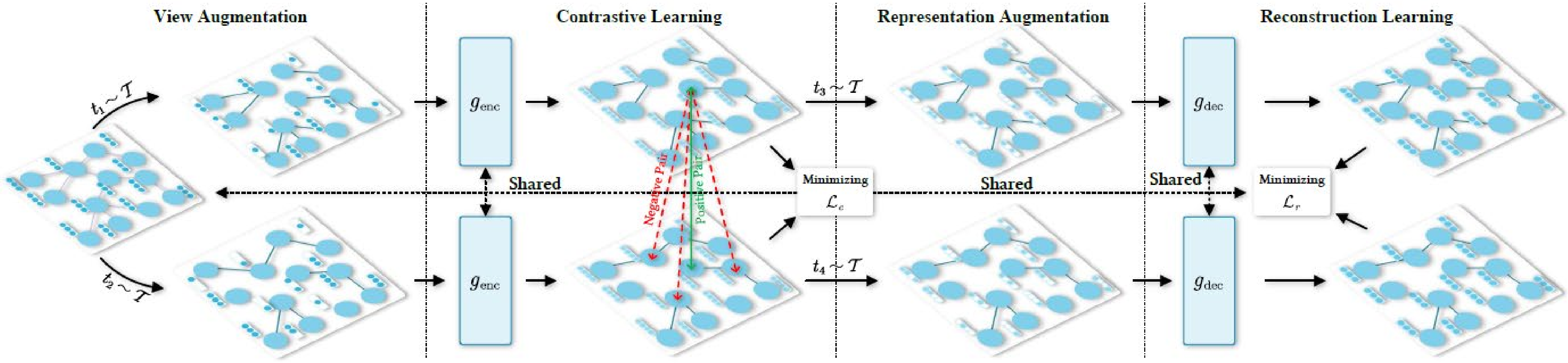


Auxiliary Model Training

- Backdoored Model

$$\mathcal{L} = \frac{1}{|\mathcal{V}^{train}|} \sum_{v_i \in \mathcal{V}^{train}} \mathrm{CE}\Big(\tilde{f}(\boldsymbol{A}, \boldsymbol{X})_i, \boldsymbol{y}_i\Big)$$

- Self-supervised Model
  - View Augmentation: Sampling two stochastic augmentation functions
  - View Encoding: Extracting latent representations of nodes
  - Contrast and Reconstruction: Distinguishing between representations of two nodes



Auxiliary Model Training

➢View Augmentation

$$P\{(v_i,v_j)\in\widehat{\mathcal{E}}\}=1-p_{ij}^s;\quad p_{ij}^s=\frac{\exp(-s(\widetilde{\boldsymbol{h}}_i,\widetilde{\boldsymbol{h}}_j))}{\sum_{(v_k,v_t)\in\widetilde{\mathcal{E}}}\exp(-s(\widetilde{\boldsymbol{h}}_k,\widetilde{\boldsymbol{h}}_t))}$$

Structure-level

$$\boldsymbol{g}_i=\frac{\partial \mathrm{CE}(f(\widetilde{\mathbf{A}},\widetilde{\mathbf{X}};\phi)_i,\boldsymbol{y}_i)}{\partial \boldsymbol{x}_i};\quad \widehat{\boldsymbol{A}}=[\boldsymbol{x}_1\odot\widehat{\boldsymbol{m}}_1;\cdots;\boldsymbol{x}_N\odot\widehat{\boldsymbol{m}}_N]^\top$$

Attribute-level

➢View Encoding

$$\widehat{\boldsymbol{H}}_1=g_{\mathrm{enc}}(\widehat{\boldsymbol{A}}_1,\widehat{\boldsymbol{X}}_1),\qquad \widehat{\boldsymbol{H}}_2=g_{\mathrm{enc}}(\widehat{\boldsymbol{A}}_2,\widehat{\boldsymbol{X}}_2)$$

➢Contrast and Reconstruction

Contrast

$$\mathcal{L}_{c_1}=\sum_{v_i\in\mathcal{V}^{\mathrm{train}}}-\frac{1}{|\mathcal{P}(v_i)|}\sum_{v_j\in\mathcal{P}(v_i)}\log\frac{\exp(s(\widehat{\boldsymbol{h}}_{1,i},\widehat{\boldsymbol{h}}_{2,j})/\tau)}{\sum_{v_k\in\mathcal{N}(v_i)}\exp(s(\widehat{\boldsymbol{h}}_{1,i},\widehat{\boldsymbol{h}}_{2,k})/\tau)};$$

Reconstruction

$$\widehat{\boldsymbol{H}}'_1=[\widehat{\boldsymbol{h}}_1\odot\widehat{\boldsymbol{m}}'_1;\widehat{\boldsymbol{h}}_2\odot\widehat{\boldsymbol{m}}'_2;\cdots;\widehat{\boldsymbol{h}}_N\odot\widehat{\boldsymbol{m}}'_N]^\top\qquad \widehat{\boldsymbol{H}}'_1=g_{\mathrm{dec}}(\widehat{\boldsymbol{A}}_1,\widehat{\boldsymbol{X}}'_1),\quad \widehat{\boldsymbol{H}}'_2=g_{\mathrm{dec}}(\widehat{\boldsymbol{A}}_2,\widehat{\boldsymbol{X}}'_2)$$

$$\mathcal{L}_r=\frac{1}{2}\sum_{i=1}^{N}(1-s(\widetilde{\boldsymbol{x}}_i,\widehat{\boldsymbol{x}}'_{1,i}))^2+\frac{1}{2}\sum_{i=1}^{N}(1-s(\widetilde{\boldsymbol{x}}_i,\widehat{\boldsymbol{x}}'_{2,i}))^2$$

