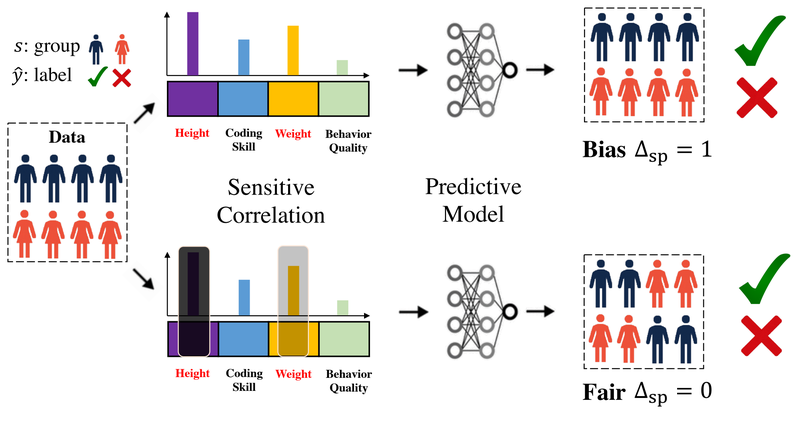
Improving Fairness in Graph Neural Networks via Mitigating Sensitive Attribute Leakage

Yu Wang[1] 
Yuying Zhao[1] 
Yushun Dong[2] 
Huiyuan Chen[3] 
Jundong Li[2] 
Tyler Derr[1] 

Network and Data Science Lab

Personal website | Wechat

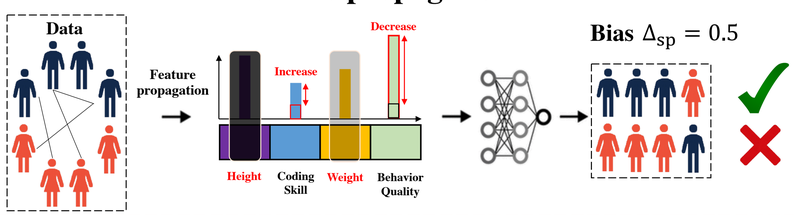
Group Fairness

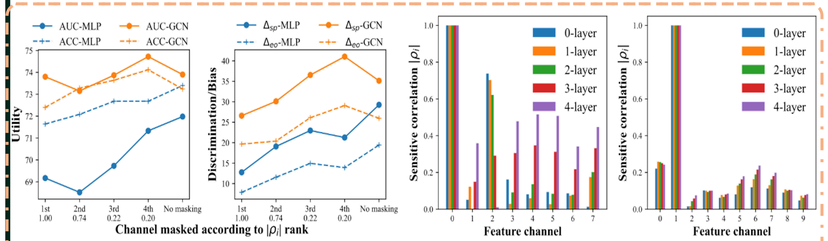


$\Delta_{sp} = |P(\hat{y}=1|s=0) - P(\hat{y}=1|s=1)|$ $\Delta_{sp} = |P(\hat{y}=1|y=1,s=0) - P(\hat{y}=1|y=1,s=1)|$

Group Fairness on Graphs

Sensitive Leakage: sensitive correlation changes after feature propagation!





(a) Model utility (b) Model bias Sensitive Correlation on (c) German and (d) Credit

(a)-(b) Masking channel with less sensitive correlation leads to more biased predictions and higher model utility.

