Demystifying and Mitigating Unfairness for Learning over Graphs

Oyku Deniz, Kose, **Yanning Shen**

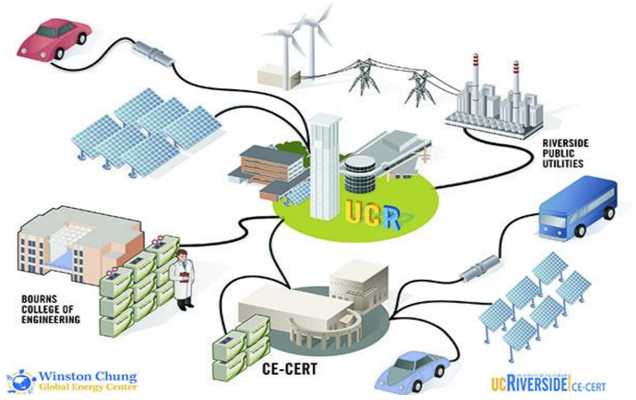
Electrical Engineering and Computer Science
University of California, Irvine

This DEGAS Webinar
Jan 22, 2025

Networks Everywhere

Social Networks

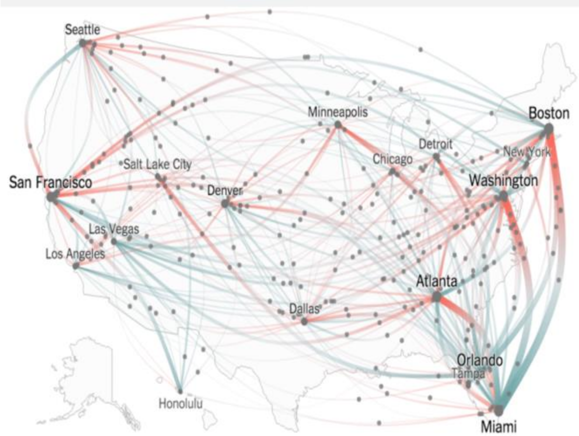
Financial Networks



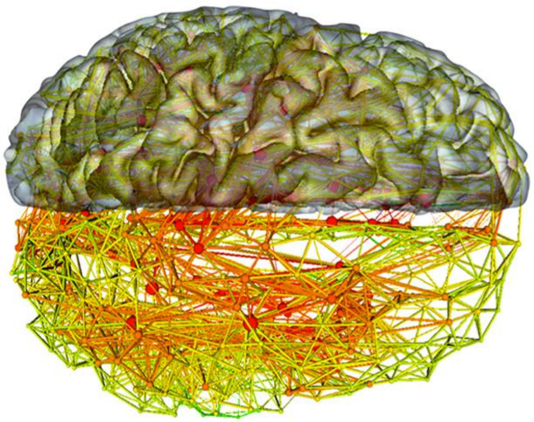
Energy Grids

Flight Networks

Graphs Definition



Flight Networks



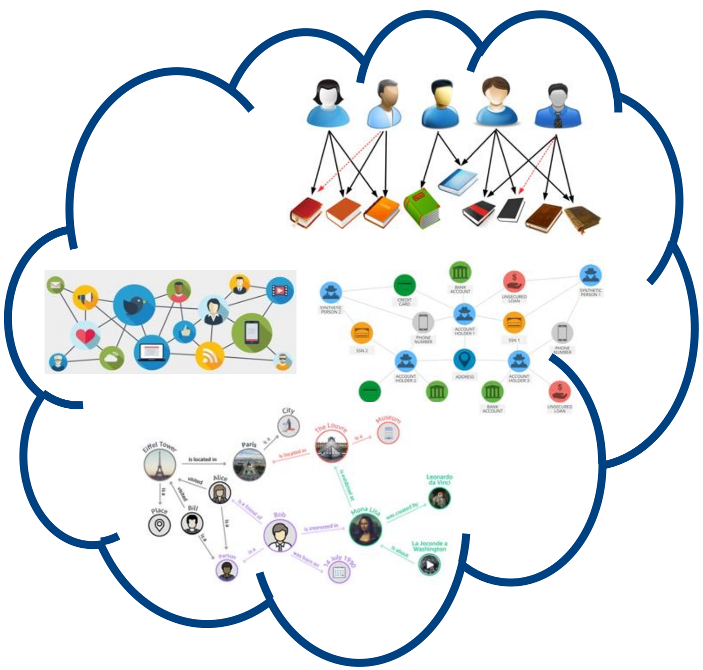
Brain networks

Traffic networks

- **Graphs :** mathematical structures to model pair-wise relations
  - **Nodes:** airports in flight networks, neurons in brain networks
  - **Edges:** flight paths between airports, roads between intersections
  - **Nodal features:** weather in airports, types of neurons (sensory/motor)

Graphs Machine Learning Algorithms

- Extract **information encoded in the graph data**
- Facilitate understanding on information over network graphs
- Gain benefits on various predictive tasks.



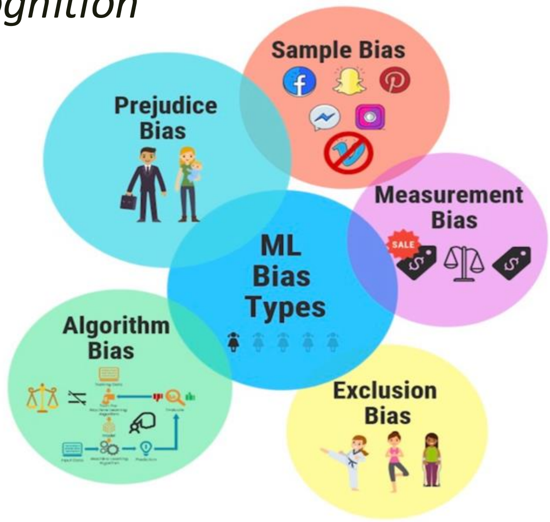
Graph ML Algorithms

- Who are potential friends?
- Which item will this customer buy?
- Which loan applicant is with the lowest risk of debt default?