Discrepancy Matrix Construction

➢Semantic Discrepancy Matrix

$$\widetilde{\boldsymbol{H}}_{\mathrm{sl}} = \tilde{f}(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}; \phi_1), \quad \widetilde{\boldsymbol{H}}_{\mathrm{ssl}} = g_{\mathrm{enc}}(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}})$$

$$\boldsymbol{D}^y_{\mathrm{sl}} = \boldsymbol{1}\,\mathrm{diag}(\widetilde{\boldsymbol{H}}^y_{\mathrm{sl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{sl}})^\top - 2\widetilde{\boldsymbol{H}}^y_{\mathrm{sl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{sl}} + \mathrm{diag}(\widetilde{\boldsymbol{H}}^y_{\mathrm{sl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{sl}})\boldsymbol{1}^\top;$$

$$\boldsymbol{D}^y_{\mathrm{ssl}} = \boldsymbol{1}\,\mathrm{diag}(\widetilde{\boldsymbol{H}}^y_{\mathrm{ssl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{ssl}})^\top - 2\widetilde{\boldsymbol{H}}^y_{\mathrm{ssl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{ssl}} + \mathrm{diag}(\widetilde{\boldsymbol{H}}^y_{\mathrm{ssl}} \widetilde{\boldsymbol{H}}^{y\top}_{\mathrm{ssl}})\boldsymbol{1}^\top,$$

$$\boldsymbol{D}^y_s = \max(\boldsymbol{D}^y_{\mathrm{ssl}} - \boldsymbol{D}^y_{\mathrm{sl}}, 0)$$

➢Attribute Importance Discrepancy Matrix

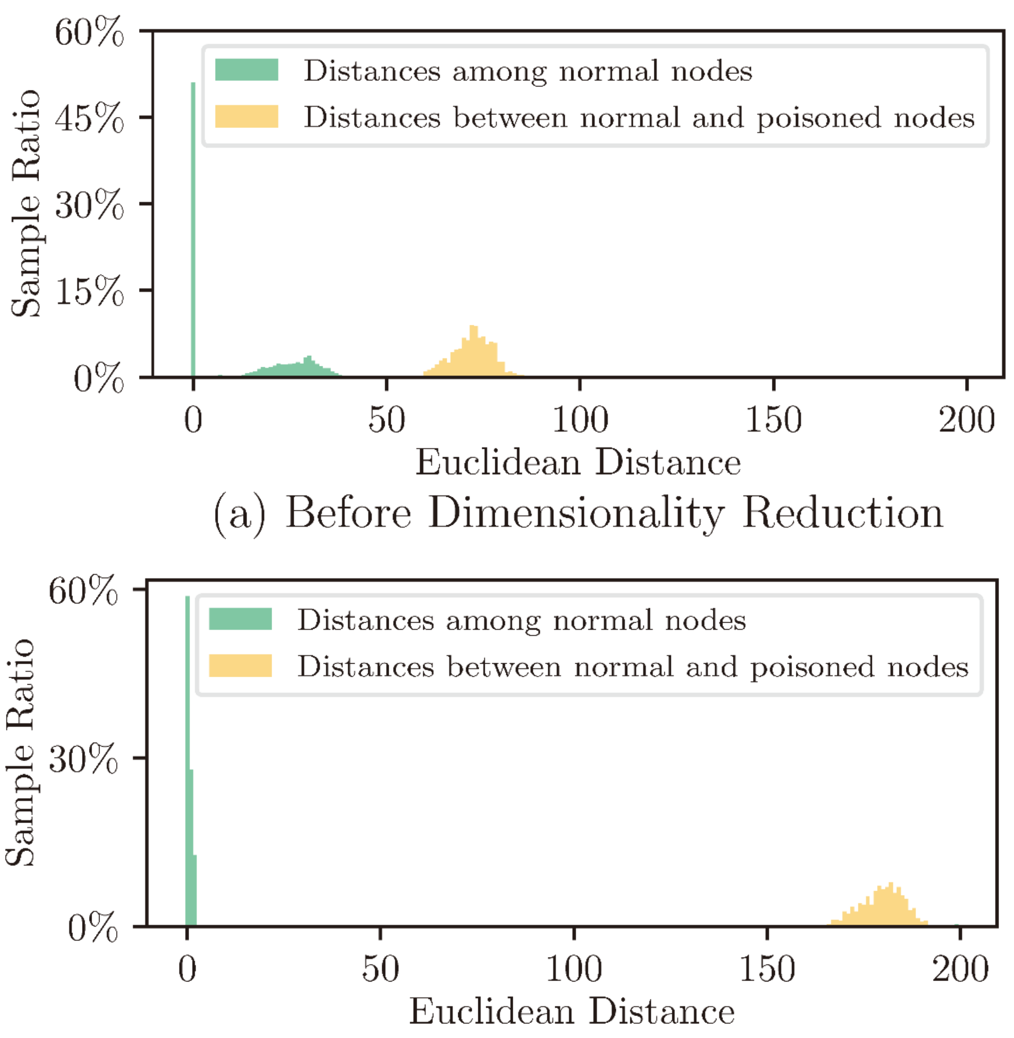
$$\boldsymbol{g}_i = \frac{\partial \mathrm{CE}(f(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}; \phi)_i, \boldsymbol{y}_i)}{\partial \boldsymbol{x}_i}, \quad w_{ij} = \mathbb{1}_{g_{ij} > 0} \qquad \widetilde{\boldsymbol{X}}^y = \widetilde{\boldsymbol{X}}^y \circ \boldsymbol{W}^y$$

$$\overline{\boldsymbol{X}}^y = \mathrm{UMAP}(\widetilde{\boldsymbol{X}}^y, d)$$

$$\boldsymbol{D}^y_a = \boldsymbol{1}\,\mathrm{diag}(\overline{\boldsymbol{X}}^y \overline{\boldsymbol{X}}^{y\top})^\top - 2\overline{\boldsymbol{X}}^y \overline{\boldsymbol{X}}^{y\top} + \mathrm{diag}(\overline{\boldsymbol{X}}^y \overline{\boldsymbol{X}}^{y\top})\boldsymbol{1}^\top$$

