TEXAS A&M UNIVERSITY
Engineering

Fairness in Machine Learning: Metrics and Algorithms

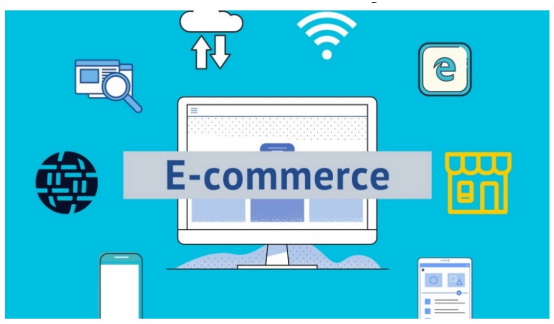
Presenter: Zhimeng Jiang
Date: 10/19/2023

About Me

- Staff Research Scientist at Visa Research
- Just defended my thesis last week at Texas A&M University
- Research Interests
  - Trustworthy ML
  - Graph Neural Networks
  - Generative AI
- Awards
  - ICML 2022 Outstanding Paper Award
  - CIKM 2022 Best Demo Paper Award

Machine Learning Systems are Everywhere

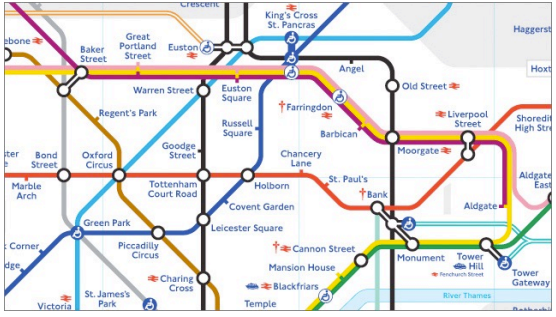
Social Networks



E-commerce Systems

Healthcare Systems

Financial Aid



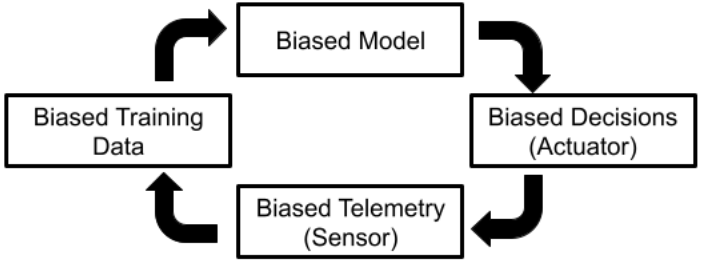
Transportation Networks

Credits Evaluation

Algorithmic Bias in ML Systems

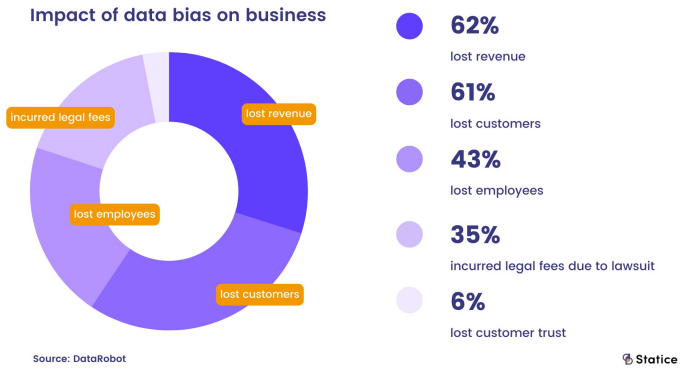
- Ethical challenges posed by ML systems
- Inherent bias presented in society
  - Reflected in training data
  - AI/ML models prone to amplifying such bias

Image recognition systems



Feedback loop
Image from Medium: link

The Consequence of ML Systems' Bias

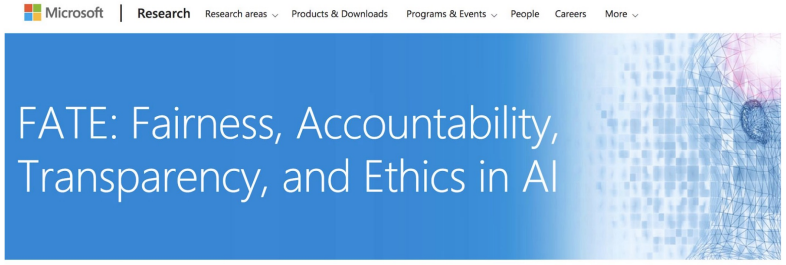


Business Impact

Microsoft's Tay AI chatbot

Bias increases the financial loss and production risk!

