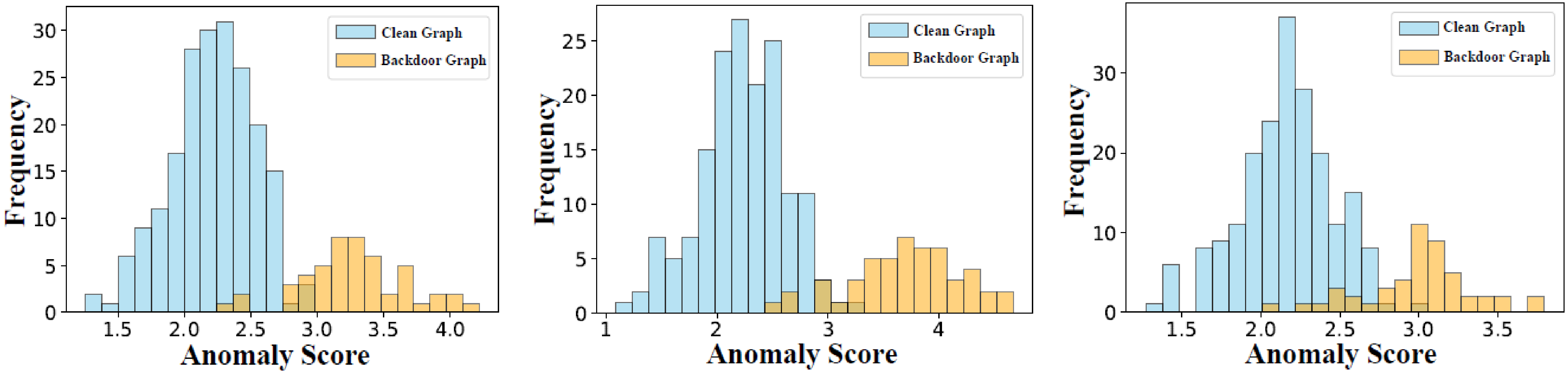
Stealthy Yet Effective: Distribution-Preserving Backdoor Attacks on Graph Classification

Xiaobao Wang, Ruoxiao Sun, Yujun Zhang, Bingdao Feng, Dongxiao He, Luzhi Wang, Di Jin

Introduction

- Graph Neural Networks (GNNs) have demonstrated strong performance across tasks, but remain vulnerable to backdoor attacks.
- Most existing graph backdoor studies focus on node classification. However, graph classification poses a fundamentally different and more complex challenge.
- Recent backdoor attacks on graph classification introduce obvious out-of-distribution (OOD) artifacts, which significantly compromise stealth and limit their practicality in real-world settings.