➢Clustering

$$\boldsymbol{D}_y = \frac{\mathrm{std}(\boldsymbol{D}^y_s)}{\mathrm{std}(\boldsymbol{D}^y_s) + \mathrm{std}(\boldsymbol{D}^y_a)}\boldsymbol{D}^y_s + \frac{\mathrm{std}(\boldsymbol{D}^y_a)}{\mathrm{std}(\boldsymbol{D}^y_s) + \mathrm{std}(\boldsymbol{D}^y_a)}\boldsymbol{D}^y_a; \quad \mathcal{V}_1, \mathcal{V}_2 = \mathrm{HDBSCAN}(\boldsymbol{D}_y)$$



(a) Before Dimensionality Reduction

(b) After Dimensionality Reduction

Importance of UMAP

Backdoor-free Model Training

➤Target Label Discovery

- Model Traning

$$\widetilde{A}_{ij}^{y,\text{pos}} \begin{cases} = \widetilde{A}_{ij}, & \text{if } v_i \notin \mathcal{V}_2^y \text{ and } v_j \notin \mathcal{V}_2^y \\ = 0, & \text{if } v_i \in \mathcal{V}_2^y \text{ or } v_j \in \mathcal{V}_2^y \end{cases} \qquad \widetilde{A}_{ij}^{y,\text{ neg}} \begin{cases} = \widetilde{A}_{ij}, & \text{if } v_i \in \mathcal{V}_2^y \text{ or } v_j \in \mathcal{V}_2^y \\ = 0, & \text{if } v_i \notin \mathcal{V}_2^y \text{ and } v_j \notin \mathcal{V}_2^y \end{cases}$$

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{V}_1^y|} \sum_{v_i \in \mathcal{V}_1^y} \text{CE}(f'(\widetilde{\boldsymbol{A}}^{y,\text{pos}}, \widetilde{\boldsymbol{X}})_i, \boldsymbol{y}_i) \qquad \mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{V}_2^y|} \sum_{v_i \in \mathcal{V}_2^y} \text{CE}(f'(\widetilde{\boldsymbol{A}}^{y,\text{ neg}}, \widetilde{\boldsymbol{X}})_i, \boldsymbol{y}_i). \qquad \mathcal{L} = \mathcal{L}_{\text{pos}} - \log(\mathcal{L}_{\text{neg}})$$

- Target Label

$$\boldsymbol{h}^y = \frac{1}{|\mathcal{V}_1^y|} \sum_{v_i \in \mathcal{V}_1^y} f'(\widetilde{\boldsymbol{A}}^{y,\text{pos}}, \widetilde{\boldsymbol{X}}; \phi_1')_i \qquad s^y = \frac{1}{|\mathcal{V}_2^y|} \sum_{v_i \in \mathcal{V}_2^y} \left\| f'(\widetilde{\boldsymbol{A}}^{y,\text{neg}}, \widetilde{\boldsymbol{X}}; \phi_1')_i - \boldsymbol{h}^y \right\|_2^2$$

$$\mathrm{y} = \left\{ y \mid y \in [1, C] \wedge \frac{|s^y - \text{median}(\boldsymbol{s})|}{\text{mad}(\boldsymbol{s})} \leq \beta \right\}$$

