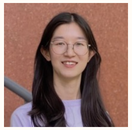
# Improving Fairness in Graph Neural Networks via Mitigating Sensitive Attribute Leakage


Yu Wang[1] 

Yuying Zhao[1] 

Yushun Dong[2] 

Huiyuan Chen[3] 

Jundong Li[2] 

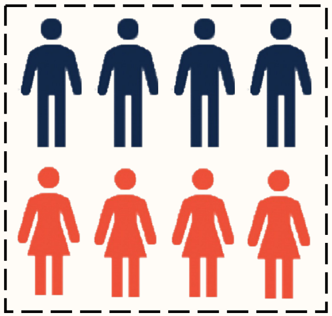
Tyler Derr[1] 

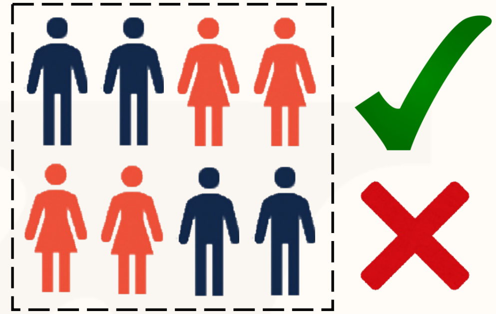
1. Network and Data Science Lab, Vanderbilt University
2. University of Virginia
3. Case Western Reserve University