…

Unfairness in Machine Learning

- ML algorithms may lead to unfair results
  - *Different error rates on female/male faces in face recognition*
  - *Different crime prediction accuracy based on ethnicity*
  - *Different credit approval rates based on gender*



- Critical for various applications and policy making

- Extensive literature on (non-graph) bias/unfairness reduction in ML
  - e.g.,[Zafar et. al., 2015][Du et. al., 2020][Zhang et al 2020][Dutta.et al., 2O21]

Group fairness Notions

- **Statistical Parity:** considers achieving the same positive rate for individuals in different sensitive subgroups.

$$\Delta_{SP} = |\mathrm{P}(\hat{Y} = 1 | S = 0) - \mathrm{P}(\hat{Y} = 1 | S = 1)|$$

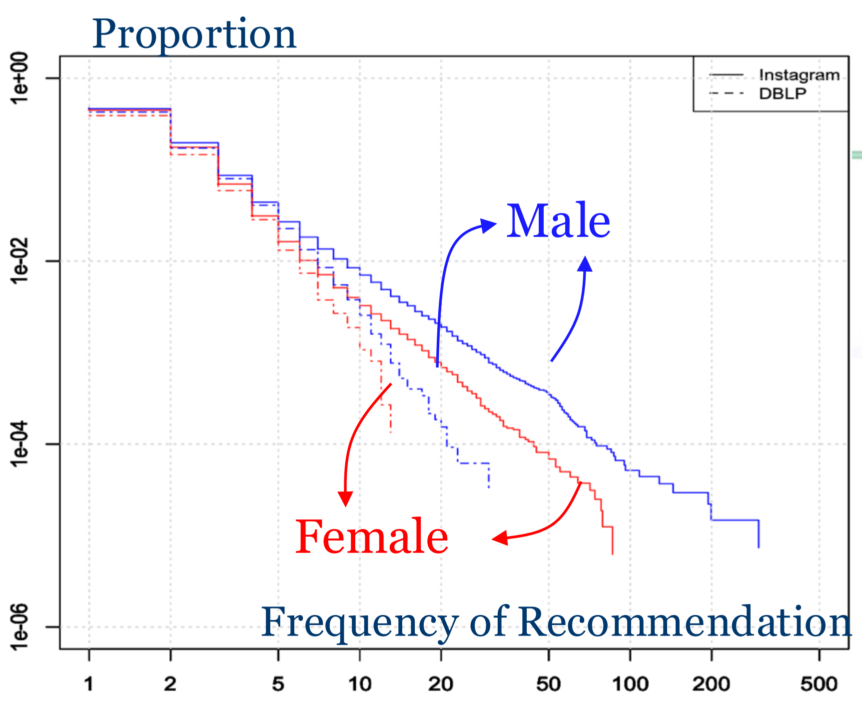
- **Equality Opportunity:** the same **true** positive rates are enforced between sensitive subgroups

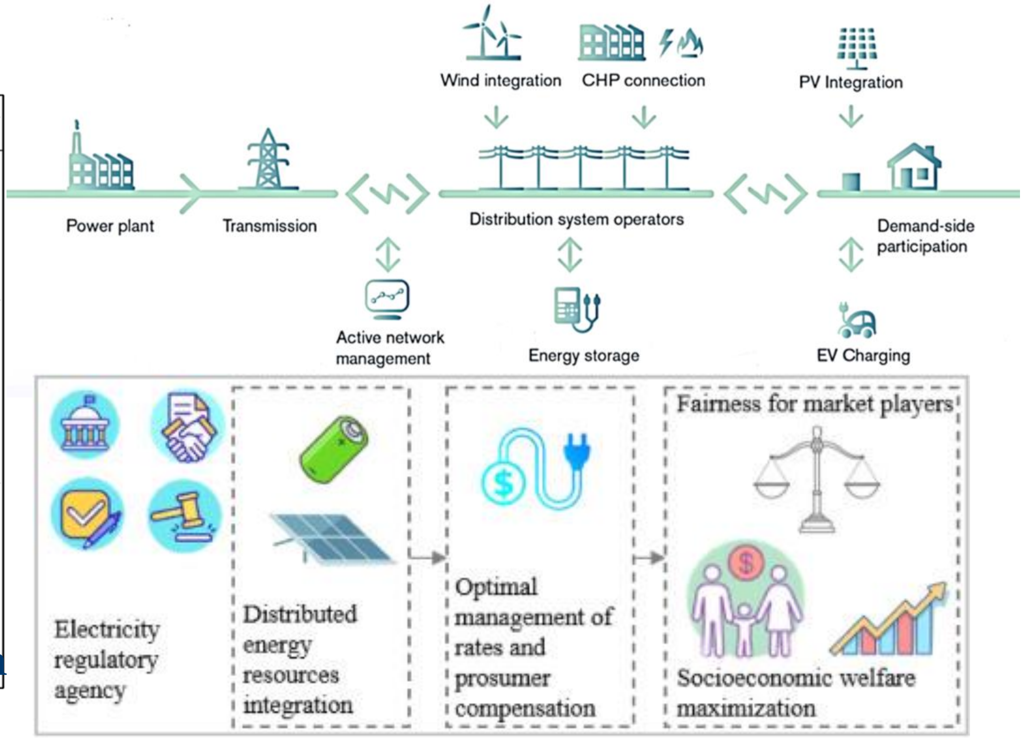
$$\Delta_{EO} = |P(\hat{Y} = 1 | S = 0, Y = 1) - P(\hat{Y} = 1 | S = 1, Y = 1)|$$

- **Smaller Δ_{SP} and Δ_{EO} are more desirable**
- **Key intuition:** Decision making **uncorrelated** with sensitive attributes
- Generalizable to graph domain

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." InNeurIPS, 2016.
[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." InNeurIPS, 2016.

Potential Unfairness in Networks





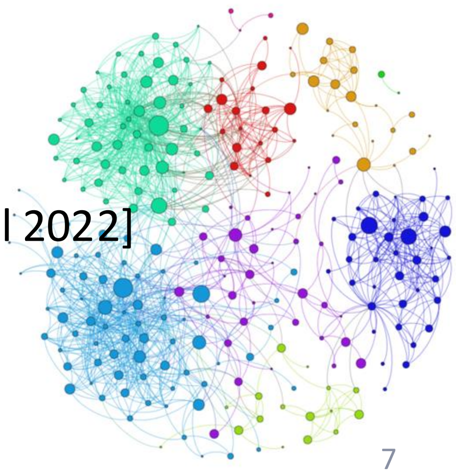
Users get recommended to be connected exhibit divergence between genders [1].