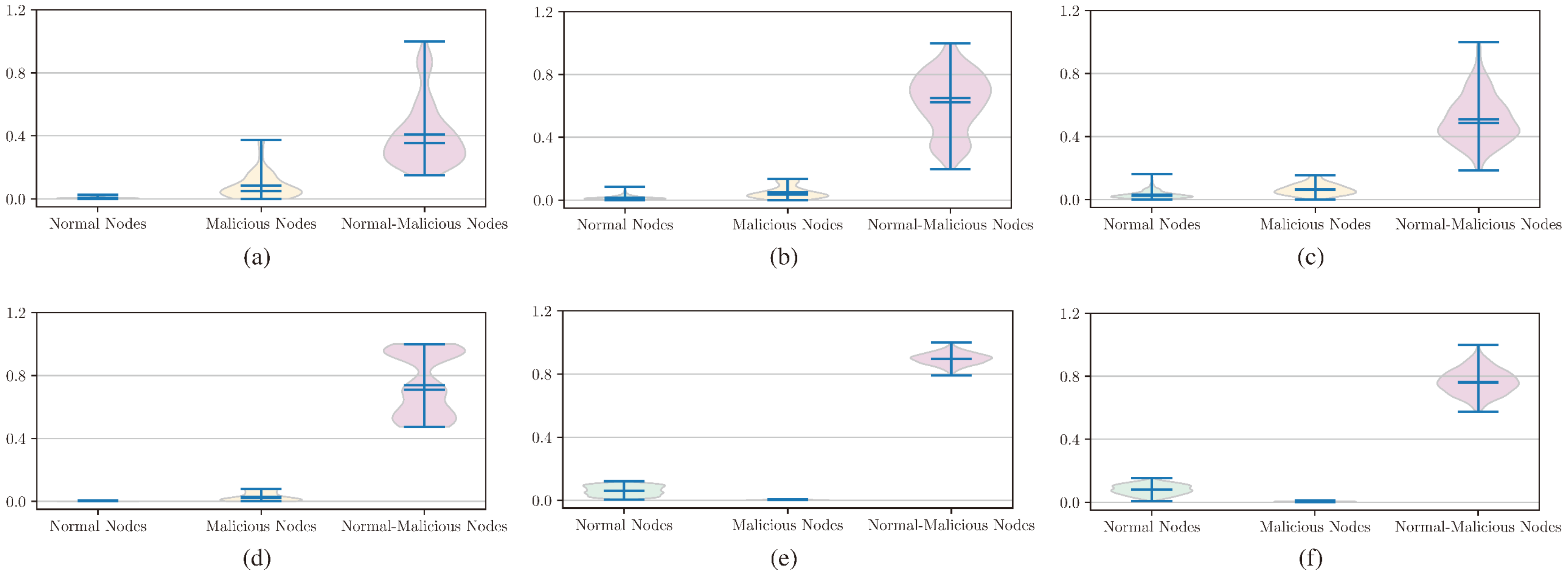
➤Backdoor-free Training

$$\mathcal{L}_{\text{clf}} = \frac{1}{|\mathcal{V}^{\dagger}|} \sum_{v_i \in \mathcal{V}^{\dagger}} \text{CE}(f(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}})_i, \boldsymbol{y}_i) \qquad \mathcal{L}_{\text{penalty}} = \frac{1}{|\mathcal{V}^{\ddagger}|} \sum_{v_i \in \mathcal{V}^{\ddagger}} \text{CE}(f(\widetilde{\boldsymbol{A}}', \widetilde{\boldsymbol{X}})_i, \boldsymbol{y}_i) \qquad \mathcal{L} = \mathcal{L}_{\text{clf}} - \gamma \log \mathcal{L}_{\text{penalty}}$$

Analysis on Discrepency Matrices

➢Elements of Discrepancy Matrices

- Differences in elements of two discrepancy matrices between normal and malicious nodes in the Cora dataset.



