Towards A Unified Framework For Fair And Stable Graph Representation Learning

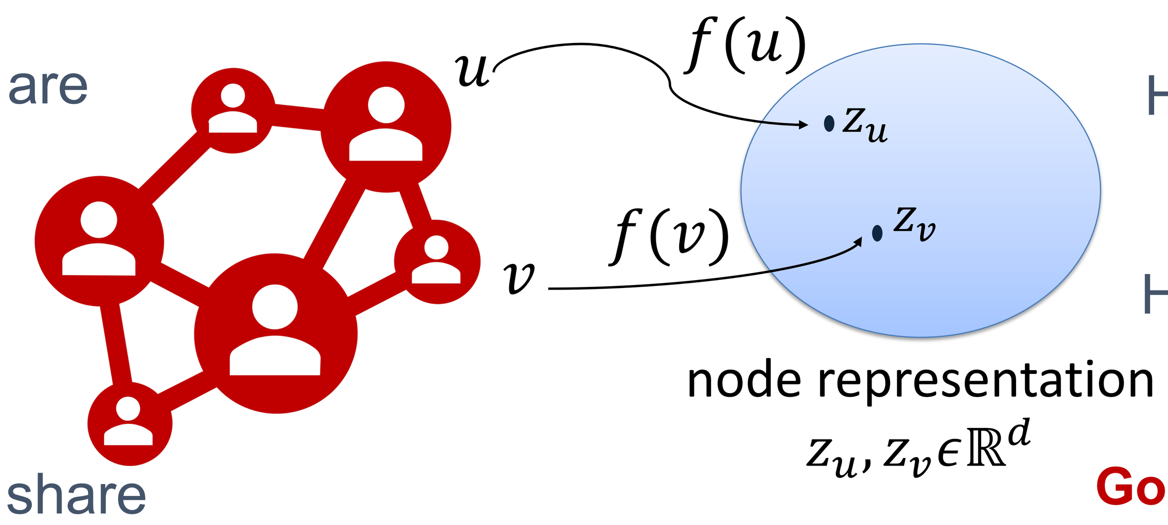
Chirag Agarwal, Himabindu Lakkaraju*, Marinka Zitnik*

chirag_agarwal@hms.harvard.edu, hlakkaraju@hbs.edu, marinka@hms.harvard.edu

Motivation

Limited existing work on learning graph representation that are Fair and stable as they present some unique challenges:

- Need for a unifying framework that jointly optimizes for Fairness and Stability
- Nodes with similar sensitive attribute values are likely to share similar representations leading to severe discriminatory biases



Open Questions

How to identify a connection between fairness and stability?

How does fairness and stability affect downstream performance?

Goal: Given a graph G, learn embeddings that are counterfactually fair and stable to attribute and structural perturbations of G

Our Framework: NIFTY

- NIFTY identifies a key connection between counterfactual fairness and stability where stability accounts for robustness *w.r.t.* small random perturbations to node attributes and/or edges, counterfactual fairness accounts for robustness *w.r.t.* modifications of the sensitive attribute
- NIFTY enforces fairness and stability both in the objective function as well as in the GNN architecture
  - The objective function maximizes the similarity between representations of the original nodes in the graph, and their counterparts in the augmented graph