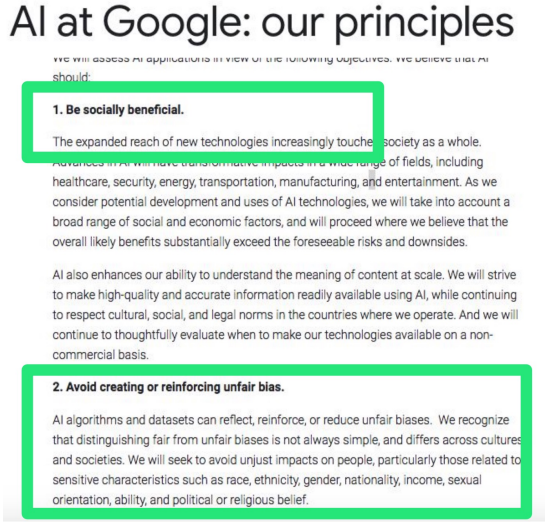
Ethical Regulation for ML Systems



OpenAI

Research Product Developers Safety Company

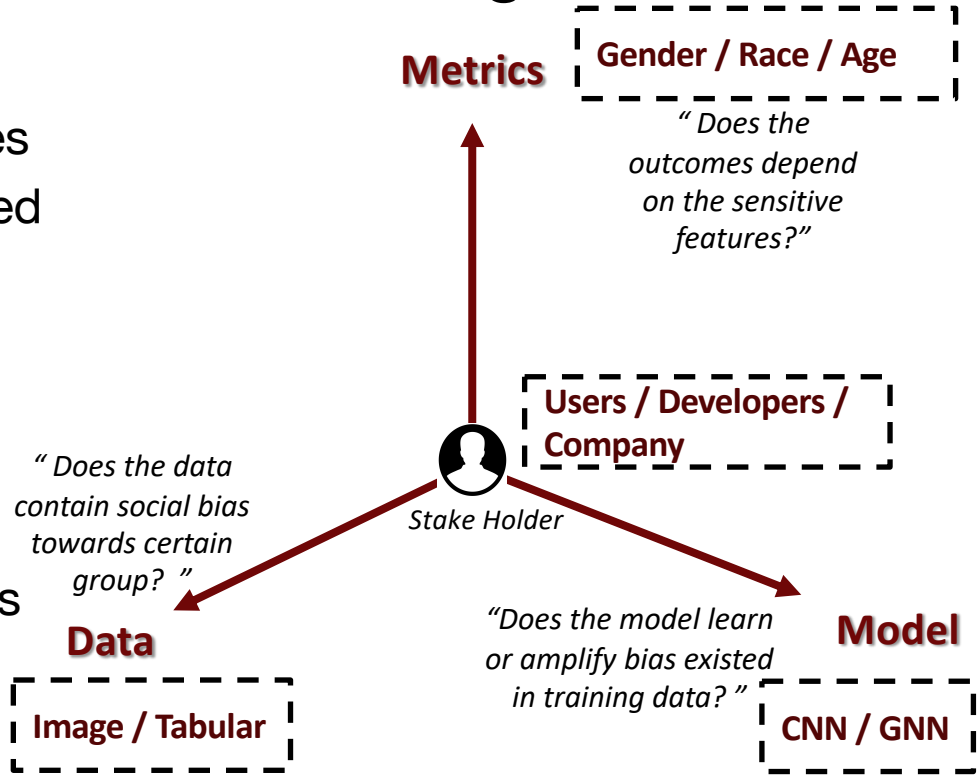
We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks. For example, it passes a simulated bar exam with a score around the top 10% of test takers; in contrast, GPT-3.5's score was around the bottom 10%. We've spent 6 months iteratively aligning GPT-4 using lessons from our adversarial testing program as well as ChatGPT, resulting in our best-ever results (though far from perfect) on factuality, steerability, and refusing to go outside of guardrails.



Fairness in Machine Learning

Goal: Develop ML/AI systems that help make decisions leading to fair outcomes

- Metrics: Evaluate outcome bias based on protected attributes
- Data: human bias leading to biased training data
- Model: ML model even amplifies bias during training