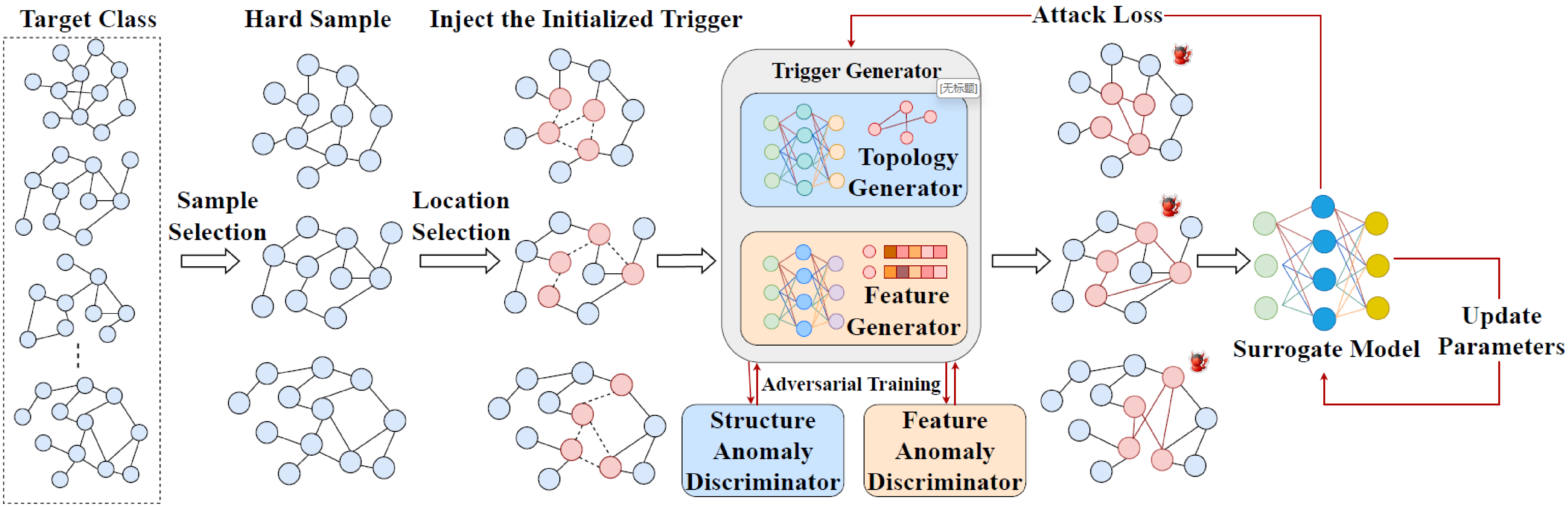


(a) ER-B (b) GTA (c) Motif

Introduction

- **Two Types of Deviations:**
  - **Structural Deviation:** Triggered by the injection of rare or unnatural subgraphs (e.g., low-frequency motifs) that diverge from the structural distribution of clean graphs.
  - **Semantic Deviation:** Caused by label flipping, this introduces a discrepancy between a graph's assigned class and its inherent structure.
- **Key Challenge:**
  - Can we design a graph-level backdoor attack that preserves the distributional properties of clean samples, avoids label manipulation, and remains both effective and stealthy?
- **Our Solution:**
  - We propose DPSBA, which utilizes clean-label setting and distribution-aware discriminator to achieve a balance between effective and stealthy.

Methodology



Methodology

- **Hard Sample Selection:**
  - Hard Samples: Samples from the target class that the model finds uncertain.

$$\mathrm{cfd}(G) = \mathrm{softmax}(f_\theta(G))_{y_t} = \frac{e^{f_\theta(G)_{y_t}}}{\sum_{j=1}^{K} e^{f_\theta(G)_j}}$$

  - We select the bottom p% of target-class graphs with the lowest cfd(G) scores as poisoned samples.

- **Trigger Location Selection:**
  - Select high degrees nodes as candidates.
  - Identify the *M* most influential nodes.

$$S(v) = |f_\theta(G + \Delta_v) - f_\theta(G)|$$

- **Trigger Generation and Injection:**
  - Topology Generator $\mathbf{H}' = \sigma(W_1\mathbf{H} + b_1)$
  - Feature Generator $\mathbf{X}' = \sigma(W_2\mathbf{X} + b_2)$

Methodology

- **Trigger Optimization:**
  - **Attack Effectiveness** $\mathcal{L}_{atk} = -\log f_{\theta^*}(G_{g_t})_{y_t}$
  - **Stealthiness via Adversarial Anomaly Minimization**
    - The topology discriminator is a GCN
    - The feature discriminator is an MLP

$$\min_{\omega_t} \max_{\theta_t} \mathcal{L}_d^{(t)} = \sum_{G \sim \mathcal{G}_c} \log D_{\theta_t}(G) + \sum_{G \sim \mathcal{G}_b} \log(1 - D_{\theta_t}(G_{g_t}(\omega_t))),$$

$$\min_{\omega_f} \max_{\theta_f} \mathcal{L}_d^{(f)} = \sum_{G \sim \mathcal{G}_c} \log D_{\theta_f}(G) + \sum_{G \sim \mathcal{G}_b} \log(1 - D_{\theta_f}(G_{g_t}(\omega_f))),$$

  - **Joint Training Objectives**

$$\min_{\omega_t} \sum_{G \in \mathcal{G}_b} \mathcal{L}_{atk}(G_{g_t}(\omega_t)) + \alpha \mathcal{L}_d^{(t)}(D_{\theta_t}(G_{g_t}(\omega_t))), \quad \text{s.t. } \theta^* = \arg\min_{\theta} \mathcal{L}_{train}(f_\theta(C))$$

$$\min_{\omega_f} \sum_{G \in \mathcal{G}_b} \mathcal{L}_{atk}(G_{g_t}(\omega_f)) + \beta \mathcal{L}_d^{(f)}(D_{\theta_f}(G_{g_t}(\omega_f))), \quad \text{s.t. } \theta^* = \arg\min_{\theta} \mathcal{L}_{train}(f_\theta(C))$$