Unfairness in user classification and resource allocation in power grids[2,3]

[1] Stoica, Ana-Andreea, et al. "Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity." In WWW 2018.
[2]. R. Du, D. Muthirayan, P. P Khargonekar, Y. Shen, "Long-term Fairness For Real-time Decision Making: A Constrained Online Optimization Approach" *IEEE Transactions on Neural Networks and Learning Systems,* accepted Oct 2024.
[3] R. D, and Y. Shen. "Fairness-aware User Classification in Power Grids." 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022

Unfairness in ML over Graphs

- Graph structure has intrinsic bias
  - Higher probability for the connections between similar users(religion, ethnicity)
- Learning over graphs amplifies already existing bias
- Information aggregation over neighbors in GNNs $\longrightarrow$ Indirect use of sensitive attributes in training!
- Fairness is in graph domain.
  - Random walk-based: [Rahman et al., 2019]
  - Fairness constraints: [Zafar et al., 2019]
  - Adversarial regularization-based: [Dai & Wang, 2020]
  - Individual fairness [Xu et al 2023], graph cut [Dinitz et al 2022]
- **Theoretical** understanding is **largely missing**, and mostly designed for **specific learning tasks**



Unbalanced Real Network Topologies

- **Pokec datasets:** *Facebook-like real social networks*

Dataset	Pokec-z	Pokec-n
# Nodes	7659	6185
# Nodes with S=0	4851	4040
# Nodes with S=1	2808	2145
# Edges	29476	21844
# Features	59	59
# Intra-group edges	28336	20901
# Inter-group edges	1140	943

Severely unbalanced edges → potential bias

- Higher probability for the connections between similar users (religion, ethnicity)

Graph Neural Networks

- **Question:** Can we explain the source of bias?
- Graph neural networks: $\mathbf{H}^l = \sigma(\mathbf{D}^{-1}(\mathbf{A}+\mathbf{I})\mathbf{H}^{l-1}\mathbf{W}^{l-1})$

$A_{ij} = 1$ if nodes i and j connected

$$\boxed{\mathbf{h}_i^l = \sigma\left(\left(\frac{1}{D_{ii}}\sum_{j\in\mathcal{N}_i}\mathbf{h}_j^{l-1}\right)\mathbf{W}^{l-1}\right)}$$

$\mathbf{D} \in \mathbb{R}^{N\times N}$: degree matrix

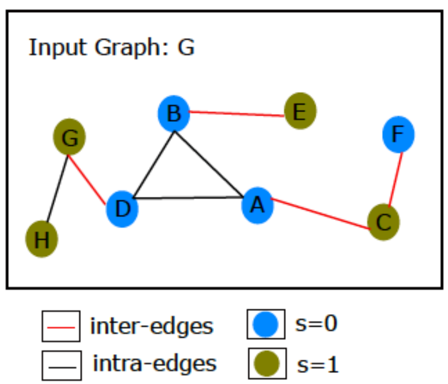
$\mathbf{W}^l$: weight matrix

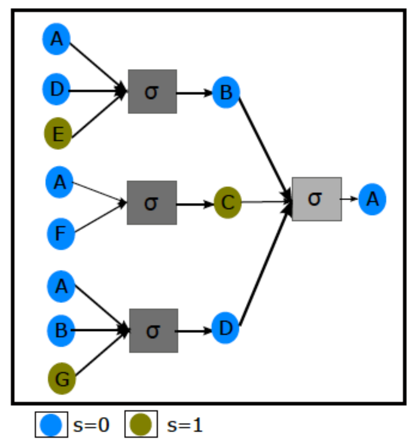
$\sigma(.)$: non-linear act

$\mathbf{H}^l$: trained node re

$\mathbf{H}^0 = \mathbf{X} \in \mathbb{R}^{N\times F}$

$\mathcal{N}_i$: neighbor set o





$\mathbf{Z}^{l-1} := \mathbf{D}^{-1}(\mathbf{A}+\mathbf{I})\mathbf{H}^{l-1}$: aggregated representation

Source of Bias

- **Idea:** measure the **correlation** between **aggregated representation** $\mathbf{z}_{:,i}$ and **sensitive attributes** $\mathbf{s} \in \mathbb{R}^N$
- **Approach :** Bound $||\boldsymbol{\rho}||_1$ with $\rho_i = \mathrm{Corr}(\mathbf{z}_{:,i}, \mathbf{s})$ for $i = \{1 \cdots F\}$

Theorem. $||\boldsymbol{\rho}||_1 \leq ||\mathbf{c}||_1 (||\boldsymbol{\delta}||_1 \max(\gamma_1, \gamma_2) + 2N\Delta)$

- $$\boldsymbol{\delta} := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_j := \mathbb{E}_{\mathbf{h}_n \sim U} \left[\overbrace{\mathbf{x}_n}^{\text{features for node n}} \mid n \in \overbrace{\mathcal{S}_j}^{\text{set of nodes with sensitive attribute j}} \right], \quad j = \{0, 1\}$$
- $$\gamma_1 := \left| 1 - \frac{\overbrace{|\mathcal{S}_0^\chi|}^{\text{nodes with at least one inter-edge}}}{|\mathcal{S}_0|} - \frac{|\mathcal{S}_1^\chi|}{|\mathcal{S}_1|} \right| \qquad \mathcal{S}_j = \mathcal{S}_j^\chi \cup \overbrace{\mathcal{S}_j^\omega}^{\text{nodes with no inter-edge}}, j = \{0, 1\}$$
- $$\gamma_2 = \left| 1 - 2\min\left(\mathrm{mean}\left(\frac{\overbrace{d_m^\chi}^{\text{Number of inter edges of node m}}}{d_m^\chi + \underbrace{d_m^\omega}_{\text{intra edges of node m}}} \middle| v_m \in \mathcal{S}_0 \right), \mathrm{mean}\left(\frac{d_n^\chi}{d_n^\chi + d_n^\omega} \middle| v_n \in \mathcal{S}_1 \right) \right) \right|$$

O. D. Kose and Y. Shen, "Demystifying and Mitigating Bias for Node Representation Learning", accepted to IEEE Transactions on Neural Networks and Learning Systems, April 2023.

