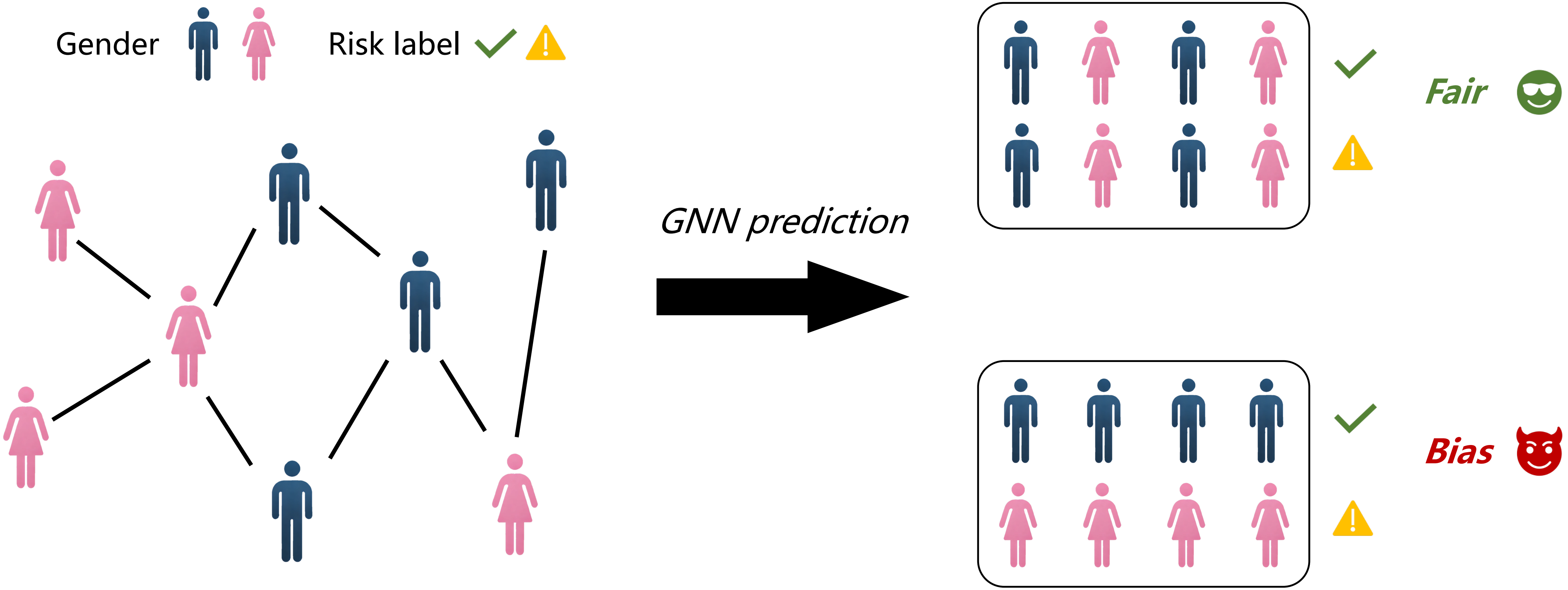
Are Your Models Still Fair? Fairness Attacks on Graph Neural Networks via Node Injections

Zihan Luo, Hong Huang, Yongkang Zhou, Jiping Zhang, Nuo Chen, Hai Jin

Huazhong University of Science and Technology

Background: Group Fairness in GNNs

- GNNs are powerful in graph representation learning, but face **fairness issues**.
- The prediction of GNNs should **be independent of sensitive attributes**, such as gender, region, age …



Background: Group Fairness in GNNs

- **Definition 1. Statistical Parity (SP).** The Statistical Parity requires the prediction probability distribution to be independent of sensitive attributes, i.e. for any class $y \in \mathcal{Y}$ and any node v $\in \mathcal{V}$:

$$\underbrace{|P(\hat{y}_v = y|s = 0) - P(\hat{y}_v = y|s = 1)|}_{\boldsymbol{\Delta}_{\mathbf{SP}}} = 0$$

- **Definition 2. Equal Opportunity (EO).** The Equal Opportunity requires that the probability of predicting **correctly** is independent of sensitive attributes, i.e. for any class $y \in \mathcal{Y}$ and any node v $\in \mathcal{V}$, we can have:

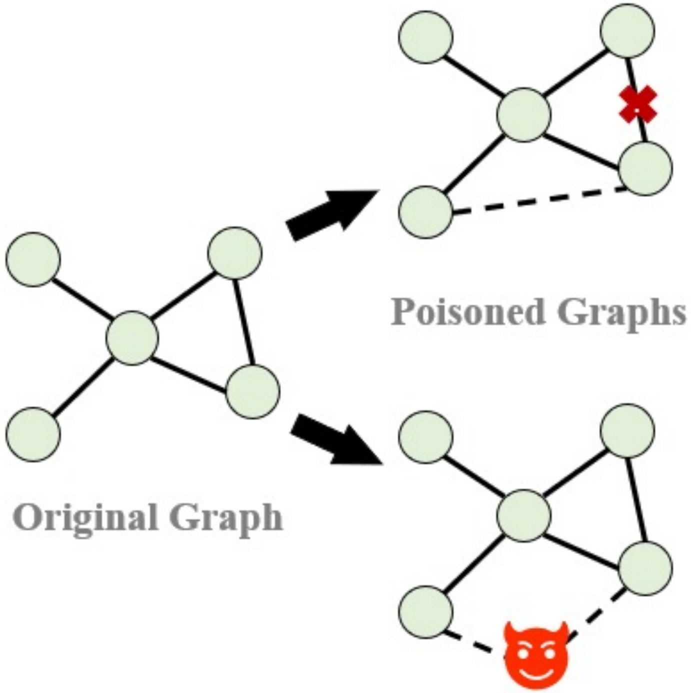
$$\underbrace{|P(\hat{y}_v = y|s = 0, y_v = y) - P(\hat{y}_v = y|s = 1, y_v = y)|}_{\boldsymbol{\Delta}_{\mathbf{EO}}} = 0$$

Motivation

- Many researchers have proposed effective fair GNN models, such as FairGNN[1], FairVGNN[2], EDITS[3] . But such fairness is actually vulnerable to adversarial attacks.

- Existing fairness attacks on GNNs need to **modify the connectivity between existing nodes**, which is hard and time-consuming in reality.

Can we launch a node-injection based fairness attack on GNNs?



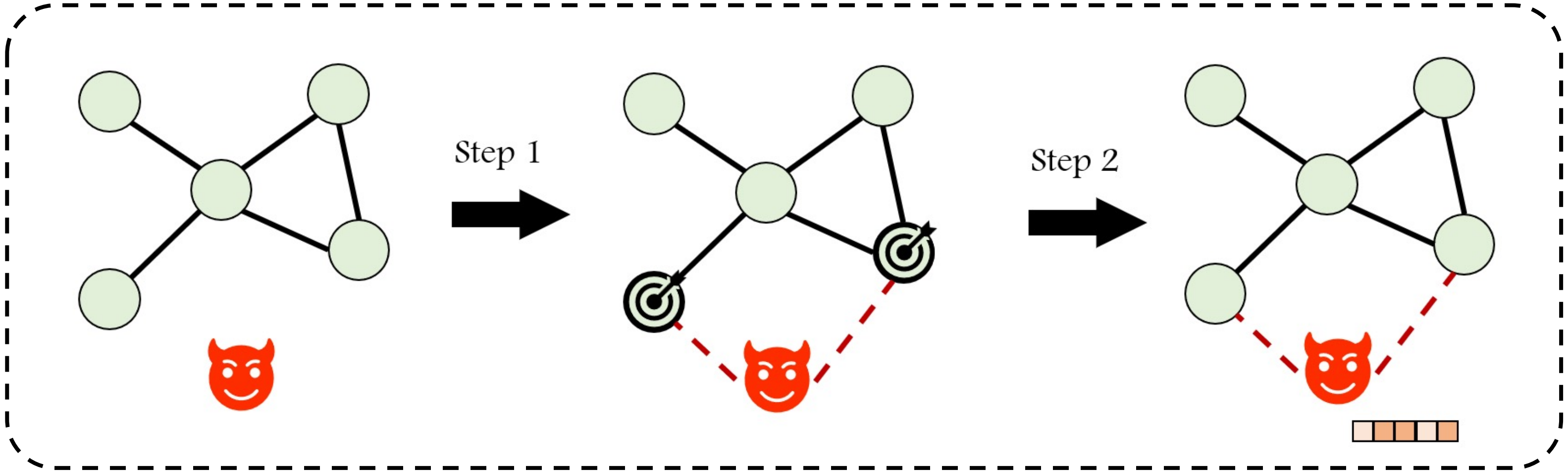
[1] Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information, WSDM 2021.
[2] Improving Fairness in Graph Neural Networks via Mitigating Sensitive Attribute Leakage, KDD 2022.
[3] EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks, WWW 2022.