Experiment

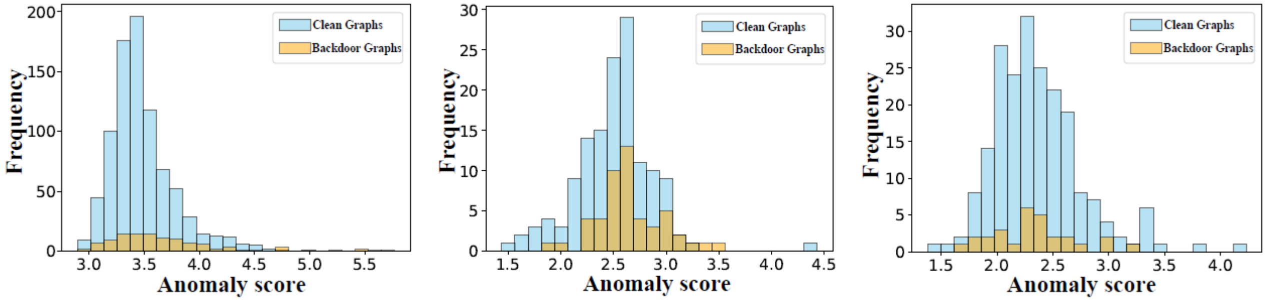
- **Main Experiment**

Table 1: Comparison results between DPSBA and each baseline model

Datasets	Surrogate Model	Metrics	ER-B	LIA	GTA	Motif	Motif-S	Ours
PROTEINS_full	GCN	ASR (%)	51.53	68.35	73.16	70.91	48.56	**73.93**
		CAD (%)	4.73	4.70	5.14	5.92	4.66	**4.62**
		AUC (%)	70.04	71.01	78.20	79.16	64.72	**60.11**
	GIN	ASR (%)	62.53	58.77	80.96	79.08	63.01	**87.91**
		CAD (%)	4.88	4.36	4.57	4.97	**4.33**	4.92
		AUC (%)	79.65	71.74	79.96	80.06	70.49	**62.95**
	SAGPool	ASR (%)	65.38	64.81	94.04	71.35	57.09	**94.15**
		CAD (%)	4.26	5.02	3.65	3.36	3.94	**3.29**
		AUC (%)	71.34	76.89	78.57	82.75	81.81	**69.20**
AIDS	GCN	ASR (%)	85.38	85.49	93.21	92.69	56.08	**94.76**
		CAD (%)	4.53	3.80	5.14	4.12	4.03	**2.38**
		AUC (%)	98.08	97.22	99.34	99.71	89.43	**72.65**
	GIN	ASR (%)	93.99	95.56	97.52	**97.75**	56.8	95.87
		CAD (%)	2.69	2.03	2.65	2.28	2.51	**1.94**
		AUC (%)	99.98	99.20	99.34	99.71	94.29	**73.66**
	SAGPool	ASR (%)	59.26	62.66	86.99	87.65	62.89	**98.90**
		CAD (%)	1.65	1.79	3.77	2.64	2.44	**-0.40**
		AUC (%)	95.79	94.56	99.67	99.02	93.43	**77.23**
FRANKEN-STEIN	GCN	ASR (%)	63.60	61.04	**99.35**	80.57	59.24	98.37
		CAD (%)	1.71	1.56	2.74	1.15	3.96	**1.01**
		AUC (%)	80.41	75.66	100.00	89.64	69.23	**68.96**
	GIN	ASR (%)	92.06	82.63	98.65	92.87	58.68	**99.84**
		CAD (%)	3.60	2.35	1.95	2.44	1.75	**1.83**
		AUC (%)	85.73	76.15	91.06	87.54	**65.77**	73.46
	SAGPool	ASR (%)	68.15	90.18	95.23	84.56	52.29	**99.99**
		CAD (%)	4.78	4.66	4.64	4.61	6.86	**4.57**
		AUC (%)	64.89	77.50	80.46	87.29	60.98	**60.12**
ENZYMES	GCN	ASR (%)	26.09	30.43	95.33	21.74	15.21	**96.67**
		CAD (%)	4.17	4.99	3.00	4.99	**-1.67**	-0.67
		AUC (%)	68.32	66.15	71.20	71.35	66.22	**66.11**
	GIN	ASR (%)	37.83	27.02	96.00	16.21	12.16	**99.33**
		CAD (%)	9.17	10.00	2.67	8.33	4.17	**-0.33**
		AUC (%)	71.40	62.01	76.42	68.18	65.78	**41.20**
	SAGPool	ASR (%)	29.54	38.63	100.00	15.91	11.37	**100.00**
		CAD (%)	4.33	6.67	5.00	10.83	**3.33**	4.00
		AUC (%)	57.73	63.98	70.37	75.47	69.48	**49.91**