Fairness-Aware Augmentation Design

- **Goal:** Design augmentation strategies $\mathcal{G}(\mathcal{V},\mathcal{E}) \rightarrow \tilde{\mathcal{G}}(\tilde{\mathcal{V}},\tilde{\mathcal{E}})$ and $\mathbf{X} \rightarrow \tilde{\mathbf{X}}$ to reduce $||\boldsymbol{\rho}||_1$

$$\boldsymbol{\delta} := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \qquad \boldsymbol{\mu}_j := \mathbb{E}_{\mathbf{h}_n \sim U}\left[\overbrace{\mathbf{x}_n}^{\text{features of node n}} \mid n \in \overbrace{\mathcal{S}_j}^{\text{set of nodes with sensitive attribute j}}\right], \quad j = \{0,1\}$$

feature masking

$$\gamma_1 := \left|1 - \frac{\overbrace{|\mathcal{S}_0^{\chi}|}^{\text{Nodes with at least one inter-edge}}}{|\mathcal{S}_0|} - \frac{|\mathcal{S}_1^{\chi}|}{|\mathcal{S}_1|}\right| \qquad \mathcal{S}_j = \mathcal{S}_j^{\chi} \cup \overbrace{\mathcal{S}_j^{\omega}}^{\text{no inter-edge}}, j = \{0,1\}$$

node sampling

$$\gamma_2 = \left|1 - 2\min\left(\text{mean}\left(\frac{\overbrace{d_m^{\chi}}^{\text{\# inter edges of node m}}}{d_m^{\chi} + \underbrace{d_m^{\omega}}_{\text{intra degree of node m}}} \middle| v_m \in \mathcal{S}_0\right), \text{mean}\left(\frac{d_n^{\chi}}{d_n^{\chi} + d_n^{\omega}} \middle| v_n \in \mathcal{S}_1\right)\right)\right|$$

edge augmentation

Node Classification

- **Performance metric:** Accuracy, Area Under the Curve(AUC)
- **Fairness metrics:**

statistical parity: $$\Delta_{SP} = |P(\hat{y} = 1 \mid s = 0) - P(\overbrace{\hat{y}}^{\text{prediction class label}} = 1 \mid s = 1)|$$

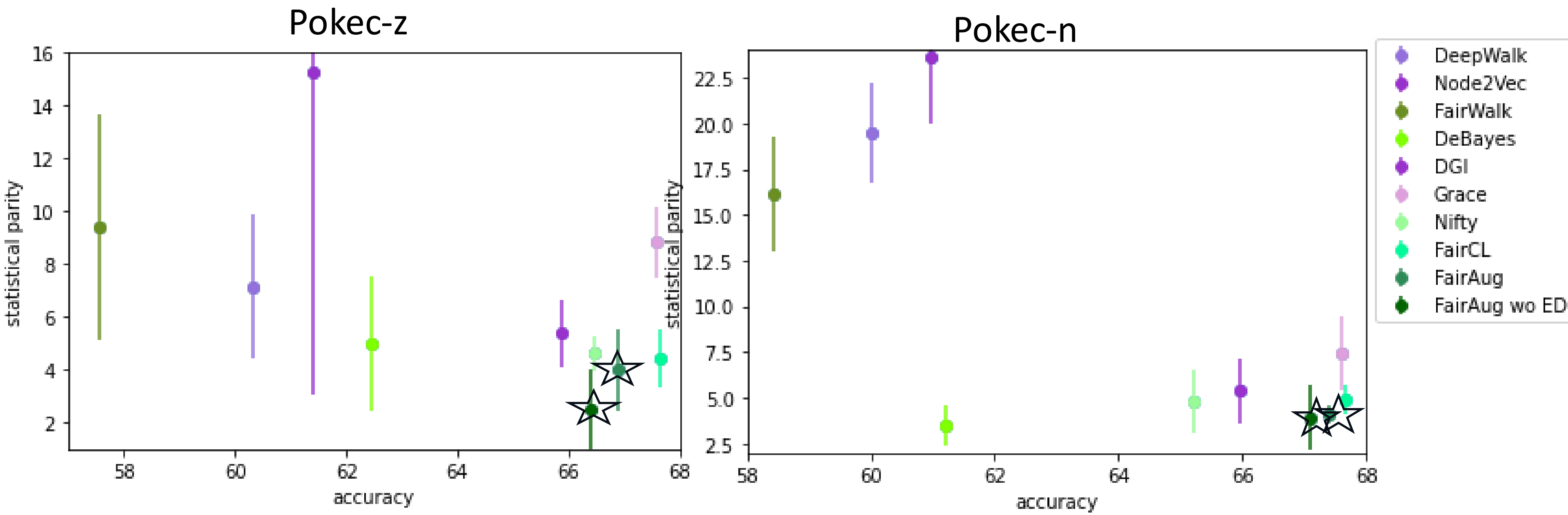
equal opportunity: $$\Delta_{EO} = |P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid \overbrace{y}^{\text{true class label}} = 1, s = 1)|$$

- **Datasets:** Real social networks

Dataset	$\lvert\mathcal{S}_0^{\chi}\rvert$	$\lvert\mathcal{S}_0^{\omega}\rvert$	$\lvert\mathcal{S}_1^{\chi}\rvert$	$\lvert\mathcal{S}_1^{\omega}\rvert$	$\lvert\mathcal{E}^{\chi}\rvert$	$\lvert\mathcal{E}_{\mathcal{S}_0}^{\omega}\rvert$	$\lvert\mathcal{E}_{\mathcal{S}_1}^{\omega}\rvert$
Pokec-z	622	4229	582	2226	1730	23428	15942
Pokec-n	423	3617	479	1666	1422	18548	10672