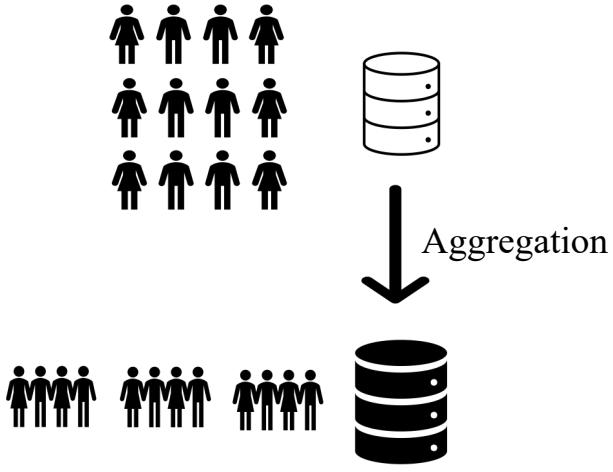


Overview

- [A] Generalized Demographic Parity for Group Fairness, ICLR'22
- [B] Learning fair graph representations via automated data augmentations, ICLR'23
- [C] Fair Graph Message Passing, under review
- [D] Topology Matters in Fair Graph Learning: a Theoretical Pilot Study, under review
- [E] Chasing Fairness under Distribution Shift: a Model Weight Perturbation Approach, NeurIPS'23

Generalized Demographic Parity [A]

- Existing group fairness metrics are either inapplicable for continuous sensitive attribute or without tractable computation.



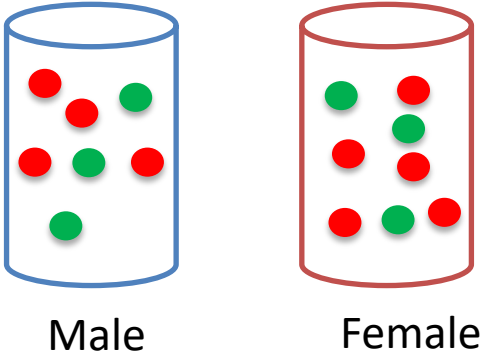
Level	Gender	Income
Person 1	Male	50k
Person 2	Female	45k

Level	Gender	Income
Region 1	45%	48.50k
Region 2	50%	46.2k

Observation: Data aggregation transforms binary sensitive attribute into **continuous** attributes

Why need a new fairness metric?

- Existing group fairness can not be applicable to continuous sensitive attribute
  - Demographic parity: the same average positive prediction rate among demographic group
  - Continuous sensitive attribute: infinite demographic groups
- Existing fairness metrics for continuous attributes are computation-intractable
  - Mutual information
  - Metric estimation methods either rely on tractable bounds or neural network approximation
  - Therefore insufficiently trustful for algorithm ranking



MINE[1]: NN replaces optimization

$$I(X;Z) \geq I_\Theta(X,Z),$$

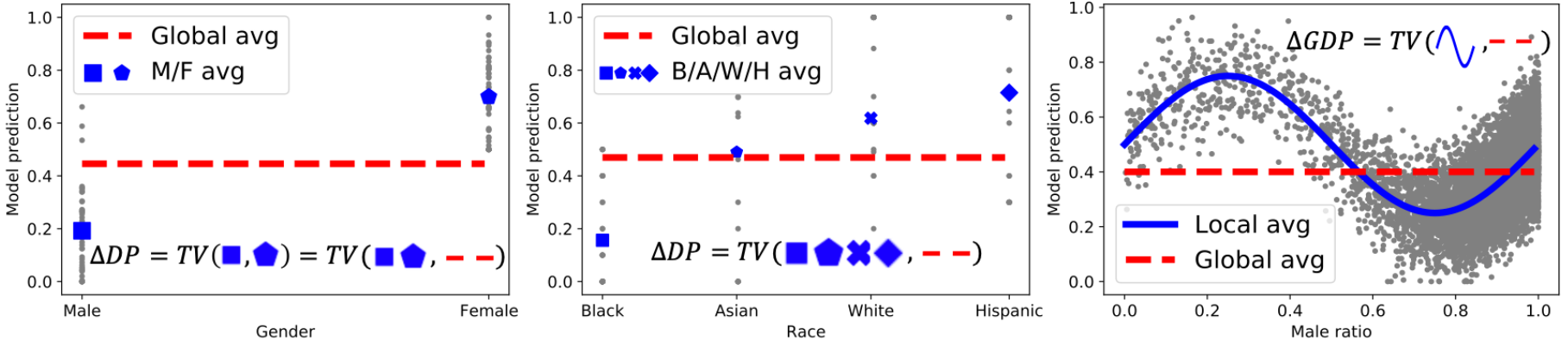
$$I_\Theta(X,Z) = \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Z}[e^{T_\theta}]).$$

Find by NN

[1] Mohamed Ishmael Belghazi, et al. "Mutual information neural estimation Mutual information neural estimation" ICML, 2018.

GDP Overview

- Demographic parity (DP) [2]: binary sensitive attribute
- Difference w.r.t. DP (DDP) [3]: categorical sensitive attribute
- Generalized DP (GDP): general version for binary/categorical/continuous sensitive attribute
  - local/global difference
  - Local average: average prediction given specific sensitive attribute



[2] Feldman, Michael, et al. "Certifying and removing disparate impact." proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015.