# Background – Group Fairness

$s$: group

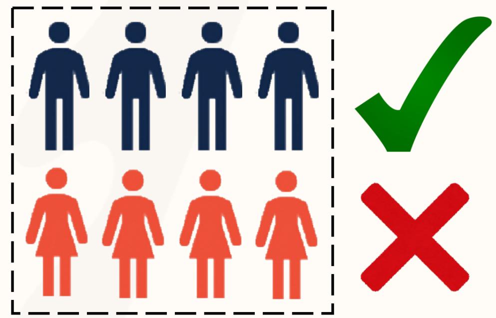
$\hat{y}$: label



**Fair!**

$\Delta_{sp} = 0$



**Bias!**

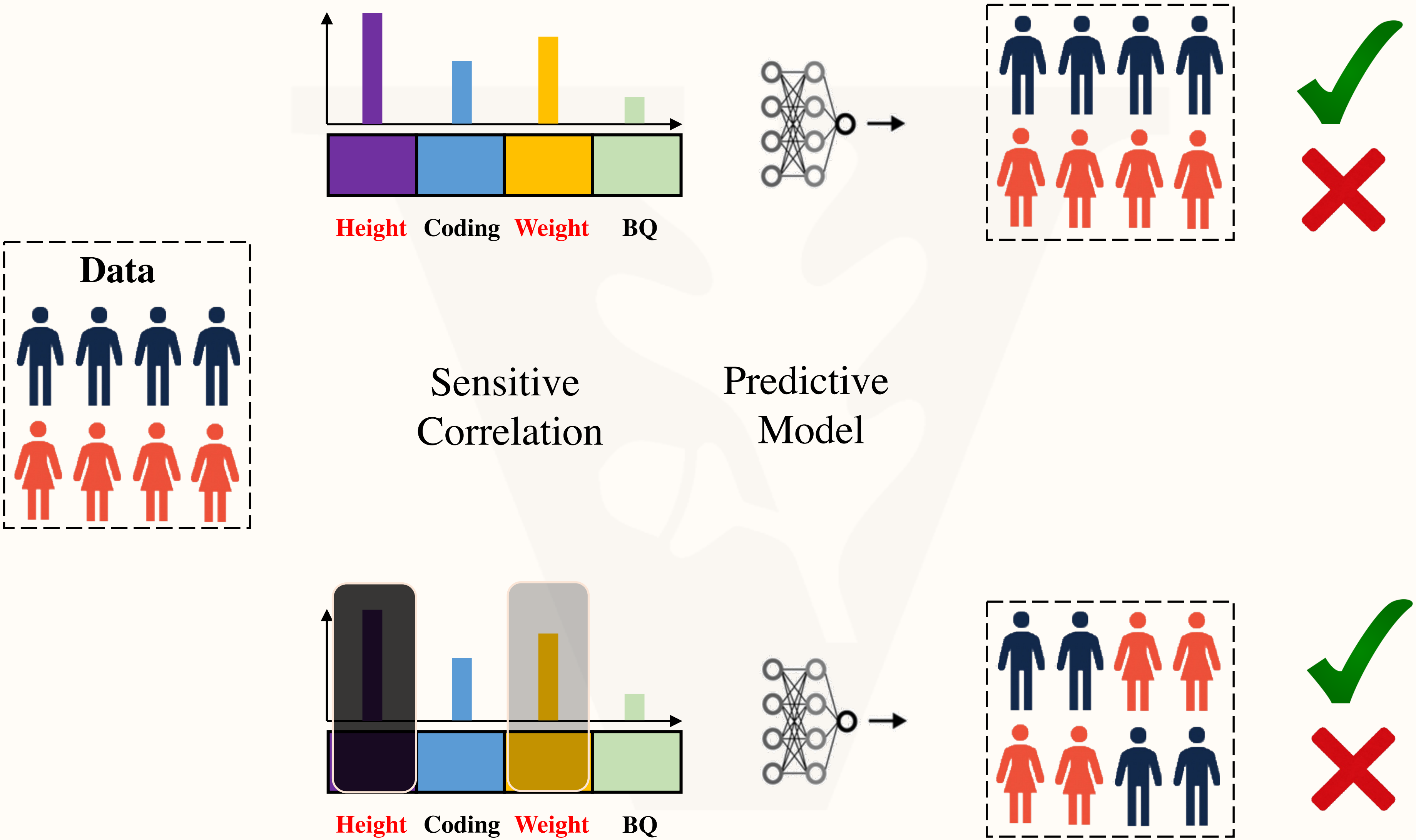
$\Delta_{sp} = 1$

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

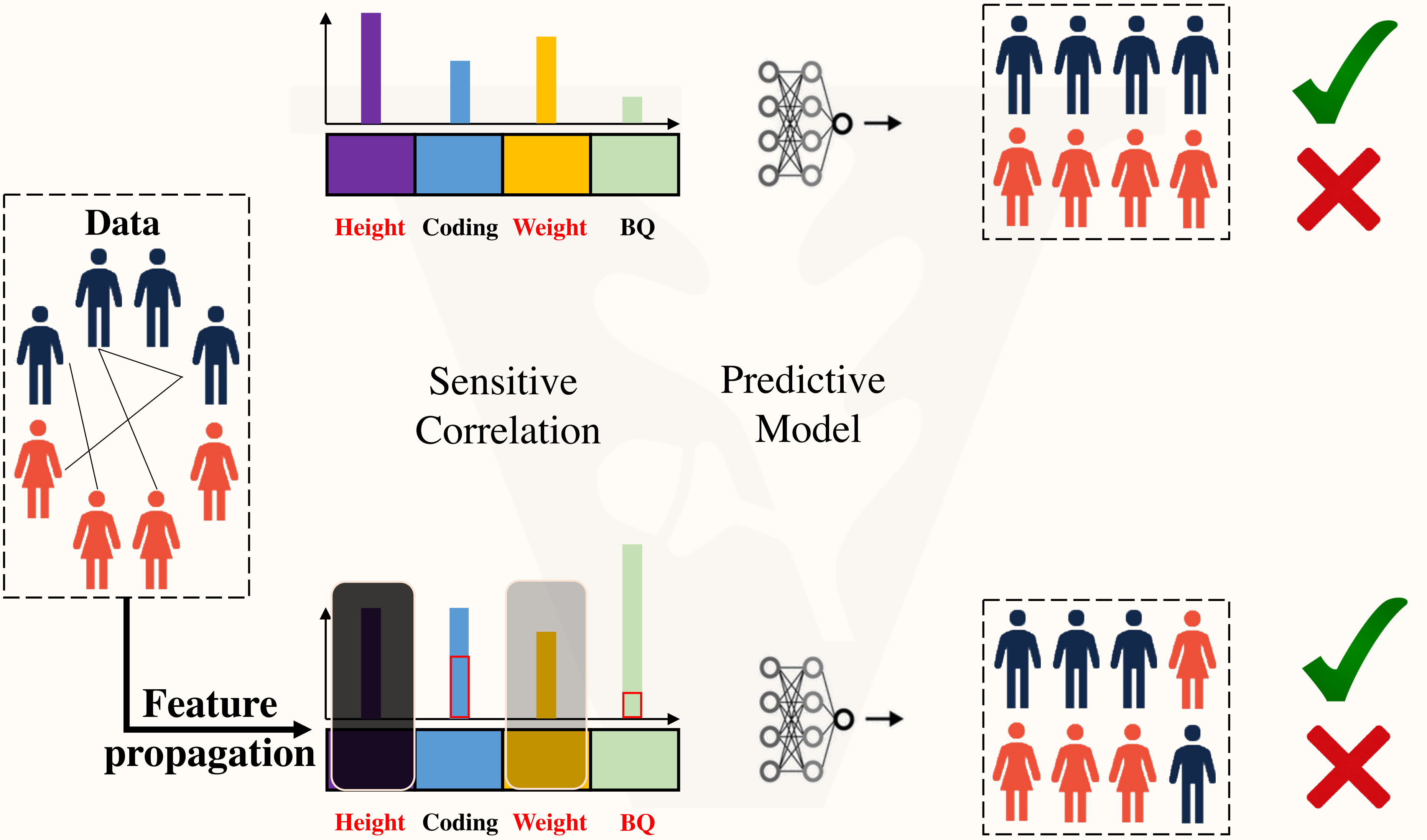
**$s$: sensitive attribute**

**Discriminative and unfair decision!**

# Background – Sensitive leakage



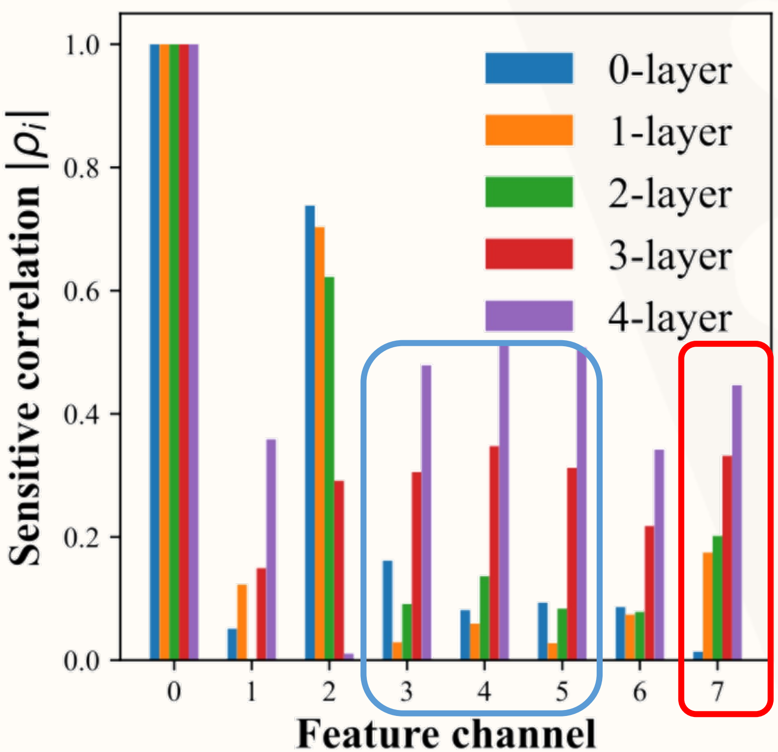