UCI Samueli
School of Engineering

Node Classification Results



- Lower right -> better
- All green tones for fairness-aware baselines

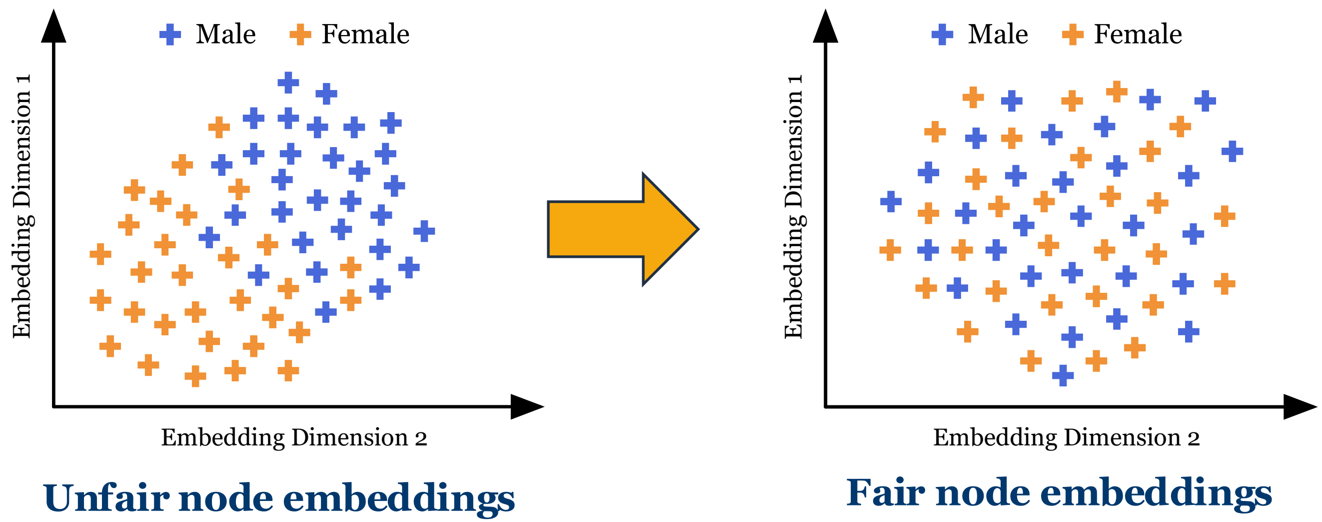
Our framework always outperforms state-of-art baselines!

Challenge: Preprocessing/augmenting data causes loss of useful information that cannot be retrieved in training

Idea: Fair normalization as in-processing

Fair Normalization

- **Theoretical Analysis:** bias in GNNs related to **distributions** of representations



Idea: shift group-wise distributions in each layer to reduce unfairness
Approach: **Fairness-aware** group-wise **trainable** batch normalization

O. D. Kose and Y. Shen, ”FairNorm: Fair and Fast Graph Neural Network Training,“ Transactions on Machine Learning Research (TMLR) May 2023.

Multiple Group-wise Normalization

- **Key Idea:** Apply **trainable** normalizations over different sensitive groups

$$\text{M-Norm}\left(a_{i,j}^{(n)}\right) = \gamma_i^{(n)} \cdot \frac{a_{i,j}^{(n)} - \alpha_i^{(n)} \cdot m_i^{(n)}}{\sigma_i^{(n)}} + \beta_i^{(n)}$$

- Normalization is applied after linear transformations

$$\mathbf{H}^{(n)} = \text{Act}\left(\text{M-Norm}^{(n)}\left((\mathbf{WHQ})^{(n)}\right)\right)$$

Acts as a preconditioner, provides **provably faster convergence**

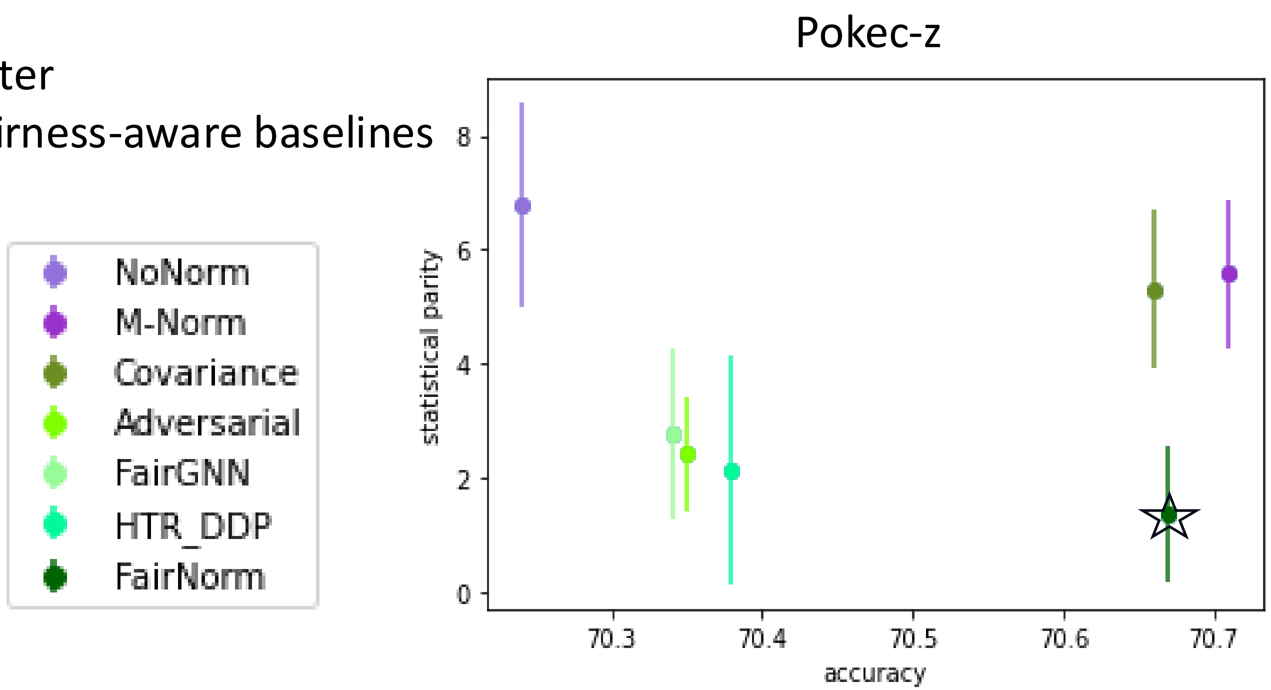
O. D. Kose and Y. Shen, " Fast&Fair: Training Acceleration and Bias Mitigation for GNNs", accepted by Transactions on Machine Learning Research (TMLR) May 2023.