NIFA – Node-Injection-based Fairness Attacks

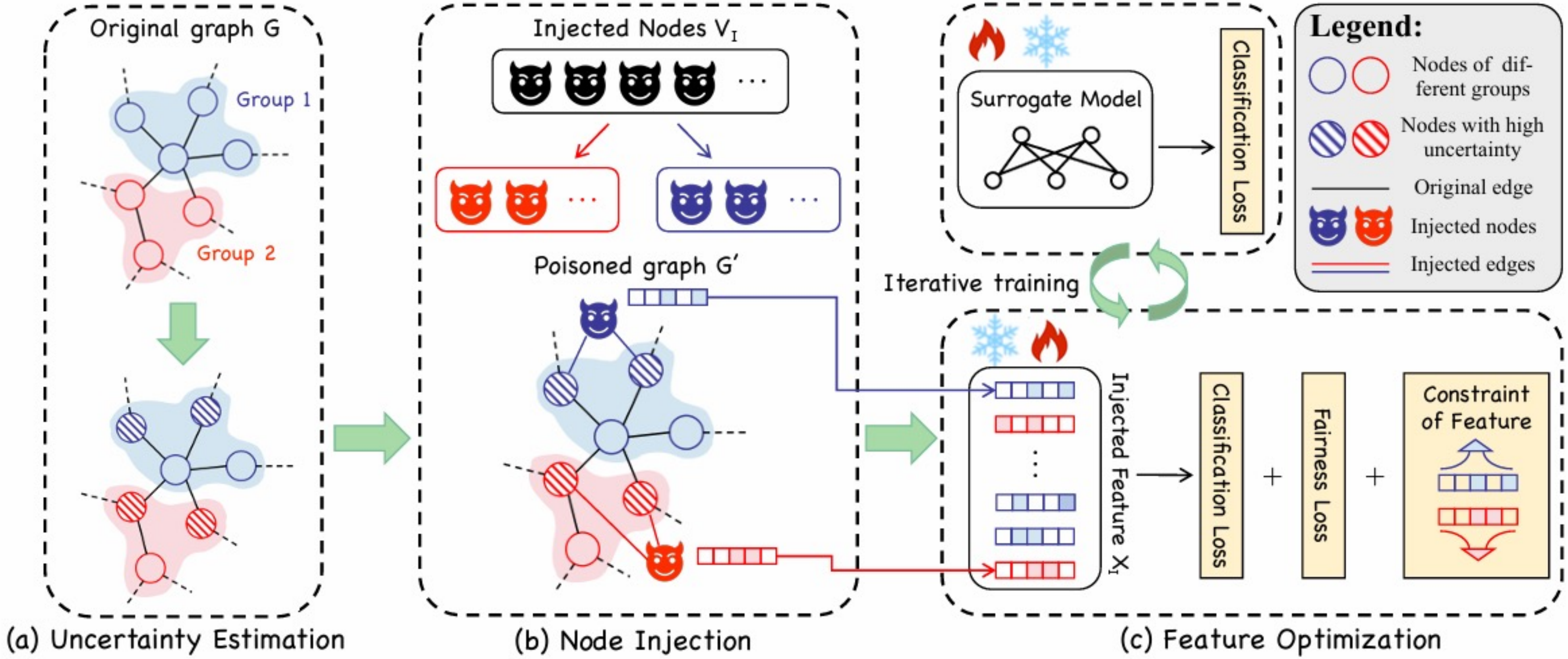
- Core idea: **Design multiple principles during the node injection, and then optimize the injected nodes' features.**

- We design two principles to guide the node injection:
  - Uncertainty-maximization principle
  - Homophily-increase principle

- Multiple objective functions are further designed for injected nodes' features.