# Background – Correlation variation



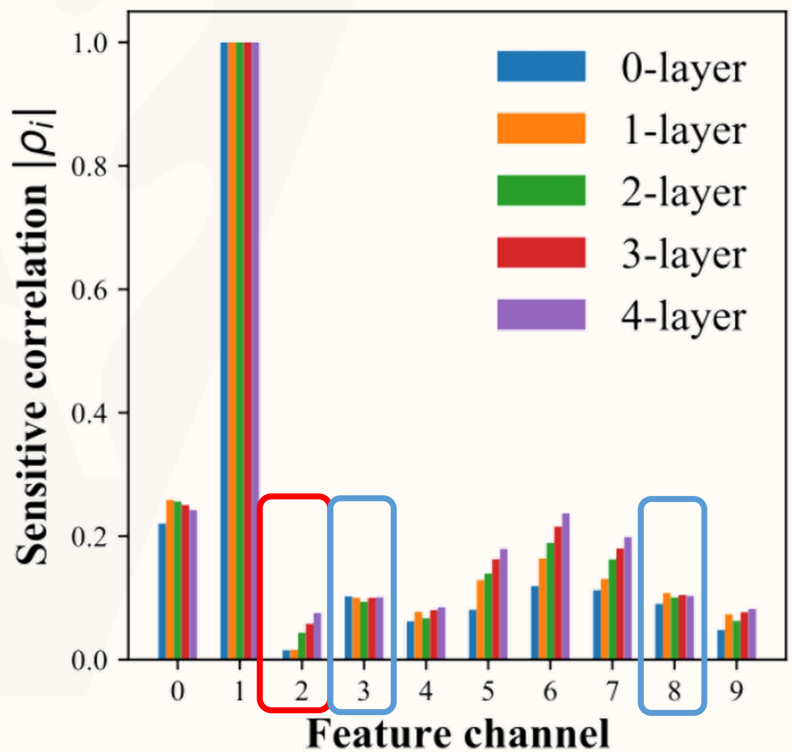
# Motivation – Correlation variation

Mask feature channels with higher correlation to the sensitive attributes

**German**

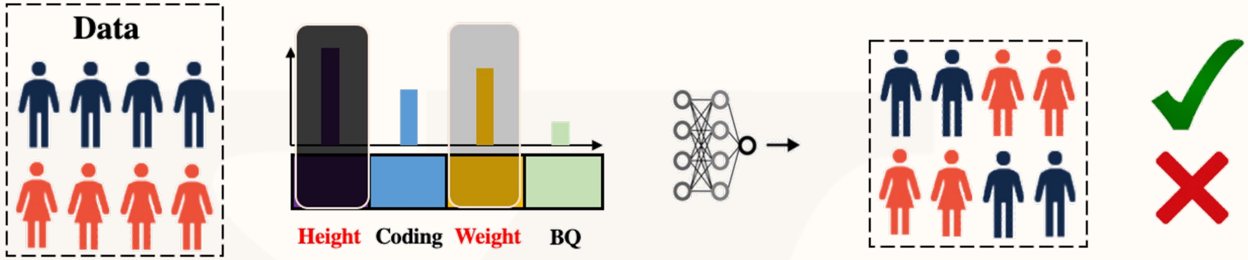



**Credit**

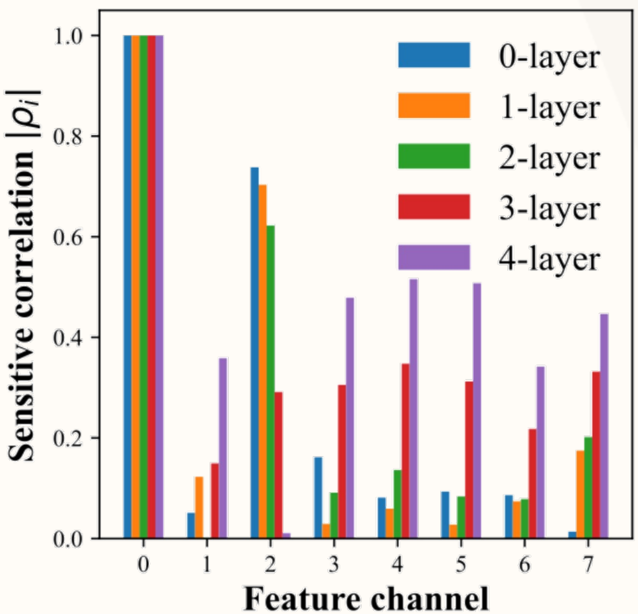


# FairVGNN – Fair View Graph Neural Network

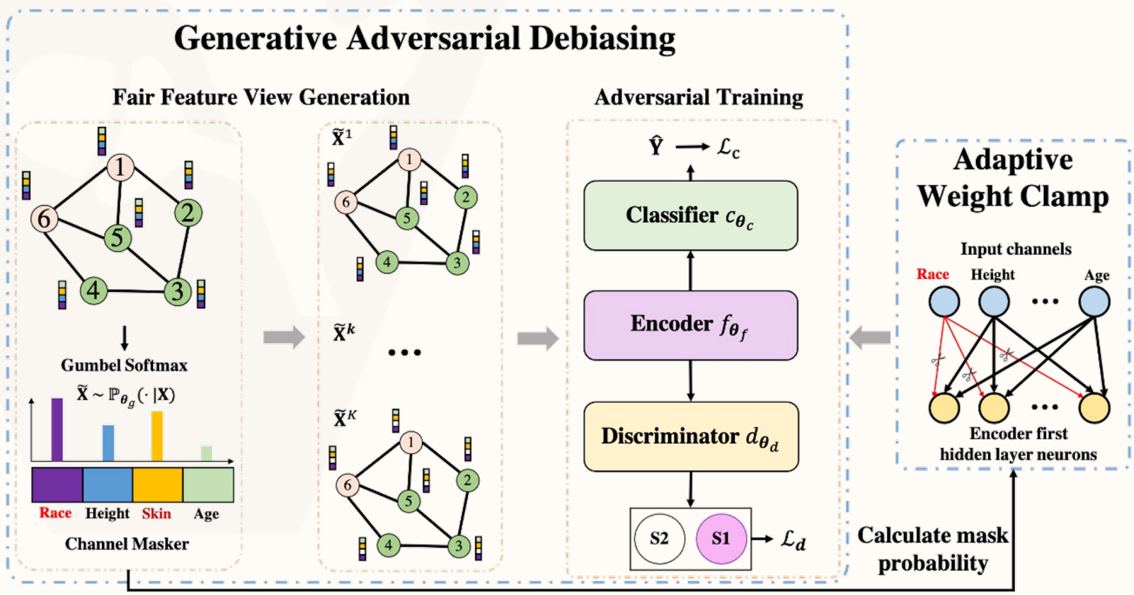
## What we desire




## Graph Domain Challenge

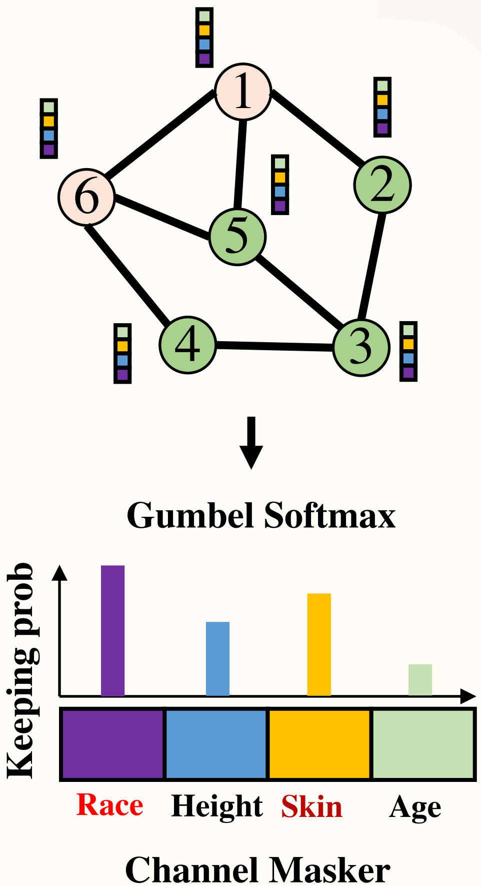


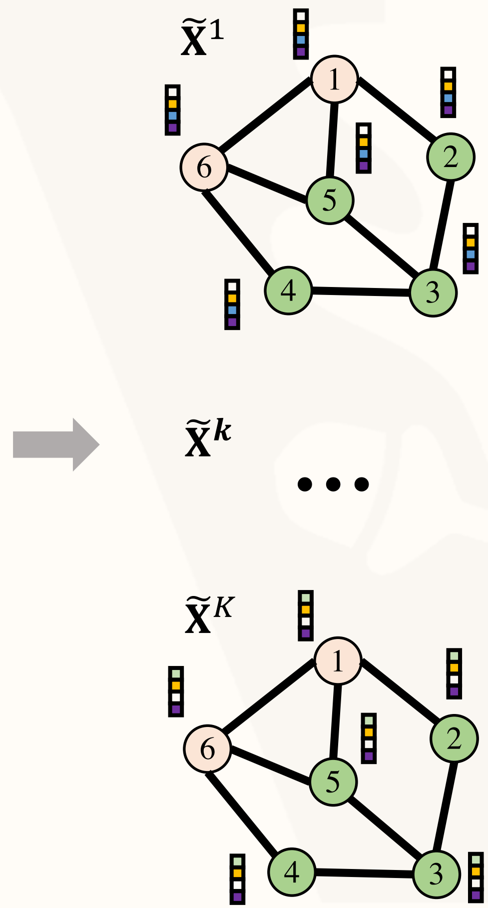