Node Classification Results

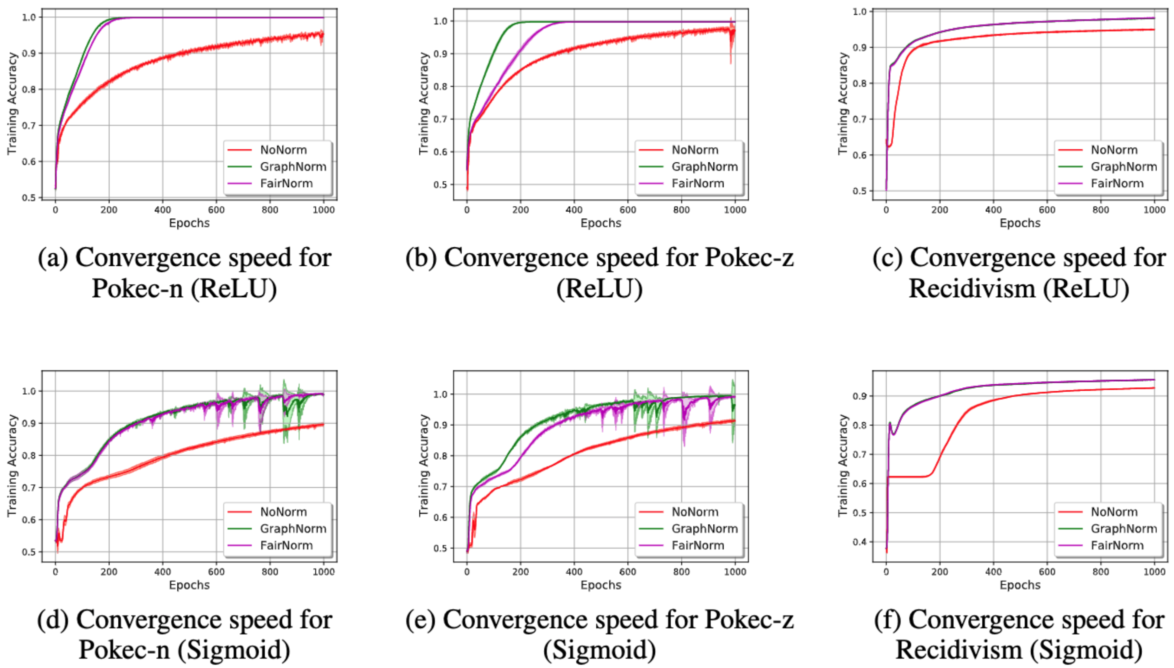
Statistical Parity

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

- Lower right -> better
- Green tones for fairness-aware baselines



Training convergence



(a) Convergence speed for Pokec-n (ReLU)

(b) Convergence speed for Pokec-z (ReLU)

(c) Convergence speed for Recidivism (ReLU)

(d) Convergence speed for Pokec-n (Sigmoid)

(e) Convergence speed for Pokec-z (Sigmoid)

(f) Convergence speed for Recidivism (Sigmoid)

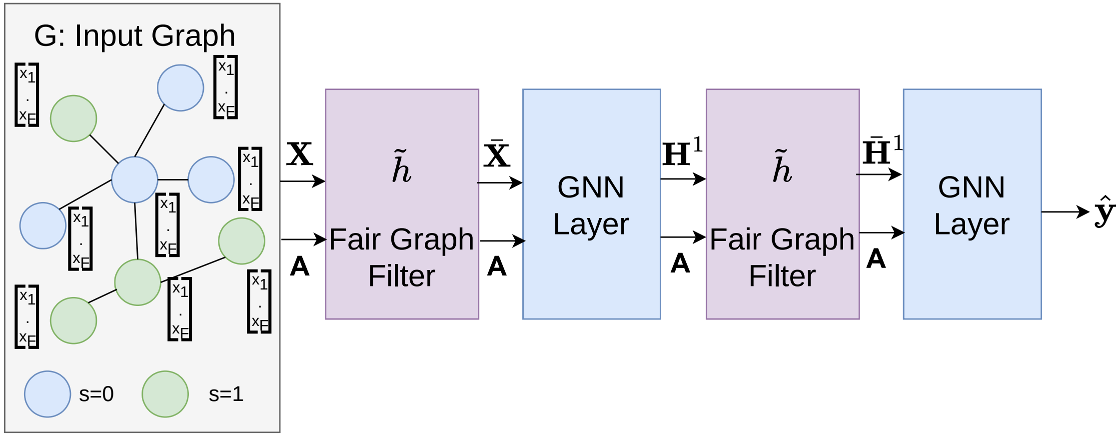
Figure 1: Convergence speed for different graph data sets when the normalization is not applied (Nonorm) and applied with/without fairness consideration (FairNorm/GraphNorm).

Provably faster convergence than NoNorm.

O. D. Kose and Y. Shen, "FairNorm: Fair and Fast Graph Neural Network Training," Transactions on Machine Learning Research (TMLR) May 2023.

Fairness-aware Graph Filtering Design

- **Idea**: Design Graph filter to filter out the bias
- **Analysis:** Graph frequency domain correlation with bias
- **Approach**: Filter out graph frequency that are correlated with bias



- Filter is pre-computed, no modification in training
  - Can be used as pre-trained bias mitigation operators before GNN layers
  - Analogy to batch normalization layers

O. D. Kose, G. Mateos, and Y. Shen, “Fair Graph Filter Design", *57th Asilomar Conference on Signals, Systems, and Computers*. IEEE 2023.
O. D. Kose, G. Mateos, and Y. Shen, “Fairness-aware Optimal Graph Filter Design”, *JSTSP,* 2024.

Question: What if we do not want to share real training data?

Idea: Graph generative models come to rescue!

Generative Models Amplify Structural Bias

- Create synthetic graphs with a SOTA diffusion model, GraphMaker [Li et al., 2023]

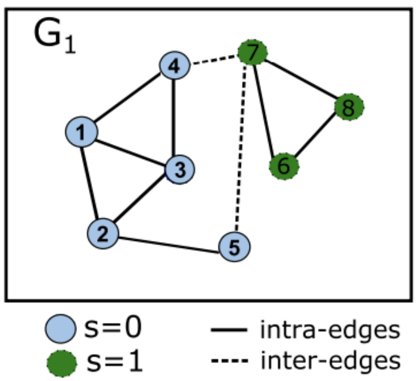
Cora	Accuracy (%)	$\Delta_{SP}(\%)$	$\Delta_{EO}(\%)$
$\mathcal{G}$	94.92	27.71	11.53
GraphMaker	87.29 ± 1.09	35.72 ± 1.74	13.27 ± 0.81
Citeseer	Accuracy (%)	$\Delta_{SP}(\%)$	$\Delta_{EO}(\%)$
$\mathcal{G}$	95.76	29.05	9.53
GraphMaker	92.19 ± 1.06	37.56 ± 1.29	13.52 ± 0.92
Amazon Photo	Accuracy (%)	$\Delta_{SP}(\%)$	$\Delta_{EO}(\%)$
$\mathcal{G}$	96.91	32.58	8.24
GraphMaker	94.45 ± 0.21	33.49 ± 0.28	10.01 ± 0.56
Amazon Computer	Accuracy (%)	$\Delta_{SP}(\%)$	$\Delta_{EO}(\%)$
$\mathcal{G}$	96.14	22.90	4.63
GraphMaker	94.04 ± 0.26	23.56 ± 0.55	6.23 ± 0.49