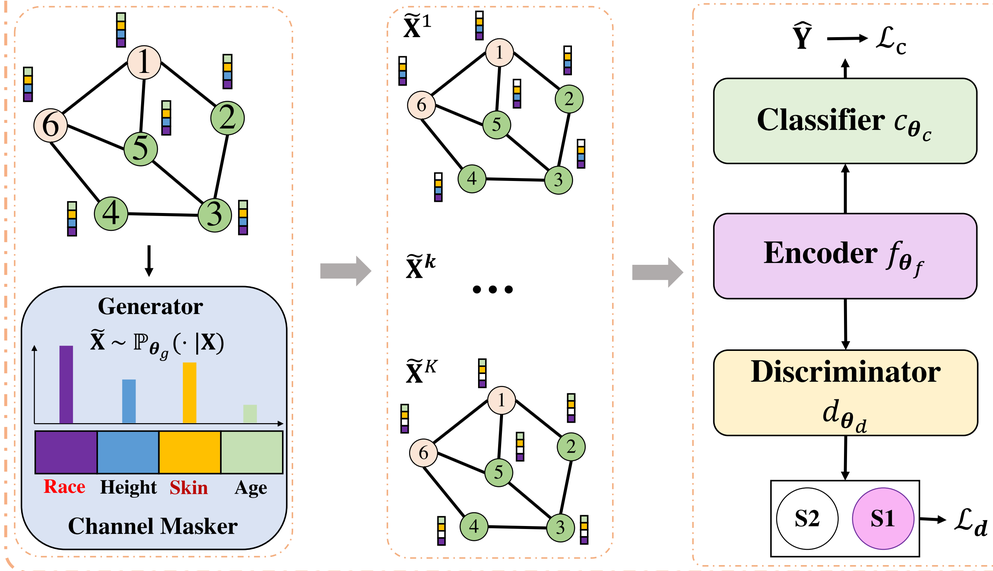
(c)-(d) Some originally innocuous feature channels become highly correlated to the sensitive feature channel.

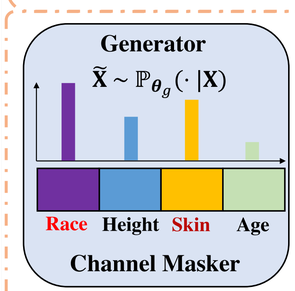
(c)-(d) Different feature channels have different changing speed on their sensitive correlation.

Fair View Graph Neural Network

Generative Adversarial Debiasing

Fair Feature View Generation



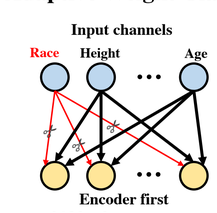


$$\tilde{\mathbf{X}}^k = \mathbf{X} \odot \mathbf{m}^k = [\mathbf{X}_1^T \odot \mathbf{m}, \ldots, \mathbf{X}_n^T \odot \mathbf{m}]$$

$$\boldsymbol{m}_i^k \sim \text{Bernoulli}(1-p_i), \forall i \in \{1,2,..,d\} \qquad \mathbf{m}_i, \mathbf{m}_j \text{ are independent}$$

$$p_{ij} = \frac{\exp(\frac{\log(\pi_{ij})+g_{ij}}{\tau})}{\sum_{k=1}^{2}\exp(\frac{\log(\pi_{ik})+g_{ik}}{\tau})}, \forall j = 1,2, i \in \{1,2,\ldots,d\}, \qquad g_{ij} \sim \text{Gumbel}(0,1)$$

Adaptive Weight Clamp



$$\mathbf{W}_{ij}^{f,1} = \begin{cases} \mathbf{W}_{ij}^{f,1}, & |\mathbf{W}_{ij}^{f,1}| \le \epsilon * \mathbf{p}_j \\ \text{sign}(\mathbf{W}_{ij}^{f,1}) * \epsilon * \mathbf{p}_j, & |\mathbf{W}_{ij}^{f,1}| > \epsilon * \mathbf{p}_j \end{cases}$$

χ: Network homophily
μ: The mean of the embedding difference
ϵ: Preset threshold
$\mathbf{W}$: Weight of the 1st hidden layer
$\mathbf{p}_i$: Keeping probability
$\Delta\mu$: The mean of the initial feature difference

$$||\boldsymbol{\mu}||_2 = ||(2\chi-1)\mathbf{W}^{f,1}\Delta\boldsymbol{\mu}||_2 \le (2\chi-1)\Big(\sum_{i=1}^{d_1}\big(\sum_{r\in\mathcal{S}}\epsilon\mathbf{p}_r\Delta\boldsymbol{\mu}_r + \sum_{k\in\mathcal{NS}}\epsilon\mathbf{p}_k\Delta\boldsymbol{\mu}_k\big)^2\Big)^{0.5}$$