## Solution: FairVGNN



# FairVGNN – Fair View Graph Neural Network

## Fair Feature View Generation



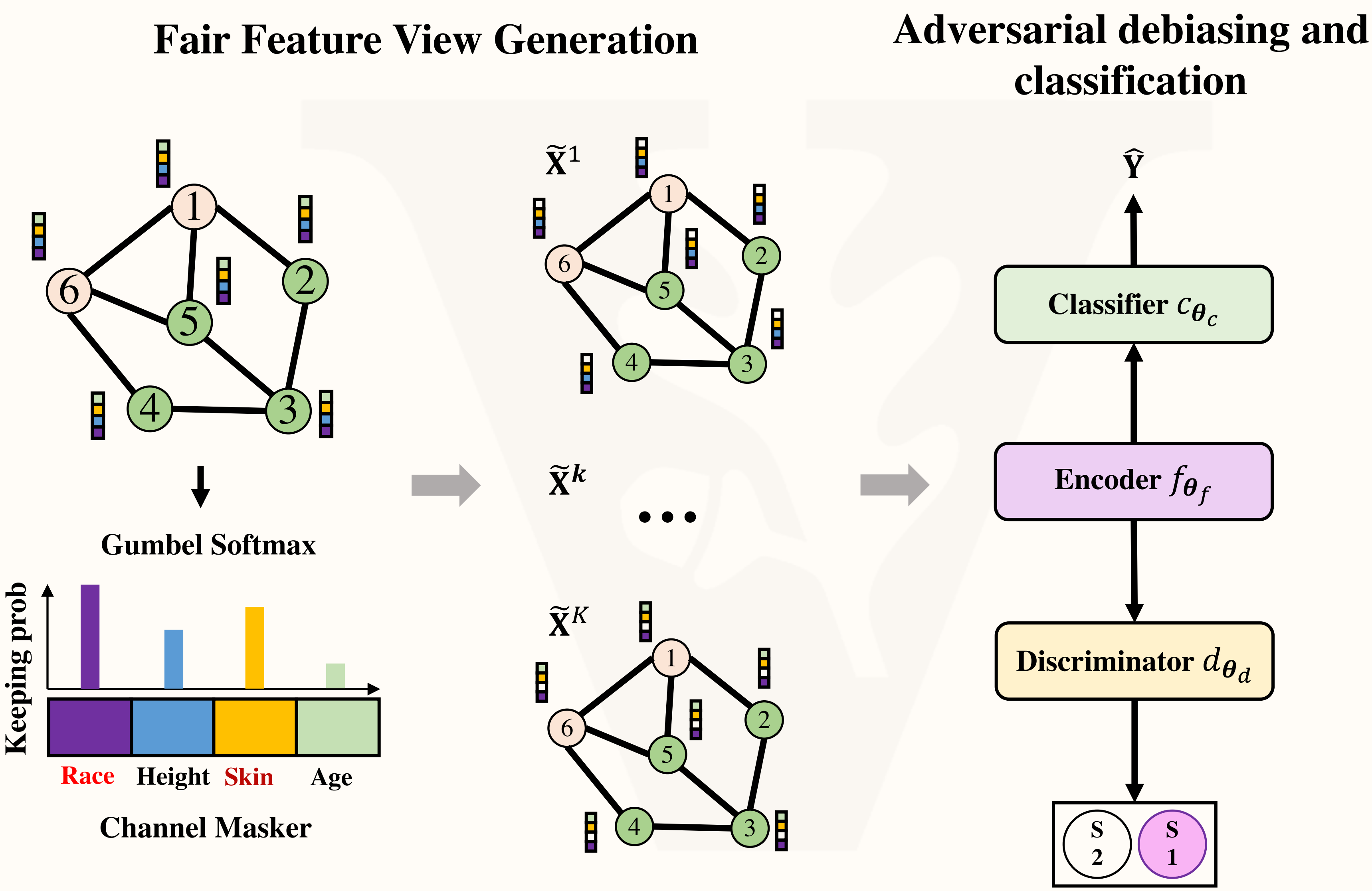

$$\tilde{\mathbf{X}}^k = \mathbf{X} \odot \mathbf{m}^k = [\mathbf{X}_1^{\mathrm{T}} \odot \mathbf{m}, \dots, \mathbf{X}_n^{\mathrm{T}} \odot \mathbf{m}]$$

$\mathbf{m}_i, \mathbf{m}_i$ are independent

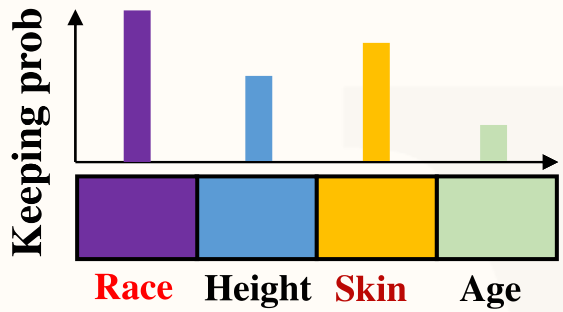
$$\boldsymbol{m}_i^{\boldsymbol{k}} \sim \mathrm{Bernoulli}(1 - p_i), \forall i \in \{1,2,..,d\}$$