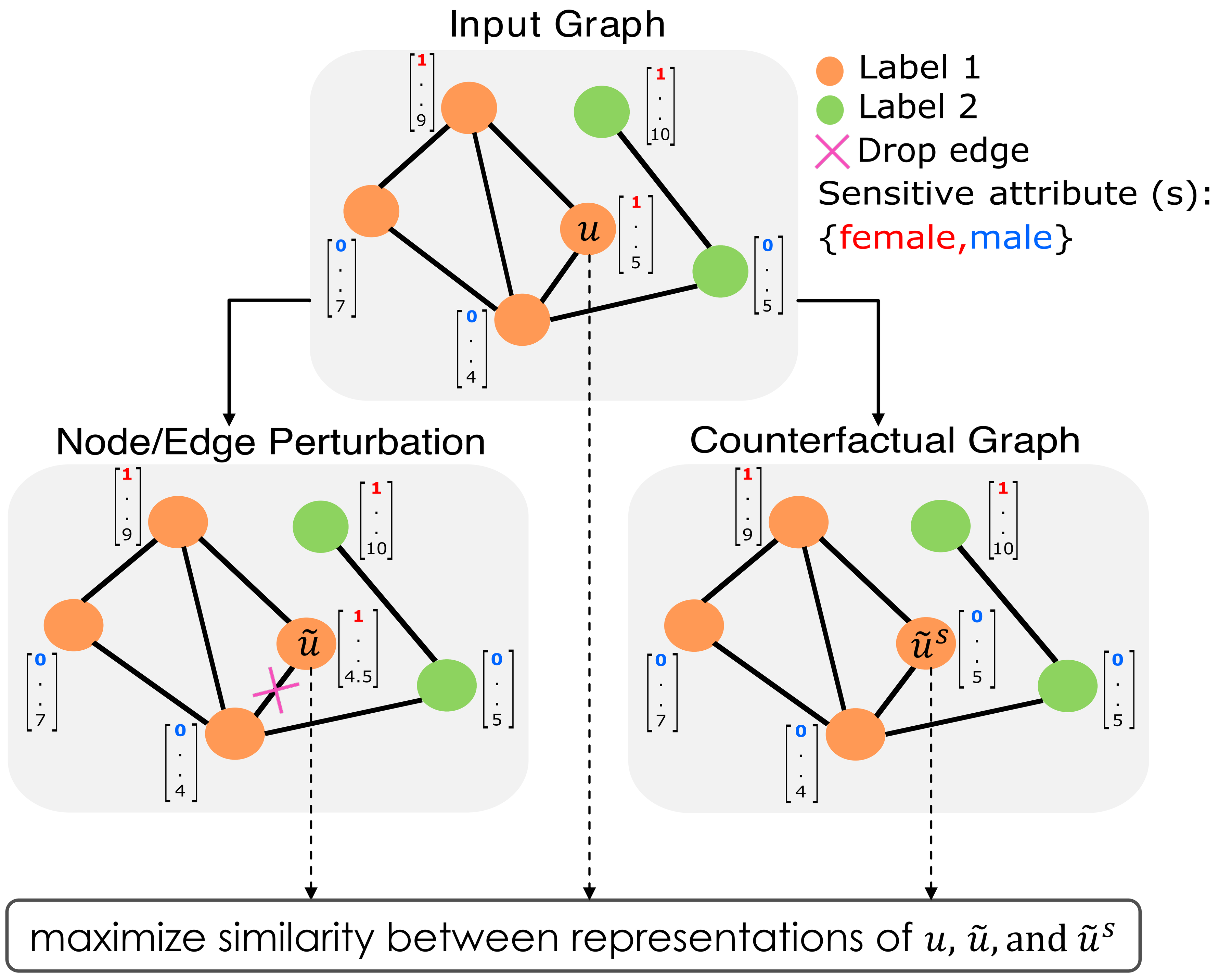
$$\mathcal{L}_s = \mathbb{E}_u \Big[\frac{1}{2} \big(D(t(\mathbf{z}_u), \text{sg}(\tilde{\mathbf{z}}_u)) + D(t(\tilde{\mathbf{z}}_u), \text{sg}(\mathbf{z}_u)) \big) \Big]$$

$$\min_{\theta_{\text{ENC}}, \theta_t, \theta_f} \mathbb{E}_u \left[(1-\lambda)\mathcal{L}_c \right] + \lambda \mathcal{L}_s$$

  - Enhancing the neural message passing step by carrying out layer-wise weight normalization using the Lipschitz constant