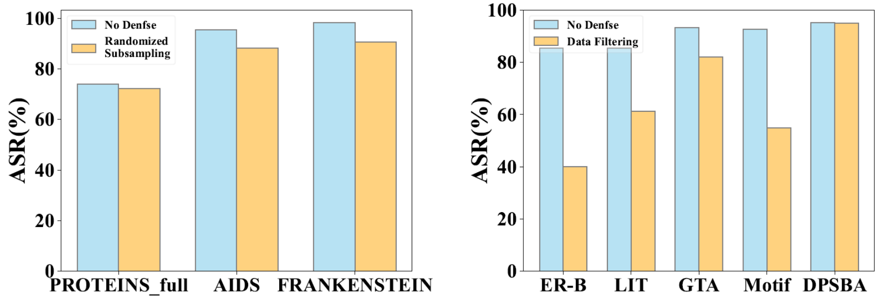
Table 2: Results of the transferability evaluation(%)

Surrogate model	Actual model	PROTEINS_full		AIDS		FRANKENSTEIN	
		ASR	CAD	ASR	CAD	ASR	CAD
GCN	GIN	81.32	4.79	99.44	1.01	98.37	0.03
	SAGPool	98.90	0.08	96.14	2.48	94.96	-0.10



(a) FRANKENSTEIN (b) AIDS (c) PROTEINS_full

Figure 3: Anomaly distribution visualization



(a) Randomized Subsampling (b) Data Filtering

Figure 4: Attack performance under defense

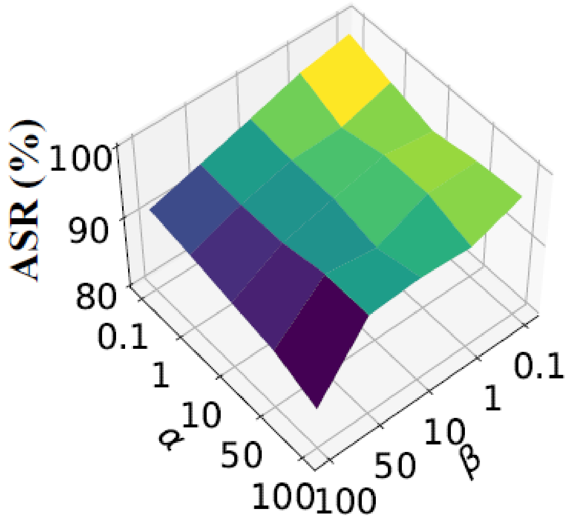
Experiment

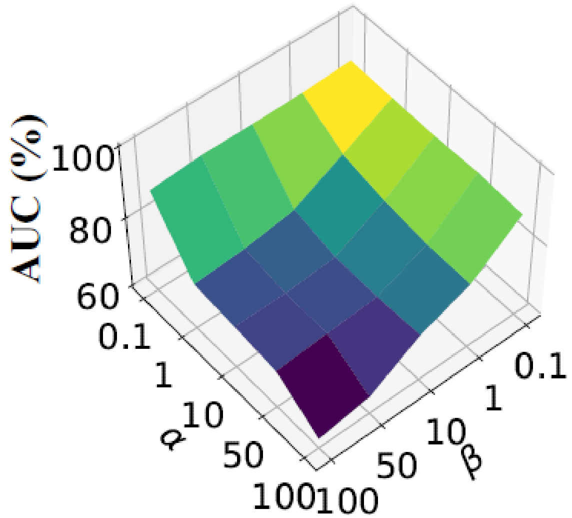
- **Ablation Experiment**

DPSBA/S	w/o hard sample selection
DPSBA/N	w/o position selection
DPSBA/F	w/o feature generator
DPSBA/T	w/o topology generator
DPSBA/OD	w/o adversarial training

Model	PROTEINS_full			AIDS		
	ASR	CAD	AUC	ASR	CAD	AUC
DPSBA	73.93	4.62	60.11	94.76	2.38	72.65
DPSBA/S	70.98	3.57	60.24	91.32	2.09	72.60
DPSBA/N	70.74	4.53	58.97	93.67	2.31	71.26
DPSBA/F	71.80	4.96	59.01	85.67	2.40	67.26
DPSBA/T	69.08	3.71	54.73	93.66	2.91	71.41
DPSBA/OD	90.88	4.90	90.23	99.46	3.54	93.72

- **Impact of the Loss Weights α and β**





The End, Thanks!