**Gumbel Softmax**

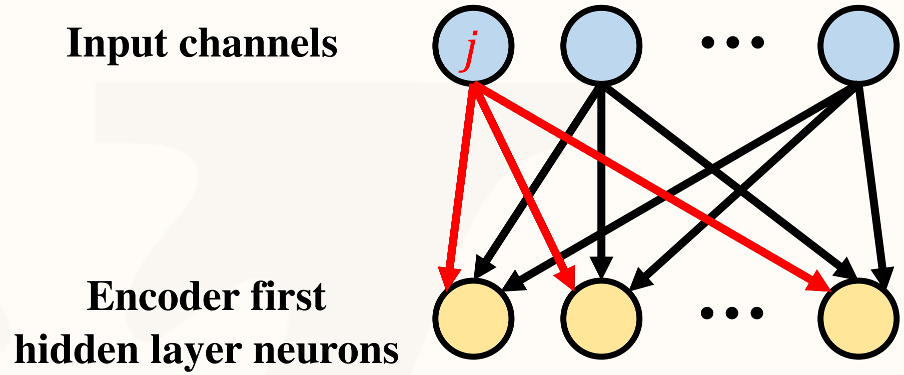
# FairVGNN – Fair View Graph Neural Network



# FairVGNN – Fair View Graph Neural Network




**Channel Masker**



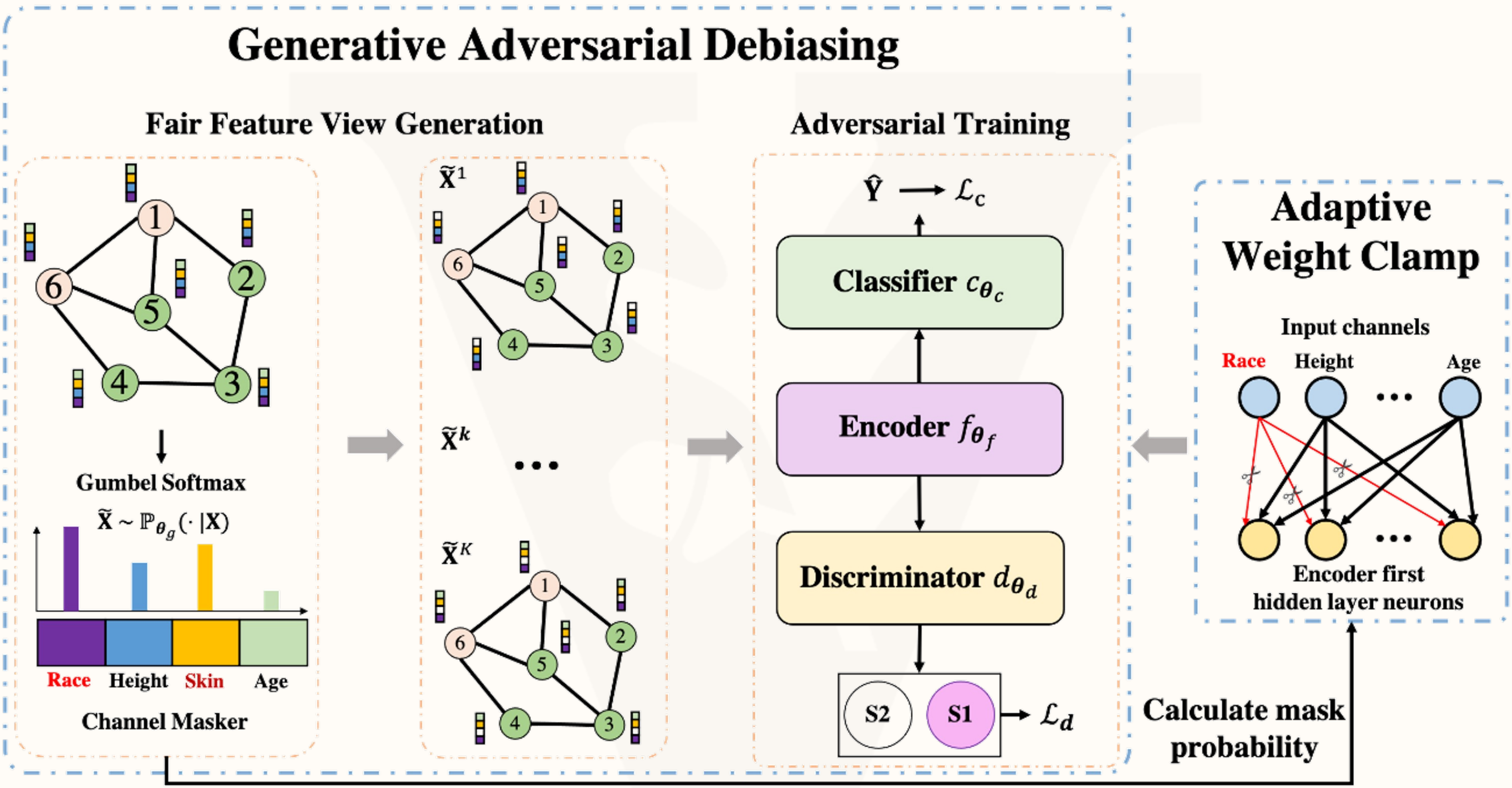

$$\mathbf{p} = \sum_{k=1}^{K} \mathbf{m}^k \in \mathbb{R}^d$$

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

**Adaptive Weight Clamp**

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

# FairVGNN – Fair View Graph Neural Network



# FairVGNN – Datasets and Baselines

Table 2: Basic dataset statistics.

| Dataset | German | Credit | Bail |
|---|---|---|---|
| #Nodes | 1000 | 30,000 | 18,876 |
| #Edges | 22,242 | 1,436,858 | 321,308 |
| #Features | 27 | 13 | 18 |
| Sens. | Gender | Age | Race |
| Label | Good/bad Credit | Default/no default Payment | Bail/no bail |