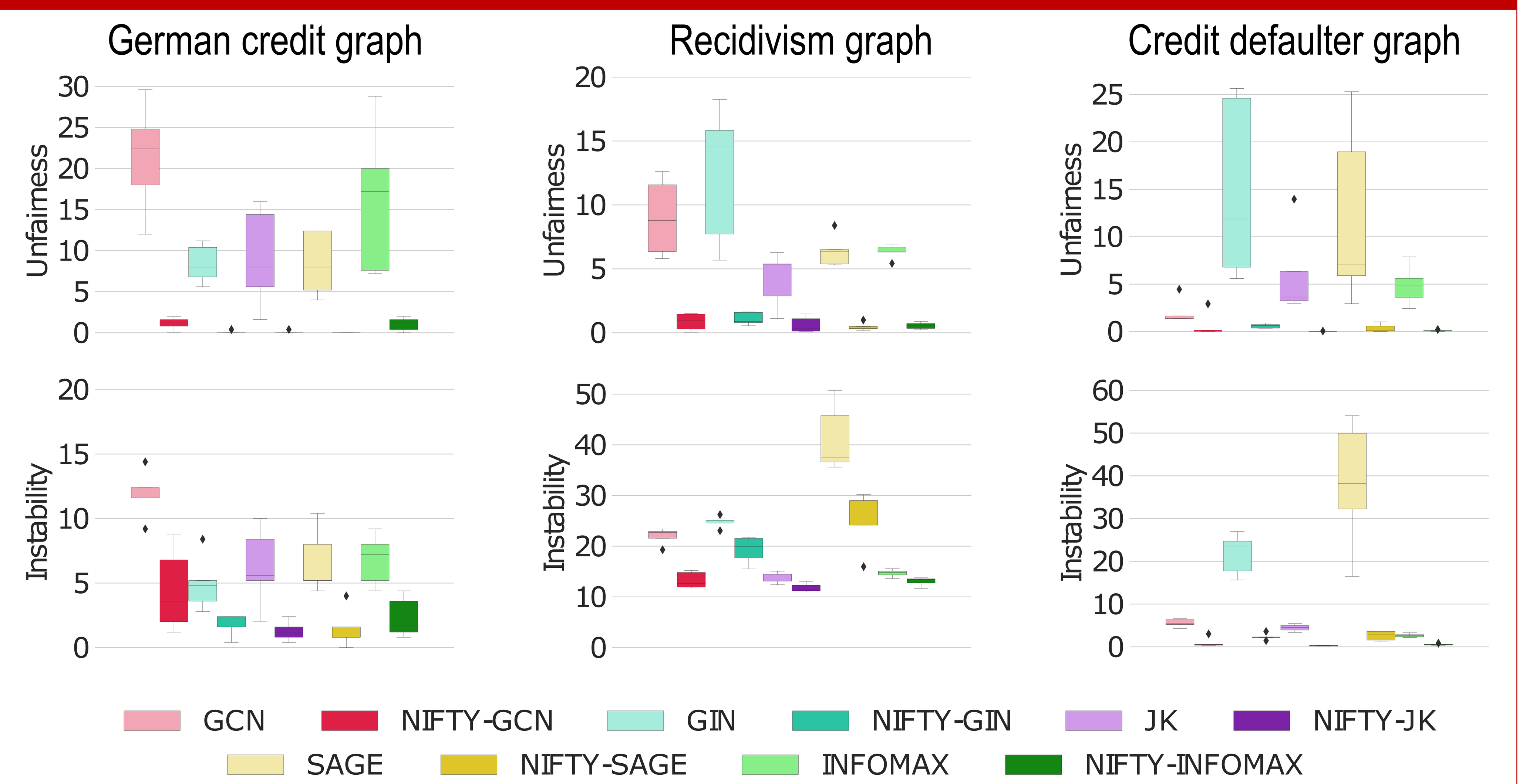
$$\tilde{\mathbf{W}}_a^k = \mathbf{W}_a^k / \sigma(\mathbf{W}_a^k)$$

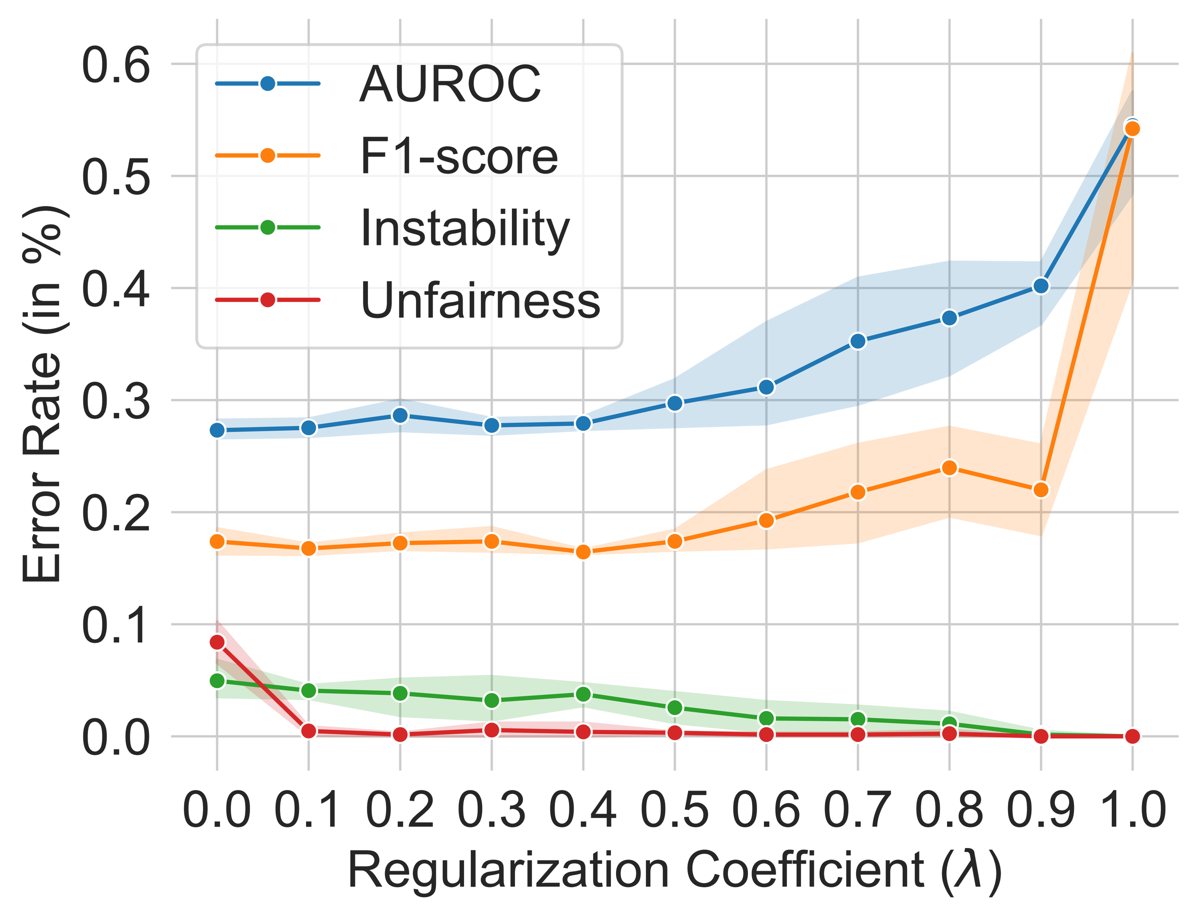
$$\mathbf{h}_u^k = \sigma(\tilde{\mathbf{W}}_a^k \, \mathbf{h}_u^{k-1} + \mathbf{W}_n^k \textstyle\sum_{v \in \mathcal{N}(u)} \mathbf{h}_v^{k-1})$$



