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

Table R1. The attack performance of NIFA with different k.

	GCN			GraphSAGE		
	Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}
clean	95.88±1.61	3.84±0.34	1.91±0.75	96.28±0.68	2.67±1.15	2.54±1.07
0.10	96.03±0.54	4.44±1.45	9.46±1.84	95.83±0.52	4.31±1.32	8.46±1.75
0.25	93.67±1.62	10.06±2.03	15.68±2.59	93.87±0.80	7.20±3.03	12.70±3.25
0.50	93.37±1.48	13.49±2.83	20.33±3.82	94.32±0.44	7.06±2.94	14.25±2.43
0.75	92.46±1.07	10.86±4.17	17.32±5.35	94.42±0.78	5.02±1.78	9.06±1.45

Table R2. The attack performance with more limited attack budget. The best fairness attack performance are **bold**.

DBLP	**Acc**	Δ_{SP}	Δ_{EO}
Clean	95.88±1.61	3.84±0.34	1.91±0.75
FA-GNN	95.13±0.61	**8.46±1.39**	5.88±1.41
FATE	95.03±1.22	3.99±0.13	2.48±0.65
G-FairAttack	93.52±0.49	7.78±1.19	4.91±1.26
Ours	93.42±1.00	6.43±0.57	**14.25±1.67**

Table R3. The performance of NIFA with limited training nodes.

		Pokec-z			Pokec-n		
		Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}
train_ratio=25%	**before**	71.47±0.59	6.26±1.63	4.23±1.55	70.16±0.45	1.03±0.70	2.87±0.53
	after	70.48±0.41	15.16±1.64	13.18±1.77	69.49±0.31	9.79±2.02	9.52±1.84
train_ratio=10%	**before**	70.40±0.28	6.29±2.77	4.63±2.43	70.25±0.57	2.56±1.32	3.72±0.32
	after	70.14±0.26	15.26±3.02	13.36±3.06	69.25±0.62	11.94±1.78	12.00±1.75
train_ratio=5%	**before**	70.03±0.54	4.90±1.98	3.75±1.01	68.61±1.15	3.79±2.30	4.21±1.94
	after	69.49±0.65	14.35±2.03	12.42±2.04	68.05±1.63	10.31±5.68	10.60±5.04

Table R6. The attack performance on GCN with different architectures. We **bold** the results when NIFA successfully deteriorates the fairness of victim models.

		Pokec-z			Pokec-n		
		Acc	Δ_{SP}	Δ_{SP}	Acc	Δ_{SP}	Δ_{SP}
Different model layers							
1-layer (128-dim)	before	71.07±0.30	6.36±1.67	4.60±1.33	71.20±0.44	0.97±0.28	1.63±0.78
	after	71.26±0.21	**15.46±2.47**	**13.48±2.53**	70.25±0.30	**7.94±3.65**	**7.67±3.36**
2-layer (128-dim)	before	71.56±0.24	6.52±0.91	4.41±0.96	70.93±0.27	1.02±0.52	1.82±1.09
	after	71.35±0.24	**15.59±1.47**	**13.77±1.50**	70.15±0.60	**9.50±3.43**	**9.19±3.17**
3-layer (128-dim)	before	70.81±0.27	9.13±1.23	7.05±1.33	70.17±0.39	1.17±0.61	2.78±0.98
	after	69.99±0.28	**13.65±3.06**	**11.83±3.12**	69.11±0.74	**8.44±4.50**	**8.59±4.02**
Different hidden dimension							
64-dim (2-layer)	before	71.63±0.19	6.52±1.64	4.47±1.68	70.26±0.56	1.09±0.69	3.09±0.33
	after	71.28±0.33	**17.05±1.35**	**15.16±1.29**	69.62±0.27	**7.94±2.15**	**8.03±2.29**
128-dim (2-layer)	before	71.56±0.24	6.52±0.91	4.41±0.96	70.93±0.27	1.02±0.52	1.82±1.09
	after	71.35±0.24	**15.59±1.47**	**13.77±1.50**	70.15±0.60	**9.50±3.43**	**9.19±3.17**
256-dim (2-layer)	before	71.33±0.24	7.06±1.16	5.01±1.18	71.52±0.30	1.13±0.51	1.94±0.97
	after	71.30±0.31	**16.28±1.28**	**14.39±1.32**	70.45±0.47	**8.02±1.34**	**7.62±1.34**

Table R4. The fairness attack comparison with baselines on additional datasets. "*" denotes pokec datasets with smaller scales.

	Pokec-z*			Pokec-n*			German		
	Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}
Clean	68.50±0.55	6.49±1.79	4.92±1.22	68.02±0.59	8.13±1.81	7.76±1.68	71.12±1.05	9.88±1.02	8.32±0.93
FA-GNN	69.32±0.59	13.53±1.75	11.65±1.76	69.49±0.31	9.79±2.02	9.52±1.84	70.33±1.32	6.10±8.24	6.55±6.87
FATE	68.57±0.80	7.33±2.02	5.48±1.90	67.68±0.73	11.05±1.88	10.56±1.86	71.04±0.60	**11.57±5.27**	9.43±4.55
G-FairAttack	69.77±0.26	8.06±1.45	6.05±1.08	68.58±0.42	9.70±1.43	8.93±1.57	72.80±0.48	5.62±0.45	9.27±2.18
NIFA	68.14±0.35	**14.41±2.93**	**12.59±2.97**	66.74±0.60	**12.74±3.73**	**12.28±3.76**	70.16±1.55	10.29±13.33	**10.54±12.98**

Table R5. The performance of defense strategies with NIFA.

	Pokec-z			Pokec-n			DBLP		
	Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}	Acc	Δ_{SP}	Δ_{EO}
Clean	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
NIFA	70.50±0.30	17.36±1.16	15.59±1.08	70.12±0.37	10.10±2.80	9.85±1.97	93.37±0.28	13.49±2.83	20.33±3.82
GNNGuard	71.50±0.35	18.13±2.31	16.32±1.44	70.55±0.43	13.82±1.76	13.38±1.42	95.73±0.28	7.42±1.82	14.94±1.58
ElasticGNN	71.36±0.20	13.53±0.92	11.36±1.23	70.32±0.28	6.90±1.02	6.44±1.32	94.22±0.37	9.76±1.65	17.04±2.01