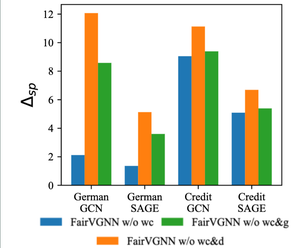
Network homophily is closely related to Graph Fairness

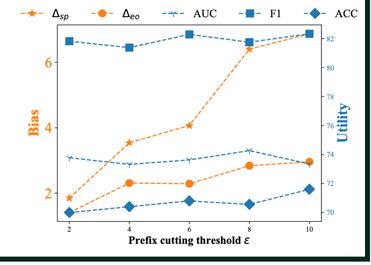
Experimental Validation

Encoder	Method	German Dataset AUC (↑)	F1 (↑)	ACC (↑)	Δ_{sp} (↓)	Δ_{eo} (↓)
GCN	Vanilla	74.11±0.37	82.46±0.89	73.44±1.09	35.17±7.27	25.17±5.89
	NIFTY	68.78±2.69	81.40±0.54	69.92±1.14	5.73±5.25	5.08±4.29
	EDITS	69.41±2.33	81.55±0.59	71.60±0.89	4.05±4.48	3.89±4.23
	FairGNN	67.35±2.13	82.01±0.26	69.68±0.30	3.49±2.15	3.40±2.15
	FairVGNN	72.41±2.10	82.14±0.42	70.16±0.86	1.71±1.68	0.88±0.58

Encoder	Method	Credit Dataset AUC (↑)	F1 (↑)	ACC (↑)	Δ_{sp} (↓)	Δ_{eo} (↓)
GIN	Vanilla	74.36±0.21	82.28±0.64	74.02±0.73	14.48±2.44	12.35±2.86
	NIFTY	70.90±0.24	84.05±0.82	75.59±0.66	7.09±4.62	6.22±3.26
	EDITS	72.35±1.11	82.47±0.85	74.07±0.98	14.11±14.45	15.40±15.76
	FairGNN	68.66±4.48	79.47±5.29	70.33±5.50	4.67±3.06	3.94±1.49
	FairVGNN	71.36±0.72	87.44±0.23	78.18±0.20	2.85±2.01	1.72±1.80

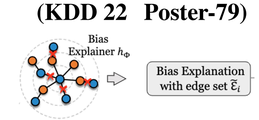
Encoder	Method	Bail Dataset AUC (↑)	F1 (↑)	ACC (↑)	Δ_{sp} (↓)	Δ_{eo} (↓)
SAGE	Vanilla	90.71±0.69	80.99±0.55	86.72±0.48	2.16±1.53	0.84±0.55
	NIFTY	92.04±0.89	77.81±6.03	84.11±5.49	5.74±0.38	4.07±1.28
	EDITS	89.07±2.26	77.83±3.79	84.42±2.87	3.74±3.54	4.46±3.50
	FairGNN	91.53±0.38	82.55±0.98	87.68±0.73	1.94±0.82	1.72±0.70
	FairVGNN	91.56±1.71	83.58±1.88	88.41±1.29	1.14±0.67	1.69±1.13



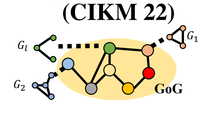


Recent/Future Work

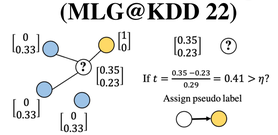
Structural explanation for bias (KDD 22 Poster-79)



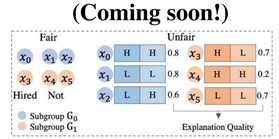
Mitigating Imbalance (CIKM 22)



Imbalanced Label Propagation (MLG@KDD 22)



Fairness and Explanation (Coming soon!)



Channel homophily, propagation and fairness