Overview of NIFA



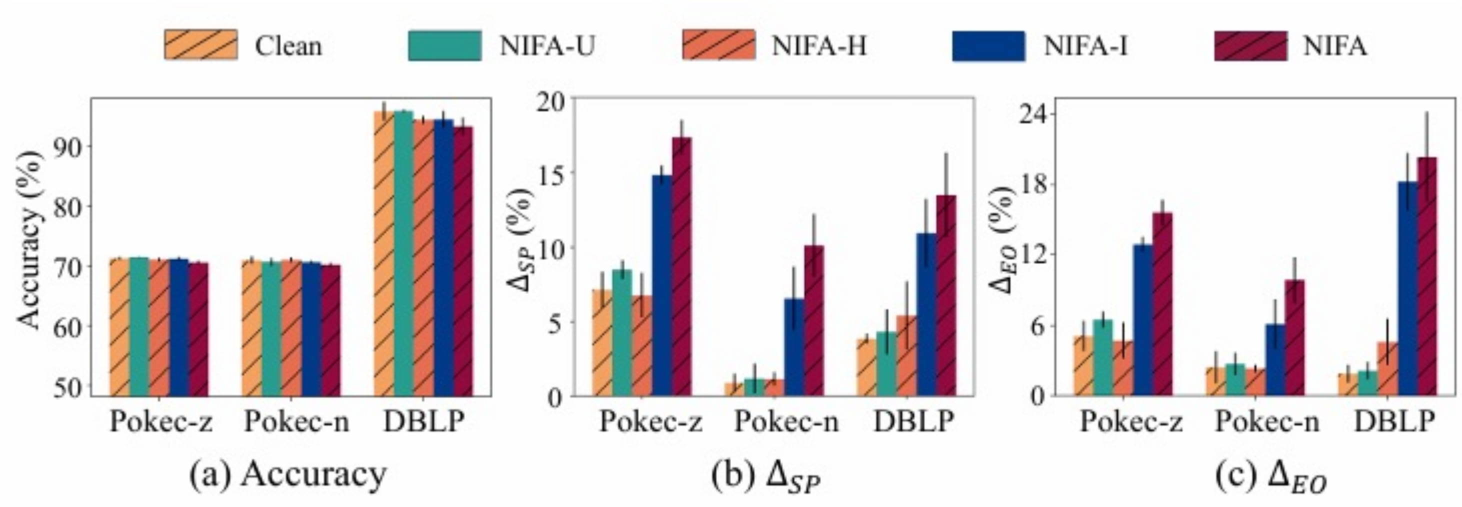
Attack Performance on GNNs

		Pokec-z			Pokec-n			DBLP		
		Accuracy	Δ_{SP}	Δ_{EO}	Accuracy	Δ_{SP}	Δ_{EO}	Accuracy	Δ_{SP}	Δ_{EO}
GCN	before	71.22±0.28	7.13±1.21	5.10±1.28	70.92±0.66	0.88±0.62	2.44±1.37	95.88±1.61	3.84±0.34	1.91±0.75
	after	70.50±0.30	**17.36±1.16**	**15.59±1.08**	70.12±0.37	**10.10±2.10**	**9.85±1.97**	93.37±1.48	**13.49±2.83**	**20.33±3.82**
GraphSAGE	before	70.79±0.62	4.29±0.84	3.46±1.12	68.77±0.34	1.65±1.31	2.34±1.04	96.58±0.29	4.27±1.09	2.78±0.91
	after	70.05±1.25	**6.20±1.63**	**4.20±1.77**	68.93±1.19	**3.32±1.88**	**3.56±1.91**	93.92±0.74	**10.16±2.24**	**16.65±3.30**
APPNP	before	69.79±0.42	6.83±1.25	5.07±1.26	68.73±0.64	3.39±0.28	3.71±0.28	96.58±0.38	3.98±1.18	2.20±1.08
	after	69.12±0.70	**18.44±1.41**	**16.85±1.50**	67.90±0.76	**13.47±3.22**	**13.52±3.56**	92.46±0.94	**13.88±3.20**	**20.20±4.25**
SGC	before	69.09±0.99	7.28±1.50	5.45±1.42	66.95±1.69	2.74±0.85	3.21±0.78	96.53±0.48	4.70±1.26	3.11±1.24
	after	67.83±0.70	**17.65±1.01**	**16.09±1.06**	66.72±1.21	**10.59±2.40**	**10.67±2.61**	92.56±1.09	**13.88±3.37**	**20.25±4.44**
FairGNN	before	68.75±1.12	1.89±0.63	1.51±0.47	69.41±0.66	1.42±0.35	2.32±0.57	93.12±1.23	1.95±0.99	3.09±1.81
	after	69.38±2.07	**5.71±2.52**	**4.22±1.89**	69.97±0.42	**6.13±5.81**	**6.33±5.77**	92.56±1.49	**5.89±2.52**	**10.48±3.82**
FairVGNN	before	68.57±0.45	3.79±0.51	2.59±0.59	67.77±1.00	1.90±1.23	3.10±1.20	95.18±0.54	1.90±0.52	2.91±1.05
	after	67.65±0.38	**11.01±2.79**	**9.28±2.87**	65.74±1.42	**3.51±1.51**	**3.65±1.56**	91.56±1.13	**7.96±1.49**	**13.57±2.57**
FairSIN	before	67.33±0.22	1.73±1.49	2.61±1.44	67.18±0.30	0.39±0.89	2.40±1.02	94.72±0.62	0.23±0.15	0.45±0.16
	after	66.55±0.44	**9.48±2.62**	**10.39±1.06**	66.20±0.12	**11.82±0.75**	**14.58±0.22**	92.46±0.32	**10.90±2.12**	**23.65±7.77**

Observations

- NIFA can successfully deteriorate the fairness of both classic GNNs and fair GNNs.
- Different from conventional attacks, NIFA only slightly influence the utility of victim models.

Ablation Study and Defense Discussions



Observations

- Both Uncertainty-maximization principle and Homophily-increase principle are crucial for NIFA.

Defense Discussions

- Select reliable training nodes.
- Strengthen connections among different groups.
- Introduce fairness audits.

Thanks for listening!

NEURAL INFORMATION PROCESSING SYSTEMS

Are Your Models Still Fair? Fairness Attacks on Graph Neural Networks via Node Injections