Using generated graph **increases unfairness!**

Sources of Structural Bias

Theorem:

$$\Delta_{SP} \propto \quad \alpha_1 := \left| \frac{p_k^{\omega}}{|\mathcal{S}_k|} - \frac{p_k^{\chi}}{N-|\mathcal{S}_k|} \right| \quad \text{and} \quad \alpha_2 := \left| \frac{\sum_{v_i, v_j \in \mathcal{V}} \tilde{A}_{ij} - p_k^{\omega} - 2p_k^{\chi}}{N-|\mathcal{S}_k|} - \frac{p_k^{\chi}}{|\mathcal{S}_k|} \right|$$



$$\underbrace{p_k^{\chi}}_{\text{E[\# inter-edges]}} := \sum_{v_i \in \mathcal{S}_k, v_j \notin \mathcal{S}_k} \tilde{A}_{i,j}, \quad \underbrace{p_k^{\omega}}_{\text{E[\# intra-edges]}} := \sum_{v_i \in \mathcal{S}_k, v_j \in \mathcal{S}_k} \underbrace{\tilde{A}_{i,j}}_{\text{Stochastic graph view}}$$

Intuition: balance between inter/intra-edges is desirable

O. D. Kose and Y. Shen, "FairWire: Fair Graph Generation", accepted to NeurIPS 2024

Novel Fair Regularizer Design

Theorem:

$$\Delta_{SP} \propto \quad \alpha_1 := \left| \frac{p_k^{\omega}}{|\mathcal{S}_k|} - \frac{p_k^{\chi}}{N-|\mathcal{S}_k|} \right| \quad \text{and} \quad \alpha_2 := \left| \frac{\sum_{v_i, v_j \in \mathcal{V}} \tilde{A}_{ij} - p_k^{\omega} - 2p_k^{\chi}}{N-|\mathcal{S}_k|} - \frac{p_k^{\chi}}{|\mathcal{S}_k|} \right|$$

Proposed Regularizer:

Batch of nodes

One-hot representation for sensitive attributes

$$\mathcal{L}_{\text{FairWire}}\,(\tilde{\mathbf{A}}, \mathcal{B}) := \sum_{k=0}^{K} \left| \frac{\sum_{v_i, v_j \in \mathcal{B}} \left(\tilde{\mathbf{A}} \odot (\mathbf{S}\mathbf{e}_k)(\mathbf{S}\mathbf{e}_k)^{\top}\right)_{ij}}{|\mathcal{S}_k|} - \frac{\sum_{v_i, v_j \in \mathcal{B}} \left(\tilde{\mathbf{A}} \odot (\mathbf{S}\mathbf{e}_k)(\mathbf{1} - (\mathbf{S}\mathbf{e}_k))^{\top}\right)_{ij}}{N-|\mathcal{S}_k|} \right|$$

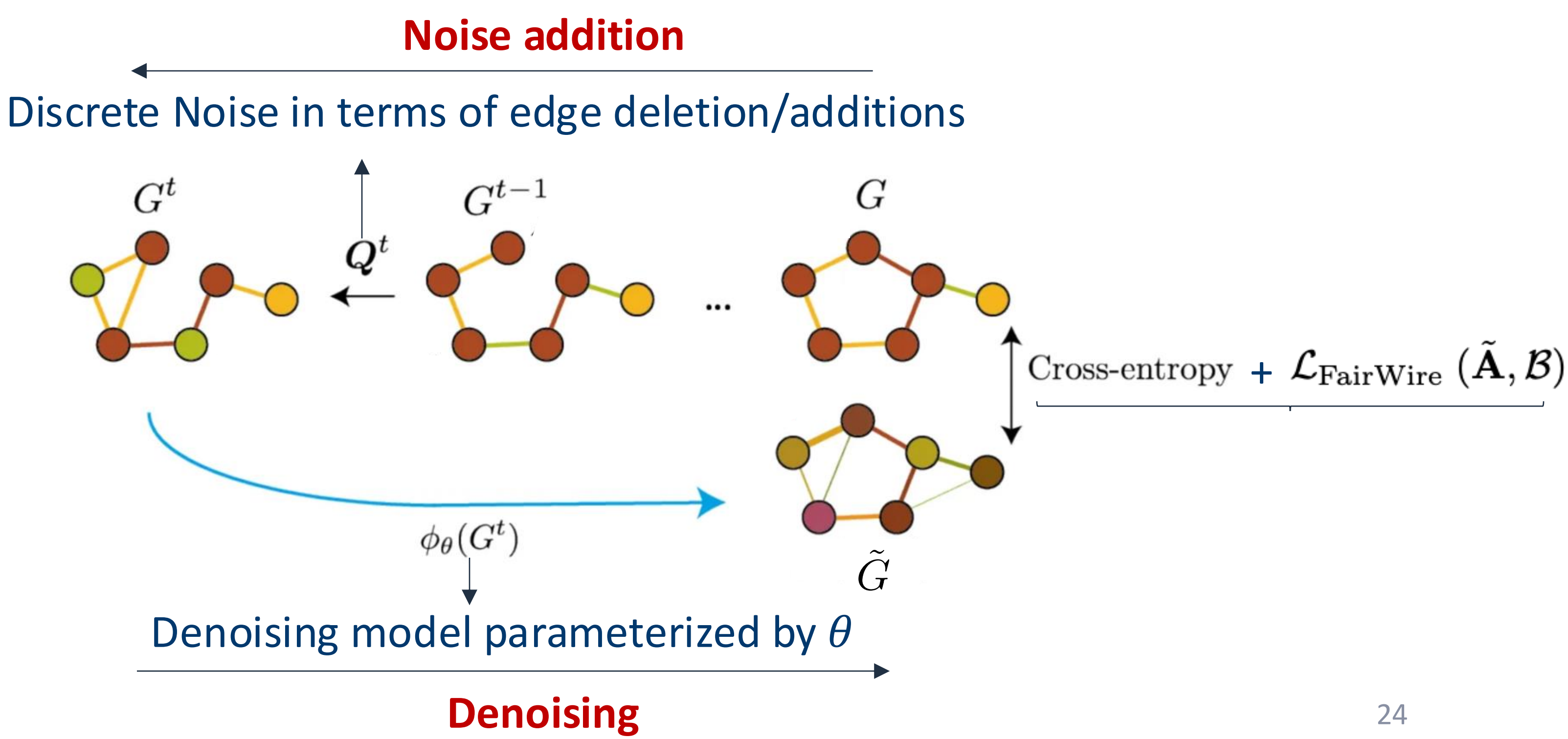
Allows a minibatch application

- Can be applicable to **any model outputting probabilities** for edges in graph
  - GNN training for link prediction
  - Graph generative models