[3] Cho, Jaewoong, et al. "A fair classifier using kernel density estimation." Advances in Neural Information Processing Systems 33 (2020): 15088-15099.

GDP Justifications

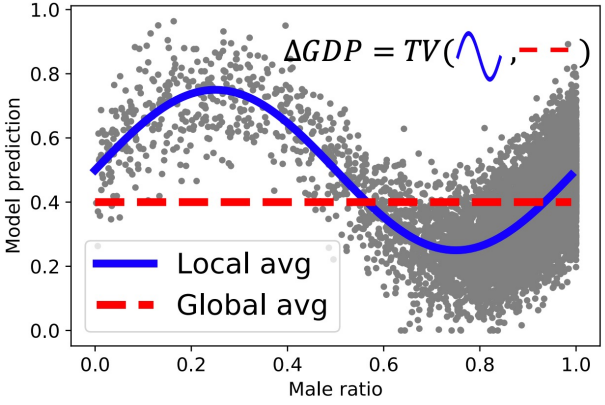
- GDP is a natural extension of DP/DDP for continuous attribute
  - GDP and DP are equivalent except for the dataset-dependent coefficient for binary attribute.
  - GDP is weighted DDP for categorical attributes.
- GDP understanding from a probabilistic view
  - Idea case: prediction $\perp$ sensitive attribute
    - Joint distribution = Product marginal distribution
  - GDP is a necessary condition for independence
    - GDP $\leq$ TV distance(joint, product margin)
- GDP regularizer v.s. adversarial debiasing
  - Adversarial debiasing leads to lower GDP

$$\mathcal{L}_{adv}\left(g^*(f(X)), S\right) \geq \Delta GDP.$$

Adversary: Predict sensitive attribute based on NN outputs

GDP Estimation

- Histogram estimation
  - Hard group: consecutive, non-overlapping intervals
  - Internal group average as local average
  - Estimation error v.s #samples: $Err_{hist} = O(N^{-\frac{2}{3}})$
- Kernel estimation
  - Soft group: closer attribute pair, higher weight
  - Normalized weighted average (Nadaraya–Watson kernel estimator)
  - Estimation error v.s #samples: $Err_{kernel} = O(N^{-\frac{4}{5}})$



$$\tilde{m}^h(s) = \frac{\sum_{n=1}^{N} \hat{y}_n K(\frac{s_n - s}{h})}{\sum_{n=1}^{N} K(\frac{s_n - s}{h})},$$

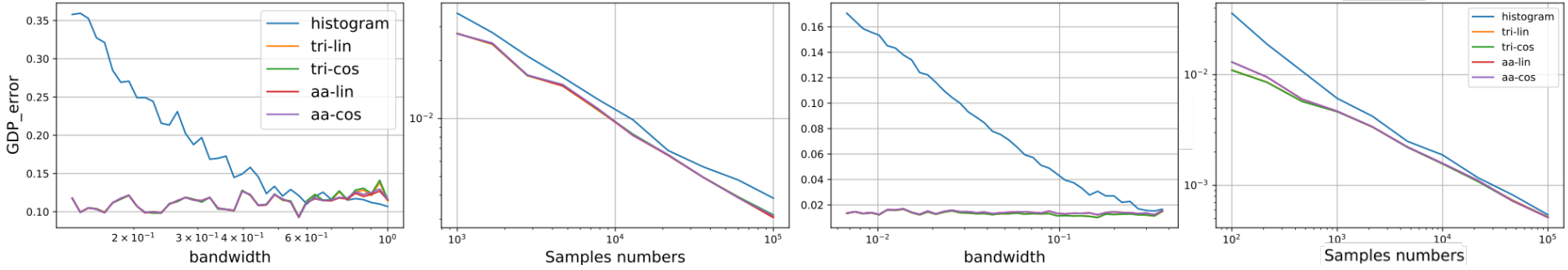
$$\tilde{m}^h_{avg} = \frac{\sum_{n=1}^{N} \hat{y}_n}{N}.$$

$$\tilde{\Delta}GDP(h) = \int_0^1 \left| \tilde{m}^h(s) - \tilde{m}^h_{avg} \right| \tilde{p}^h_S(s) \mathrm{d}s.$$