**Augmentation-based:** NIFTY, EDITS

**Adversarial-based:** FairGNN

# FairVGNN – Experiments

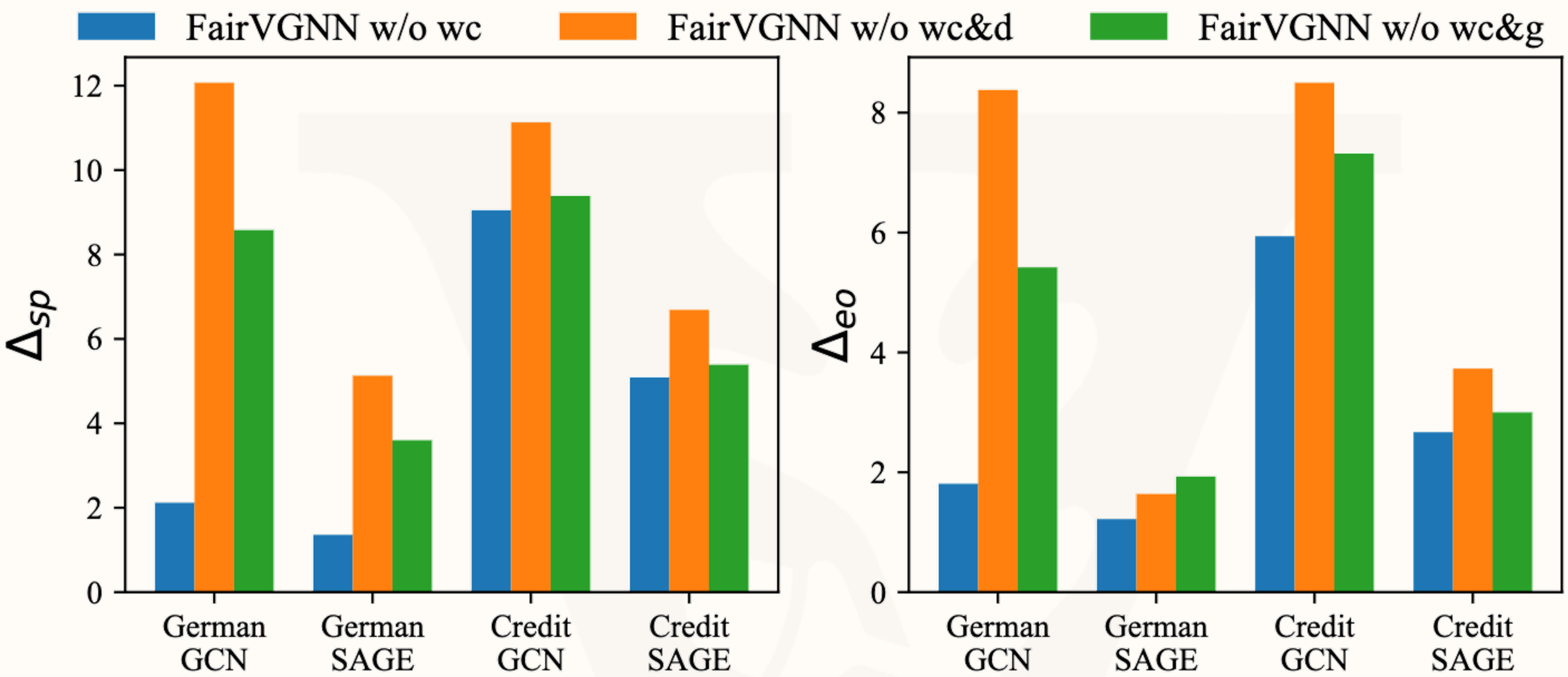
| Encoder | Method | German | | | | |
|---|---|---|---|---|---|---|
| | | AUC (↑) | F1 (↑) | ACC (↑) | $\Delta_{sp}$ (↓) | $\Delta_{eo}$ (↓) |
| GCN | Vanilla | 74.11±0.37 | 82.46±0.89 | 73.44±1.09 | 35.17±7.27 | 25.17±5.89 |
| | NIFTY | 68.78±2.69 | 81.40±0.54 | 69.92±1.14 | 5.73±5.25 | 5.08±4.29 |
| | 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 | | | | |
|---|---|---|---|---|---|---|
| | | AUC (↑) | F1 (↑) | ACC (↑) | $\Delta_{sp}$ (↓) | $\Delta_{eo}$ (↓) |
| 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 |

(1) Compared with vanilla GNN model, the bias-mitigating model can achieve lower bias

(2) Compared with other baselines, FairVGNN can achieve even better trade-off between fairness and utility performance