O. D. Kose and Y. Shen, "FairWire: Fair Graph Generation", accepted to NeurIPS 2024

FairWire: Fair Synthetic Graph Generation

- **FairWire**: a diffusion model for graph generation together with sensitive attributes
  - Generate graphs with similar distribution to original G alleviated structural bias
  - Allows fair model training without sharing sensitive information



Experimental Settings

- Results obtained over 6 real-world datasets

Dataset	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	F	K
Cora	2708	10556	1433	7
Citeseer	3327	9228	3703	6
Amazon Photo	7650	238163	745	8
Amazon Computer	13752	491722	767	10
Credit	1000	22242	27	2
Pokec-n	6185	21844	59	2

- Fairness metrics:

Node classification

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

$$\Delta_{EO} = |P(\underbrace{\hat{y}}_{\text{model predictions}} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid \underbrace{y}_{\text{labels}} = 1, s = 1)|$$

Link prediction

$$\Delta_{SP} = |P(\hat{y} = 1 \mid e \in \mathcal{E}^{\chi}) - P(\hat{y} = 1 \mid e \in \mathcal{E}^{\omega})|$$

$$\Delta_{EO} = |P(\hat{y} = 1 \mid y = 1, e \in \underbrace{\mathcal{E}^{\chi}}_{\text{set of inter-edges}}) - P(\hat{y} = 1 \mid y = 1, e \in \overbrace{\mathcal{E}^{\omega}}^{\text{set of intra-edges}})|$$

Graph Generation Evaluation

- Link prediction and node classification models trained on generated graphs
- Evaluated on same real graphs

Link Prediction

	Cora			Citeseer		
	AUC (%)	Δ_{SP} (%)	Δ_{EO} (%)	AUC (%)	Δ_{SP} (%)	Δ_{EO}(%)
$\mathcal{G}$	94.92	27.71	11.53	95.76	29.05	9.53
$\tilde{\mathcal{G}}$	$\mathbf{87.29} \pm 1.09$	35.72 ± 1.74	13.27 ± 0.81	$\mathbf{92.19} \pm 1.06$	37.56 ± 1.29	13.52 ± 0.92
FairAdj	82.13 ± 1.07	15.47 ± 2.39	6.26 ± 2.05	82.67 ± 2.78	$\mathbf{15.45} \pm 2.68$	7.98 ± 1.47
Adversarial	83.66 ± 5.64	16.35 ± 9.80	7.82 ± 5.84	89.59 ± 2.70	24.20 ± 5.82	10.34 ± 1.66
FairWire	86.49 ± 2.79	$\mathbf{12.91} \pm 6.35$	$\mathbf{4.31} \pm 3.59$	91.27 ± 2.78	18.35 ± 6.91	$\mathbf{7.80} \pm 2.76$

Node Classification

	German			Pokec-n		
	Acc (%)	Δ_{SP} (%)	Δ_{EO} (%)	Acc (%)	Δ_{SP} (%)	Δ_{EO}(%)
$\mathcal{G}$	70.00	2.13	1.78	68.73	8.58	9.68
FairGen	**73.60**	28.71	15.34	51.73	0.00	0.00
$\tilde{\mathcal{G}}$	68.92 ± 2.37	2.61 ± 5.83	2.29 ± 5.06	66.19 ± 2.05	3.63 ± 2.58	2.66 ± 2.50
FairAdj	70.08 ± 1.08	2.17 ± 4.49	1.11 ± 2.24	-	-	-
Adversarial	70.00 ± 0.62	1.57 ± 2.70	1.34 ± 2.86	$\mathbf{69.36} \pm 0.70$	2.16 ± 1.73	2.73 ± 2.01
FairWire	69.76 ± 0.51	$\mathbf{0.63} \pm 1.53$	$\mathbf{0.30} \pm 0.61$	68.23 ± 0.45	$\mathbf{1.91} \pm 0.92$	$\mathbf{1.35} \pm 0.92$

Achieves better fairness/utility trade-off compared to fairness-aware baselines

Conclusions

- **Theoretical analyses for the sources of bias** in multiple GNN frameworks
- **Fair Model Designs:** Multiple fairness-aware strategies: augmentation, normalization.
    - Applicable in different stages of learning (pre-processing, in-processing)
- **Fair Graph Generation:**
    - Diffusion-based fairness-aware generative framework
    - Enables private fair model training without sharing sensitive information
- Experimental results on real-world datasets validate the improvements in fairness measures with similar utility

Related papers

O. D. Kose and Y. Shen, ”FairWire: Fair Graph Generation", NeurIPS 2024

R. Du, D. Muthirayan, P. P Khargonekar, Y. Shen, "Long-term Fairness For Real-time Decision Making: A Constrained Online Optimization Approach" IEEE Transactions on Neural Networks and Learning Systems, Oct 2024.

O. D. Kose, G. Mateos and Y. Shen, ”Fairness-aware Graph Filter Design,” *IEEE Journal of Selected Topics in Signal Processing*, Jan 2024.

O. D. Kose and Y. Shen. "FairGAT: Fairness-aware Graph Attention Networks”, Transactions on Knowledge Discovery from Data (TKDE), 2024.

O. D. Kose and Y. Shen, ” Fast&Fair: Training Acceleration and Bias Mitigation for GNNs”, the *Transactions on Machine Learning Research* (TMLR) May 2023.

O. D. Kose and Y. Shen, ”Demystifying and Mitigating Bias for Node Representation Learning“, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), April 2023.

O. D. Kose, and Y. Shen, “Fairness-aware Graph Contrastive Learning,” O. D. Kose and Y. Shen, IEEE Transactions on Signal and Information Processing over Networks, May 2022.

R. D, and Yanning Shen. "Fairness-aware User Classification in Power Grids." 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022.

Thank you!

UCI Samueli
School of Engineering
University of California, Irvine

Hellman Fellows

Microsoft

Google

National
Science
Foundation

Questions?

Email: yannings@uci.edu