Improved Fairness and Stability across **3 datasets and 5 GNNs**

- 3 new graph datasets comprising of high-stakes decisions in criminal justice and financial lending domains designed to analyze fairness and stability properties of GNNs
- Across 3 datasets and 5 GNNs, NIFTY improves stability and fairness of GNNs by 60.87% and 92.01%, respectively, without sacrificing the predictive performance
- Enforcing fairness and stability both using the objective function and layer-wise normalization of GNN architecture using the Lipschitz constant are important
- We observe that increasing regularization coefficient λ in NIFTY decreases the error rates for counterfactual fairness and stability steadily





Ablation study

Method	AUROC (↑)	F1-score (↑)	Unfairness (↓)	Instability (↓)	Δ_{SP}(↓)	Δ_{EO}(↓)
GCN [Kipf and Welling, 2017]	86.52±0.42	77.50±0.87	9.02±3.04	21.97±1.63	8.49±0.73	5.93±0.56
NIFTY-GCN w/o obj. changes (Sec. 4.1)	80.02 ±0.20	67.51 ±0.23	2.61±0.64	13.69±0.60	5.86±0.85	4.65±0.49
NIFTY-GCN w/o arch. changes (Sec. 4.2)	84.83 ±2.85	76.15±5.74	1.64 ±1.58	13.98 ±1.38	4.29 ±1.32	3.48 ±1.37
NIFTY-GCN	81.40 ±0.89	69.24±0.70	**0.84**±0.68	**13.28**±1.62	**3.16**±0.60	**2.99**±0.40

Comparison of NIFTY to baseline methods

Dataset	Method	AUROC (↑)	F1-score (↑)	Unfairness (↓)	Instability (↓)	Δ_{SP} (↓)	Δ_{EO} (↓)
German credit graph	FairGCN	75.21±0.36	81.52±0.68	N/A	7.84±2.20	38.12±4.87	26.70±4.27
	RobustGCN	71.06±1.48	78.85±6.39	7.68±4.69	4.48±1.07	25.78±10.92	18.47±9.87
	NIFTY-GCN	70.32±4.42	81.98±0.82	**1.12**±0.77	**4.48**±3.23	**15.08**±8.22	**12.56**±8.60
Recidivism graph	FairGCN	87.55±0.60	78.14±0.94	N/A	24.37±2.33	6.51±0.77	4.51±1.10
	RobustGCN	87.25±1.67	79.02±2.84	2.61±1.58	**13.02**±6.06	5.36±1.28	4.20±1.88
	NIFTY-GCN	81.40±0.89	69.24±0.70	**0.84**±0.68	13.28±1.62	**3.16**±0.60	**2.99**±0.40
Credit defaulter graph	FairGCN	72.69±1.23	80.16±2.03	N/A	5.73±0.60	15.86±5.16	14.43±6.06
	RobustGCN	72.98±0.26	81.79±0.60	0.94±0.60	1.68±0.83	12.41±0.54	10.16±0.49
	NIFTY-GCN	71.92±0.19	81.99±0.63	**0.63**±1.28	**0.95**±1.16	**12.40**±1.62	**10.09**±1.55