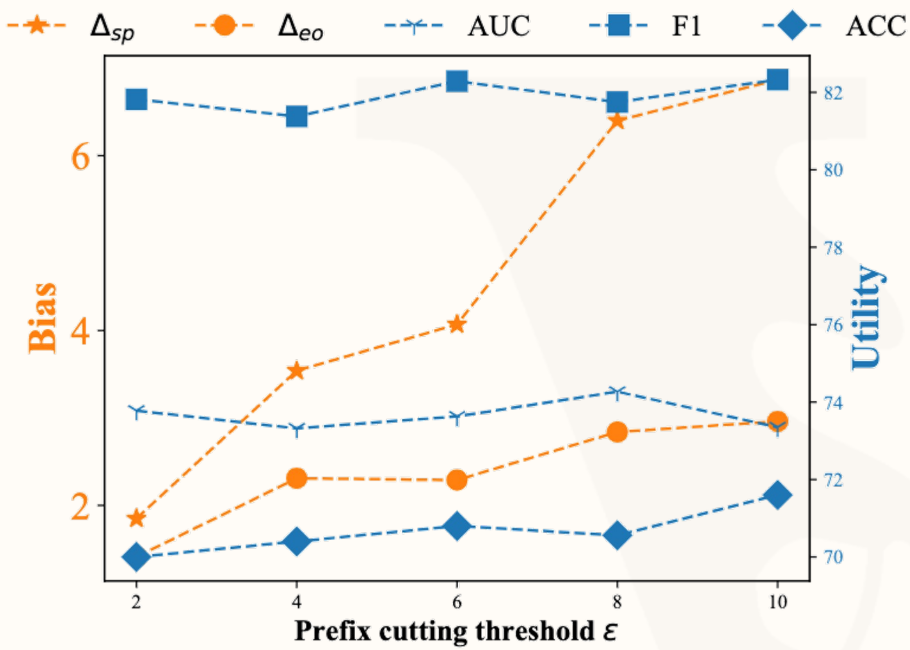
# FairVGNN – Adversarial training



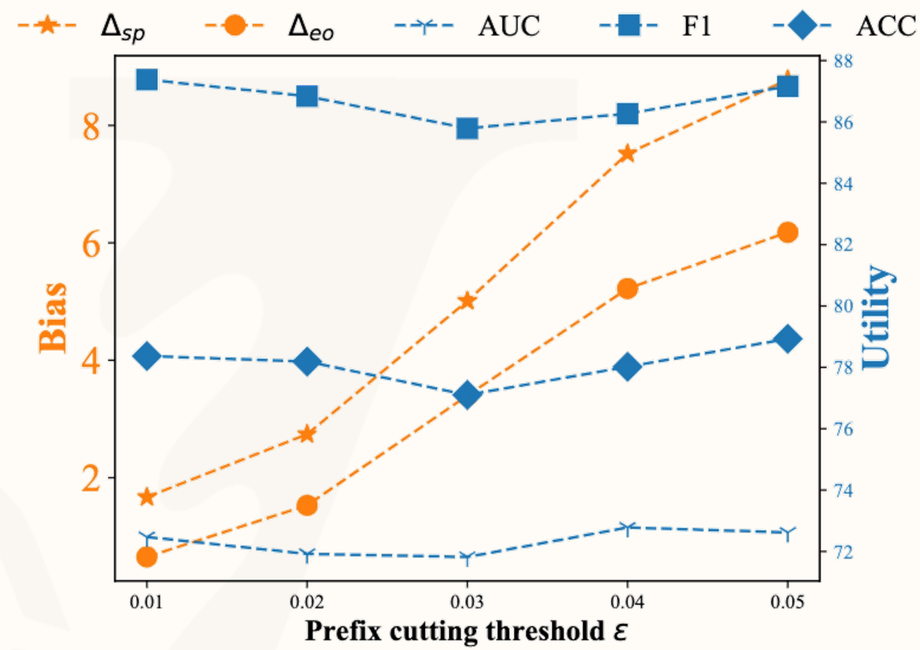

(1) Removing either the generator or the discriminator would lower the fairness

(2) Removing the discriminator causes the highest bias

# FairVGNN – Adaptive weight clamping



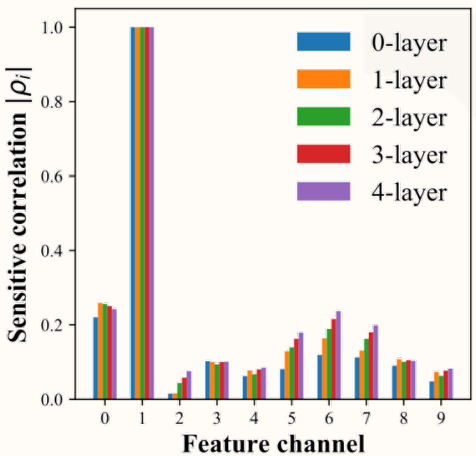

(a) German

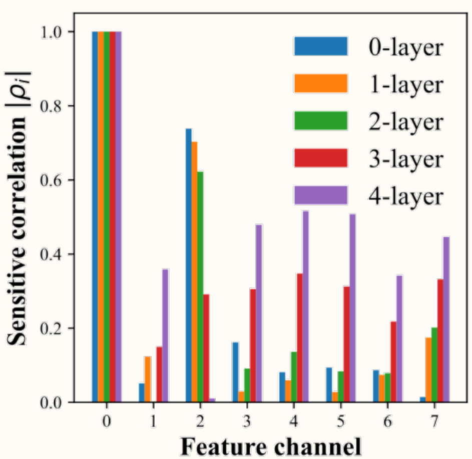



(b) Credit

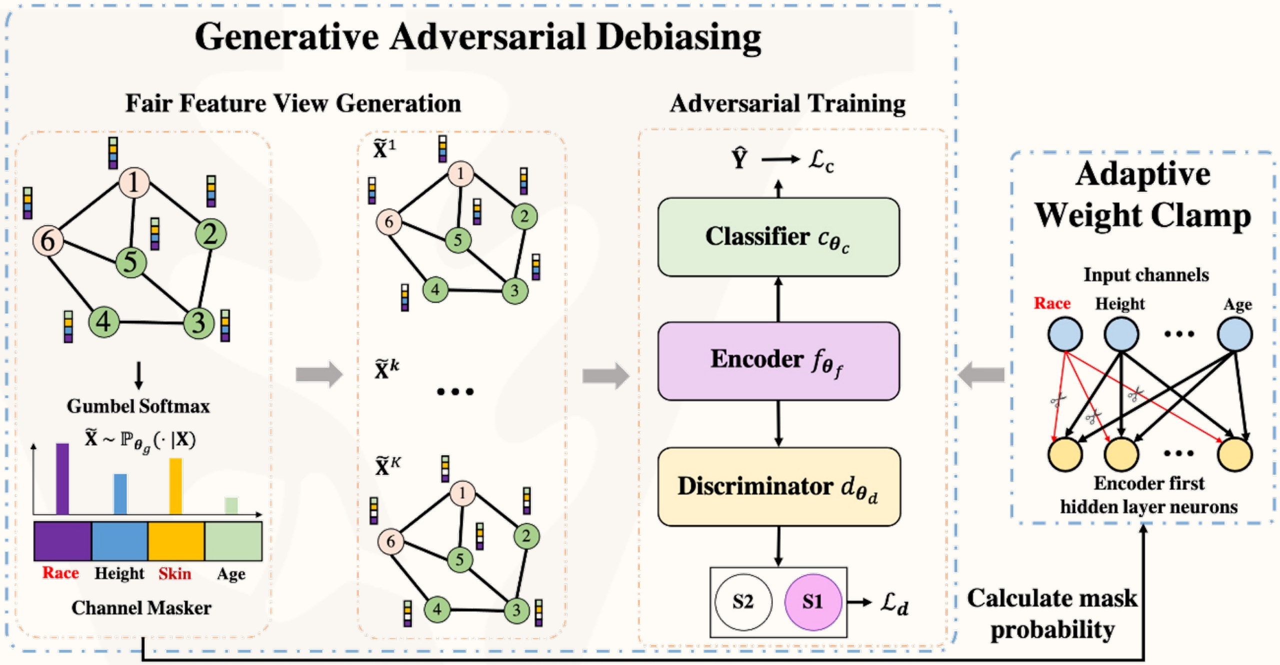
# Contributions

## Novel problem




## Solution: FairVGNN

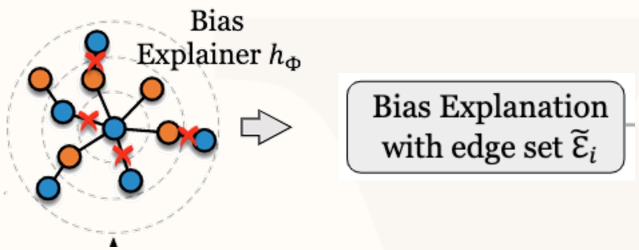



## New Finding: bias and homophily

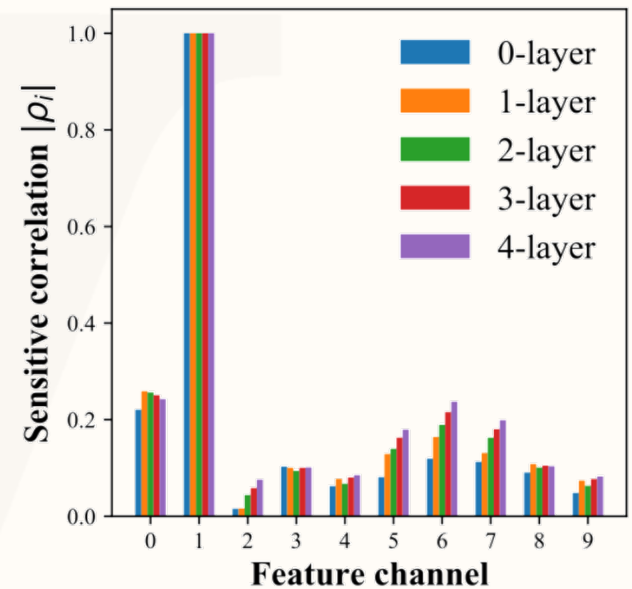
$$||\boldsymbol{\mu}||_2 = ||(2\chi - 1)\mathbf{W}^{f,1}\Delta\boldsymbol{\mu}||_2 \leq (2\chi - 1)\Big(\sum_{i=1}^{d_1}\big(\sum_{r\in\mathcal{S}}\epsilon \mathbf{p}_r\Delta\boldsymbol{\mu}_r + \sum_{k\in\mathcal{NS}}\epsilon \mathbf{p}_k\Delta\boldsymbol{\mu}_k\big)^2\Big)^{0.5}$$

# Concurrent and Future work

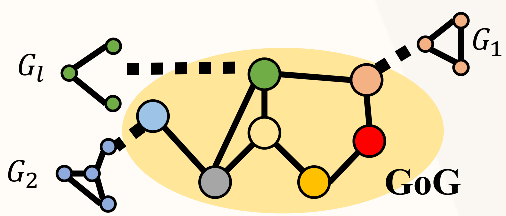
## Structural explanation for bias (KDD 22)



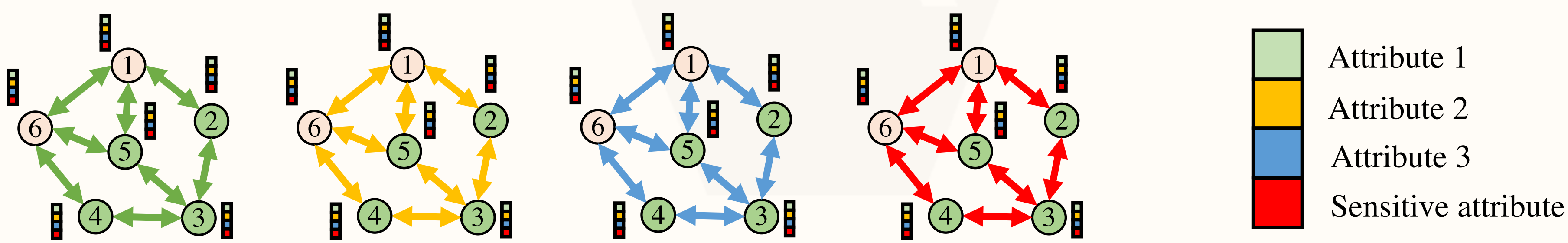

## Correlation variation




## Mitigating Labeling bias (CIKM 22)

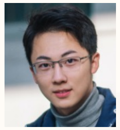



## Network channel homophily, propagation and fairness

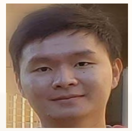


# Acknowledgement

Yuying Zhao



Yushun Dong



Huiyuan Chen

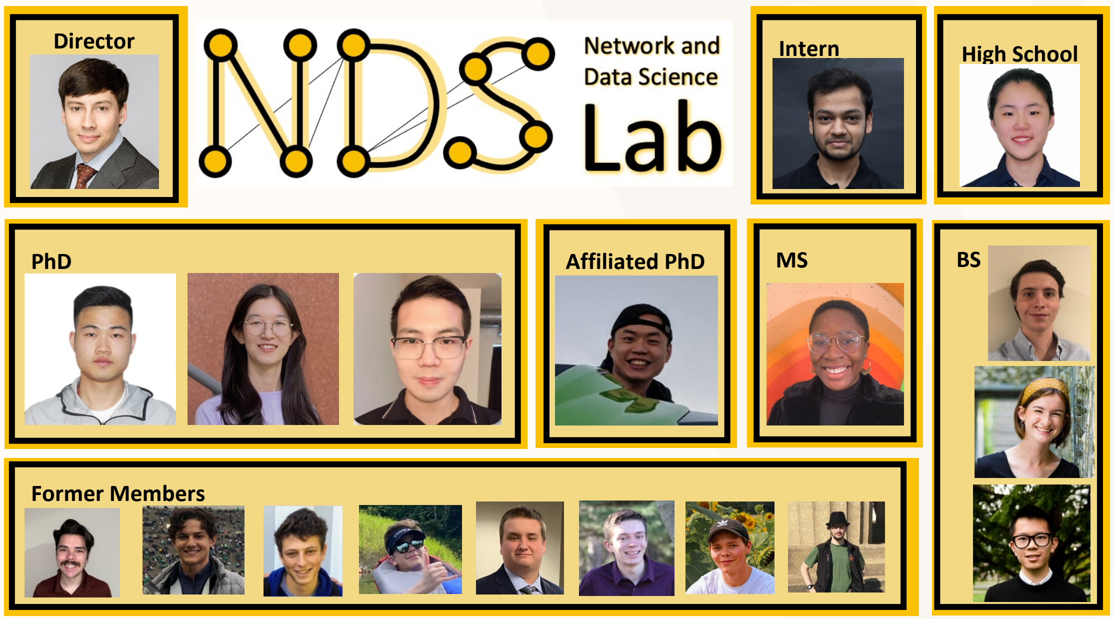
Jundong Li

Tyler Derr

KDD 2022

Association for Computing Machinery

NSF




https://nds-vu.github.io/

More about me

https://yuwvandy.github.io/