Effectiveness Evaluation

➢Seven DLBAs & Four Datasets

Datasets	Defenses	SBA [50]		GTA [38]		EBA [43]		GB-FGSM [4]		LGCB [4]		UGBA [5]		TRAP [44]	
		ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑
Cora	*no-defense*	53.14$_{\pm2.10}$	84.07$_{\pm1.02}$	96.22$_{\pm4.41}$	82.74$_{\pm2.38}$	71.81$_{\pm9.44}$	**83.48**$_{\pm1.32}$	97.34$_{\pm1.64}$	82.52$_{\pm1.53}$	97.79$_{\pm1.01}$	83.70$_{\pm0.52}$	97.51$_{\pm2.07}$	83.03$_{\pm1.32}$	19.86$_{\pm2.42}$	**84.15**$_{\pm1.09}$
	Prune [5]	30.26$_{\pm2.78}$	83.26$_{\pm0.96}$	16.42$_{\pm2.35}$	**84.00**$_{\pm1.80}$	100.0$_{\pm0.00}$	80.74$_{\pm1.57}$	56.98$_{\pm1.00}$	81.92$_{\pm3.36}$	56.98$_{\pm0.93}$	81.92$_{\pm1.95}$	54.47$_{\pm2.19}$	83.41$_{\pm1.52}$	11.36$_{\pm2.08}$	81.48$_{\pm1.87}$
	Isolate [5]	04.06$_{\pm2.42}$	37.19$_{\pm1.07}$	03.54$_{\pm1.85}$	77.78$_{\pm1.48}$	96.43$_{\pm1.66}$	80.67$_{\pm1.29}$	56.09$_{\pm0.78}$	81.85$_{\pm1.08}$	56.46$_{\pm0.78}$	81.63$_{\pm1.97}$	56.02$_{\pm0.80}$	80.74$_{\pm3.21}$	**10.15**$_{\pm1.94}$	67.26$_{\pm1.63}$
	PXGBD [9]	46.37$_{\pm2.03}$	84.00$_{\pm0.85}$	97.05$_{\pm2.46}$	83.40$_{\pm1.74}$	71.04$_{\pm9.65}$	82.74$_{\pm1.40}$	97.86$_{\pm1.44}$	82.74$_{\pm0.93}$	98.38$_{\pm1.64}$	82.37$_{\pm1.13}$	97.86$_{\pm2.17}$	**84.00**$_{\pm0.55}$	22.14$_{\pm2.31}$	81.55$_{\pm1.37}$
	GXGBD [9]	56.28$_{\pm1.31}$	**84.17**$_{\pm1.30}$	83.77$_{\pm1.56}$	83.42$_{\pm1.10}$	88.19$_{\pm2.81}$	73.52$_{\pm2.75}$	98.16$_{\pm1.16}$	**84.07**$_{\pm1.09}$	98.43$_{\pm1.01}$	**84.26**$_{\pm0.98}$	80.81$_{\pm2.30}$	83.15$_{\pm2.31}$	19.70$_{\pm3.84}$	81.18$_{\pm0.80}$
	ABL [21]	54.86$_{\pm2.37}$	82.81$_{\pm1.40}$	97.79$_{\pm2.09}$	82.59$_{\pm2.38}$	81.43$_{\pm3.98}$	83.26$_{\pm1.37}$	97.56$_{\pm1.34}$	83.11$_{\pm0.81}$	98.08$_{\pm0.84}$	83.70$_{\pm0.59}$	97.42$_{\pm1.04}$	83.89$_{\pm2.19}$	18.38$_{\pm2.52}$	83.33$_{\pm0.64}$
	RS [49]	08.12$_{\pm1.51}$	80.81$_{\pm1.03}$	94.83$_{\pm2.93}$	80.67$_{\pm0.99}$	58.92$_{\pm2.72}$	81.11$_{\pm1.05}$	97.27$_{\pm0.85}$	80.15$_{\pm1.24}$	96.97$_{\pm1.09}$	80.15$_{\pm1.84}$	90.41$_{\pm1.56}$	80.81$_{\pm1.03}$	20.81$_{\pm1.78}$	79.85$_{\pm1.13}$
	DShield	**01.33**$_{\pm2.56}$	81.78$_{\pm1.93}$	**00.74**$_{\pm0.74}$	81.92$_{\pm1.98}$	**02.95**$_{\pm2.32}$	81.50$_{\pm1.08}$	**02.51**$_{\pm2.83}$	82.29$_{\pm0.71}$	**00.89**$_{\pm1.15}$	82.57$_{\pm0.51}$	**01.33**$_{\pm1.21}$	82.15$_{\pm0.80}$	13.38$_{\pm2.12}$	82.07$_{\pm1.65}$
PubMed	*no-defense*	62.84$_{\pm5.05}$	85.17$_{\pm0.20}$	96.54$_{\pm2.28}$	85.09$_{\pm0.26}$	84.79$_{\pm2.10}$	85.23$_{\pm0.17}$	99.67$_{\pm0.66}$	85.03$_{\pm0.15}$	97.14$_{\pm1.44}$	85.16$_{\pm0.22}$	93.27$_{\pm3.23}$	85.20$_{\pm0.25}$	48.34$_{\pm0.93}$	85.21$_{\pm0.38}$
	Prune [5]	53.01$_{\pm2.59}$	**85.33**$_{\pm0.37}$	46.15$_{\pm0.40}$	**85.30**$_{\pm0.21}$	99.95$_{\pm0.00}$	**85.44**$_{\pm0.18}$	70.28$_{\pm0.40}$	**85.31**$_{\pm0.25}$	69.36$_{\pm0.56}$	**85.36**$_{\pm0.18}$	66.85$_{\pm1.47}$	**85.38**$_{\pm0.29}$	**41.77**$_{\pm0.28}$	**85.33**$_{\pm0.25}$
	Isolate [5]	40.69$_{\pm0.47}$	81.25$_{\pm0.28}$	45.46$_{\pm0.22}$	84.47$_{\pm0.21}$	99.36$_{\pm0.38}$	84.56$_{\pm0.22}$	70.16$_{\pm0.31}$	85.14$_{\pm0.15}$	69.80$_{\pm0.90}$	85.11$_{\pm0.13}$	67.62$_{\pm1.34}$	85.14$_{\pm0.20}$	43.30$_{\pm3.41}$	83.34$_{\pm1.59}$
	PXGBD [9]	55.26$_{\pm3.06}$	83.55$_{\pm1.01}$	99.67$_{\pm0.54}$	83.17$_{\pm0.92}$	86.26$_{\pm5.89}$	82.64$_{\pm1.57}$	100.0$_{\pm0.00}$	82.32$_{\pm1.47}$	100.0$_{\pm0.00}$	82.38$_{\pm0.99}$	91.80$_{\pm0.87}$	82.61$_{\pm1.35}$	49.80$_{\pm1.70}$	81.98$_{\pm1.45}$
	GXGBD [9]	56.20$_{\pm1.18}$	83.93$_{\pm0.76}$	99.71$_{\pm0.35}$	82.57$_{\pm0.64}$	90.21$_{\pm2.27}$	75.45$_{\pm2.23}$	100.0$_{\pm0.00}$	82.31$_{\pm0.85}$	100.0$_{\pm0.00}$	82.41$_{\pm0.82}$	81.95$_{\pm0.00}$	82.82$_{\pm0.91}$	49.36$_{\pm3.16}$	82.84$_{\pm1.47}$
	ABL [21]	65.16$_{\pm4.55}$	85.21$_{\pm0.38}$	99.98$_{\pm0.03}$	85.13$_{\pm0.21}$	89.08$_{\pm2.61}$	84.98$_{\pm0.16}$	99.98$_{\pm0.03}$	84.89$_{\pm0.13}$	100.0$_{\pm0.00}$	85.04$_{\pm0.17}$	91.45$_{\pm5.44}$	85.00$_{\pm0.30}$	49.36$_{\pm0.68}$	84.64$_{\pm1.01}$
	RS [49]	51.10$_{\pm3.63}$	83.13$_{\pm0.45}$	97.53$_{\pm1.33}$	82.99$_{\pm0.37}$	89.86$_{\pm1.20}$	83.68$_{\pm0.32}$	100.0$_{\pm0.00}$	83.39$_{\pm0.58}$	100.0$_{\pm0.00}$	83.40$_{\pm0.57}$	59.68$_{\pm4.13}$	83.17$_{\pm0.31}$	48.48$_{\pm3.50}$	82.83$_{\pm0.71}$
	DShield	**35.91**$_{\pm3.43}$	82.01$_{\pm0.47}$	**00.73**$_{\pm0.33}$	85.28$_{\pm0.12}$	**02.78**$_{\pm1.58}$	85.02$_{\pm0.29}$	**00.66**$_{\pm0.72}$	83.71$_{\pm0.24}$	**03.54**$_{\pm1.97}$	84.11$_{\pm0.56}$	**02.24**$_{\pm0.97}$	84.70$_{\pm0.42}$	45.77$_{\pm0.76}$	82.17$_{\pm0.36}$

Effectiveness Evaluation

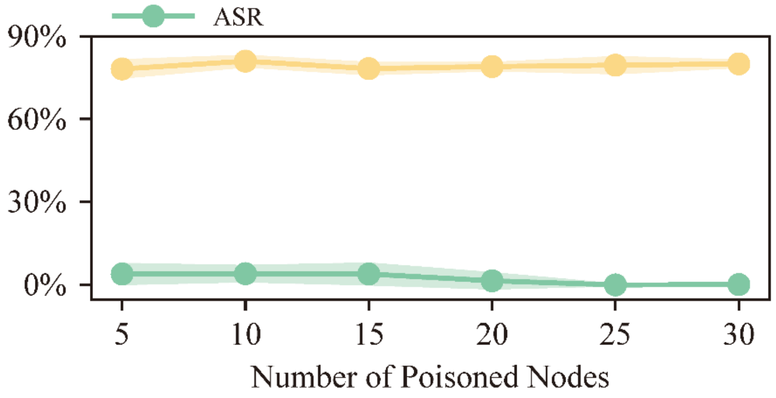
➢Two CLBAs & Four Datasets

Defenses	Cora				PubMed				Flickr				OGBN-arXiv			
	GCBA [49]		PerCBA [45]		GCBA [49]		PerCBA [45]		GCBA [49]		PerCBA [45]		GCBA [49]		PerCBA [45]	
	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑
no-defense	$81.74_{\pm3.93}$	**84.74**$_{\pm1.27}$	$84.87_{\pm3.13}$	$83.18_{\pm1.54}$	$78.86_{\pm7.76}$	$85.30_{\pm0.13}$	$100.0_{\pm0.00}$	$85.23_{\pm0.13}$	$60.48_{\pm4.72}$	**50.84**$_{\pm0.16}$	$07.05_{\pm5.11}$	**50.86**$_{\pm0.18}$	$26.84_{\pm2.79}$	**60.43**$_{\pm0.30}$	$18.06_{\pm1.48}$	$59.81_{\pm0.15}$
Prune [5]	$99.70_{\pm0.31}$	$82.37_{\pm0.42}$	$69.01_{\pm1.31}$	$82.66_{\pm0.55}$	$99.83_{\pm0.18}$	**85.55**$_{\pm0.16}$	$95.58_{\pm4.50}$	**85.34**$_{\pm0.30}$	$78.50_{\pm3.63}$	$49.97_{\pm0.27}$	$30.38_{\pm2.70}$	$50.12_{\pm0.15}$	$23.14_{\pm0.30}$	$60.40_{\pm0.35}$	$19.03_{\pm1.15}$	$59.78_{\pm0.17}$
Isolate [5]	$99.88_{\pm0.21}$	$71.92_{\pm2.05}$	$81.73_{\pm2.35}$	$69.04_{\pm2.07}$	$99.10_{\pm1.09}$	$84.12_{\pm0.14}$	$97.88_{\pm1.05}$	$84.26_{\pm1.01}$	$66.44_{\pm5.40}$	$50.34_{\pm0.08}$	$30.15_{\pm3.17}$	$50.52_{\pm0.18}$	$22.24_{\pm0.98}$	**60.43**$_{\pm0.08}$	$19.00_{\pm0.64}$	$59.75_{\pm0.18}$
PXGBD [8]	$82.29_{\pm3.66}$	$82.89_{\pm0.88}$	$27.68_{\pm2.58}$	$82.74_{\pm1.19}$	$96.91_{\pm3.08}$	$82.82_{\pm1.34}$	$93.11_{\pm0.36}$	$83.31_{\pm1.54}$	$26.50_{\pm0.79}$	$47.74_{\pm0.98}$	$00.16_{\pm0.28}$	$47.29_{\pm2.67}$	$11.87_{\pm0.71}$	$60.42_{\pm0.17}$	$00.78_{\pm0.49}$	**60.42**$_{\pm0.89}$
GXGBD [8]	$04.61_{\pm1.77}$	$82.96_{\pm0.52}$	$80.69_{\pm2.98}$	**84.30**$_{\pm1.07}$	$97.30_{\pm0.51}$	$82.43_{\pm1.11}$	$72.90_{\pm1.18}$	$82.78_{\pm0.32}$	$49.90_{\pm2.39}$	$40.57_{\pm0.30}$	$02.42_{\pm4.19}$	$40.61_{\pm0.44}$	$09.21_{\pm1.54}$	$55.94_{\pm0.16}$	$00.34_{\pm0.14}$	$48.51_{\pm0.43}$
ABL [20]	$79.70_{\pm3.89}$	$84.30_{\pm1.35}$	$88.68_{\pm4.25}$	$83.70_{\pm0.91}$	$83.40_{\pm1.46}$	$84.98_{\pm0.23}$	$82.50_{\pm3.53}$	$85.13_{\pm0.36}$	$66.88_{\pm0.84}$	$50.49_{\pm0.21}$	$01.35_{\pm0.60}$	$50.68_{\pm0.04}$	$02.73_{\pm3.84}$	$58.40_{\pm1.61}$	$22.04_{\pm2.39}$	$60.23_{\pm0.79}$
RS [49]	$42.44_{\pm4.11}$	$81.26_{\pm1.92}$	$85.22_{\pm3.15}$	$80.74_{\pm1.14}$	$74.05_{\pm1.70}$	$83.28_{\pm0.58}$	$88.46_{\pm3.41}$	$83.51_{\pm0.27}$	**00.02**$_{\pm0.03}$	$40.48_{\pm0.09}$	$23.72_{\pm0.59}$	$40.50_{\pm0.16}$	$27.76_{\pm0.08}$	$55.36_{\pm0.10}$	$43.48_{\pm5.32}$	$55.38_{\pm0.19}$
DShield	**01.11**$_{\pm1.04}$	$81.18_{\pm1.32}$	**03.41**$_{\pm2.53}$	$81.63_{\pm2.04}$	**01.91**$_{\pm1.66}$	$84.24_{\pm0.24}$	**08.09**$_{\pm2.04}$	$83.69_{\pm2.01}$	$01.66_{\pm2.34}$	$50.42_{\pm0.25}$	**00.04**$_{\pm0.05}$	**50.86**$_{\pm0.06}$	**00.00**$_{\pm0.00}$	$59.96_{\pm0.54}$	**00.00**$_{\pm0.00}$	$59.92_{\pm0.44}$

- DShield exhibits a **notable** performance against most backdoor attacks, surpassing the efficacy of conventional defenses
- Although DShield demonstrates superior defense performance in most scenarios, a **slight decline** in model performance on normal nodes is observed

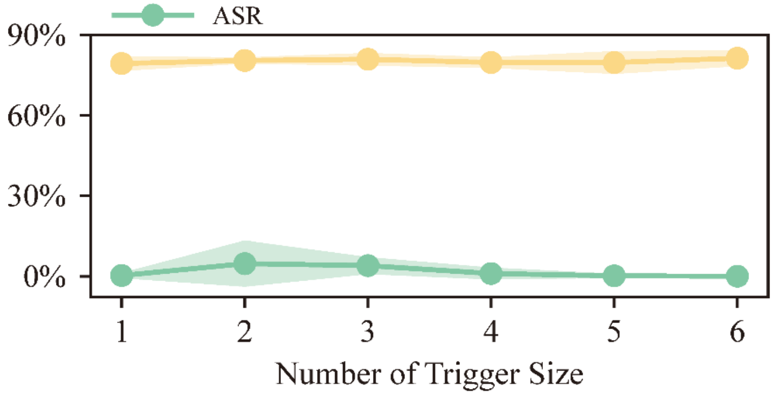
Robustness Evaluation

➢Six Poisoning Rates



(a) UGBA

➢Six Trigger Sizes



(a) UGBA

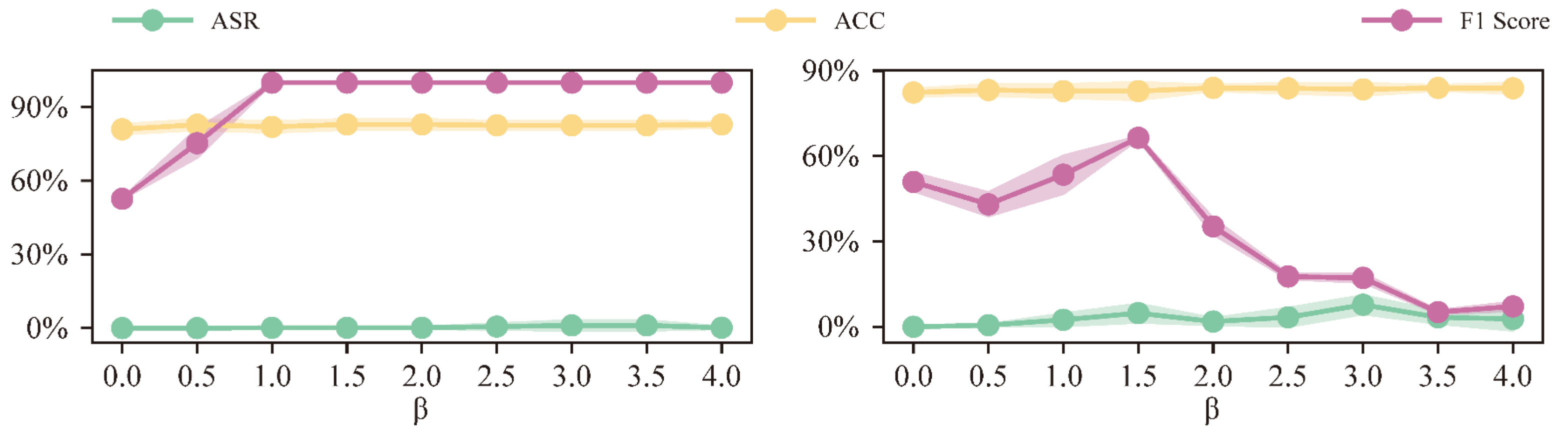
➢Five Adaptive Attacks

Datasets	Attacks	*no-defense*		*no-poison*		DShield	
		ASR ↓	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC ↑
Cora	UGBA+LGCB	$97.91_{\pm 0.93}$	$84.32_{\pm 1.19}$	$08.00_{\pm 0.21}$	$78.02_{\pm 5.00}$	$00.49_{\pm 4.16}$	$82.22_{\pm 6.72}$
	UGBA+GCBA	$88.07_{\pm 1.89}$	$83.83_{\pm 2.04}$	$08.00_{\pm 0.43}$	$78.15_{\pm 4.98}$	$03.94_{\pm 4.49}$	$81.97_{\pm 5.46}$
	GCBA+PerCBA	$74.78_{\pm 3.77}$	$83.82_{\pm 1.13}$	$08.24_{\pm 0.93}$	$77.53_{\pm 3.36}$	$08.49_{\pm 2.82}$	$80.99_{\pm 4.64}$
	AdaDA	$84.51_{\pm 2.61}$	$84.07_{\pm 0.52}$	$07.75_{\pm 0.64}$	$84.07_{\pm 1.29}$	$04.43_{\pm 3.13}$	$76.86_{\pm 0.26}$
	AdaCA	$89.85_{\pm 6.52}$	$82.59_{\pm 0.52}$	$09.47_{\pm 5.18}$	$84.69_{\pm 0.94}$	$00.74_{\pm 0.98}$	$81.23_{\pm 2.46}$
PubMed	UGBA+LGCB	$88.73_{\pm 5.16}$	$84.98_{\pm 0.26}$	$37.92_{\pm 1.71}$	$82.23_{\pm 1.97}$	$00.15_{\pm 0.27}$	$82.50_{\pm 0.74}$
	UGBA+GCBA	$91.89_{\pm 3.07}$	$85.42_{\pm 0.03}$	$39.61_{\pm 4.17}$	$83.66_{\pm 0.43}$	$02.92_{\pm 1.17}$	$82.51_{\pm 1.19}$
	GCBA+PerCBA	$89.01_{\pm 5.25}$	$84.53_{\pm 0.62}$	$39.52_{\pm 3.38}$	$83.49_{\pm 0.61}$	$13.39_{\pm 2.08}$	$81.28_{\pm 1.17}$
	AdaDA	$88.72_{\pm 3.34}$	$85.22_{\pm 0.04}$	$40.23_{\pm 0.72}$	$83.17_{\pm 3.57}$	$32.61_{\pm 8.89}$	$79.17_{\pm 1.47}$
	AdaCA	$87.88_{\pm 5.12}$	$85.13_{\pm 0.31}$	$03.82_{\pm 1.68}$	$85.39_{\pm 0.13}$	$05.68_{\pm 0.79}$	$82.60_{\pm 1.87}$

- The efficacy of DShield stems from the execution of the self-supervised learning framework and UMAP on the manipulated graph

Sensitivity Analysis

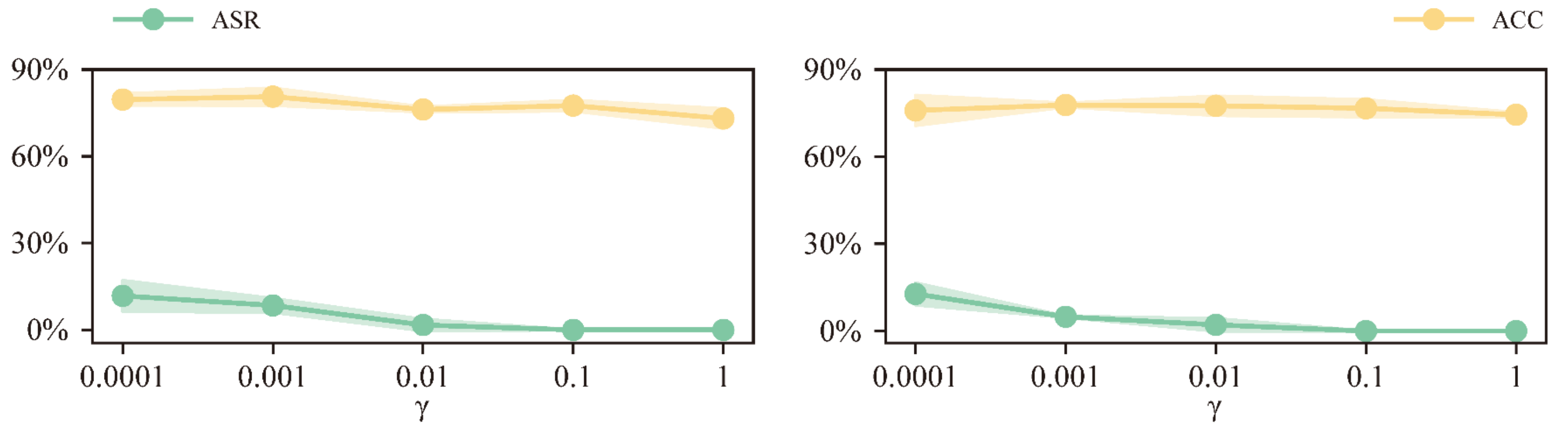
➢Parameter β



(a) UGBA (b) GCBA

- The parameter β significantly affects identification precision

➢Parameter γ



(a) UGBA (b) GCBA

- Increasing the parameter γ results in a decline in the victim model's performance on poisoned nodes

Extension to Graph Classification Tasks

➢Defending against Attacks on Graph Classification Tasks

Datasets	Defenses	G-SBA ASR ↓	G-SBA ACC ↑	G-EBA ASR ↓	G-EBA ACC ↑	G-GCBA ASR ↓	G-GCBA ACC ↑
ENZYMES	*no-defense*	$100.0_{\pm 0.00}$	$32.67_{\pm 0.94}$	$100.0_{\pm 0.00}$	$27.50_{\pm 3.54}$	$100.0_{\pm 0.00}$	$23.33_{\pm 0.00}$
	G-DShield	$01.67_{\pm 2.35}$	$35.00_{\pm 2.36}$	$02.50_{\pm 3.54}$	$31.67_{\pm 2.35}$	$00.00_{\pm 0.00}$	$30.84_{\pm 1.18}$
PROTEINS	*no-defense*	$100.0_{\pm 0.00}$	$66.67_{\pm 1.27}$	$99.56_{\pm 0.63}$	$71.62_{\pm 3.18}$	$99.11_{\pm 0.00}$	$73.43_{\pm 5.73}$
	G-DShield	$00.00_{\pm 0.00}$	$74.63_{\pm 4.04}$	$04.29_{\pm 6.06}$	$74.33_{\pm 3.19}$	$00.98_{\pm 1.39}$	$71.17_{\pm 1.27}$
MNIST	*no-defense*	$100.0_{\pm 0.00}$	$36.33_{\pm 0.67}$	$100.0_{\pm 0.00}$	$36.89_{\pm 0.02}$	$100.0_{\pm 0.00}$	$38.34_{\pm 0.66}$
	G-DShield	$00.00_{\pm 0.00}$	$38.86_{\pm 0.37}$	$01.87_{\pm 2.64}$	$39.77_{\pm 1.65}$	$00.00_{\pm 0.00}$	$40.57_{\pm 1.89}$

- Experimental results show that semantic drift and attribute over-emphasis also occur in graph classification tasks and the defense efficiency is not limited to node classification backdoor attacks

Thank You!

Hao Yu

csyuhao@gmail.com

National University of Defense Technology

Source Code： https://github.com/csyuhao/DShield