Synthetic Experiments

GDP estimation error w.r.t. bandwidth/kernel/#samples

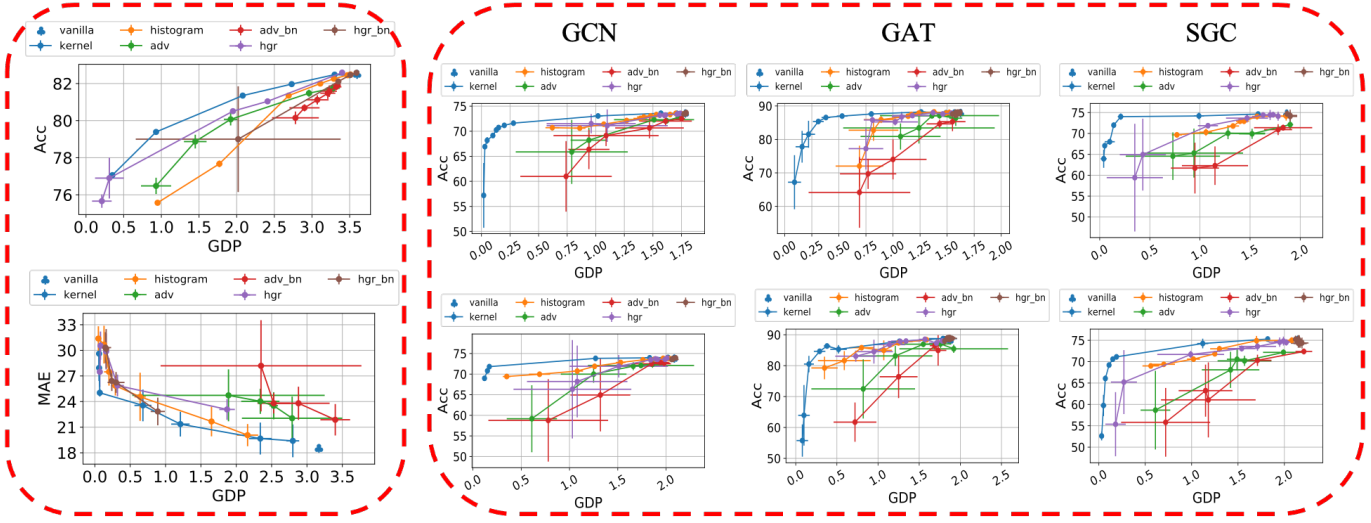
- Kernel estimation is robust over kernel/bandwidth
- Estimation error convergence rate
  - Kernel estimation > histogram estimation



Experiments on Real data

GDP regularizer achieves the best fairness-prediction trade-off performance

- Tabular/graph/dynamic graph data
- Classification/regression tasks
- Single/compositional sensitive attributes



Tabular dataset

Graph dataset

Learning fair graph representations via automated data augmentations [B]

- Data Augmentation to mitigate data bias

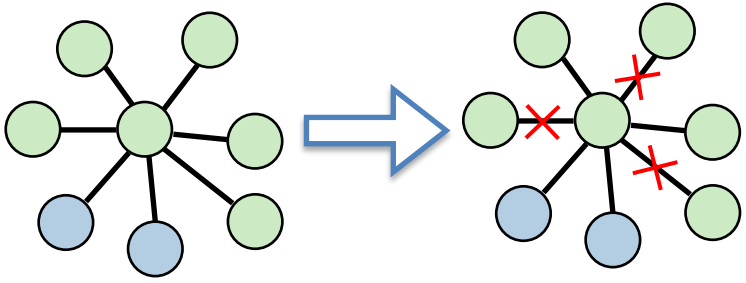
- Existing fairness-aware graph data augmentations
  - Heuristic graph properties that are beneficial to fair representation learning
  - Not optimal from data augmentation perspective.

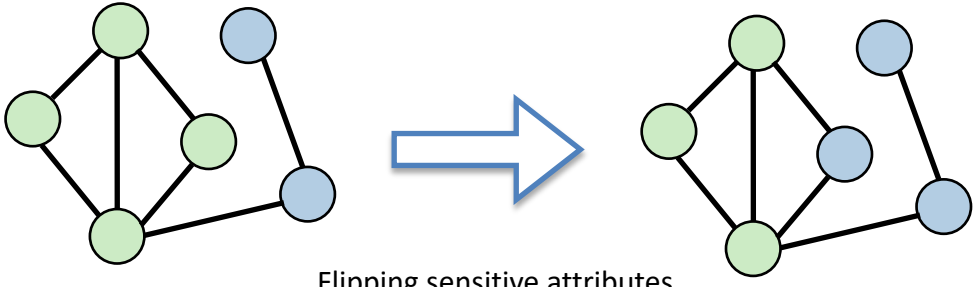
Automated Data Augmentation!

Heuristic Examples

Balanced inter/intra edges

NIFTY [4]





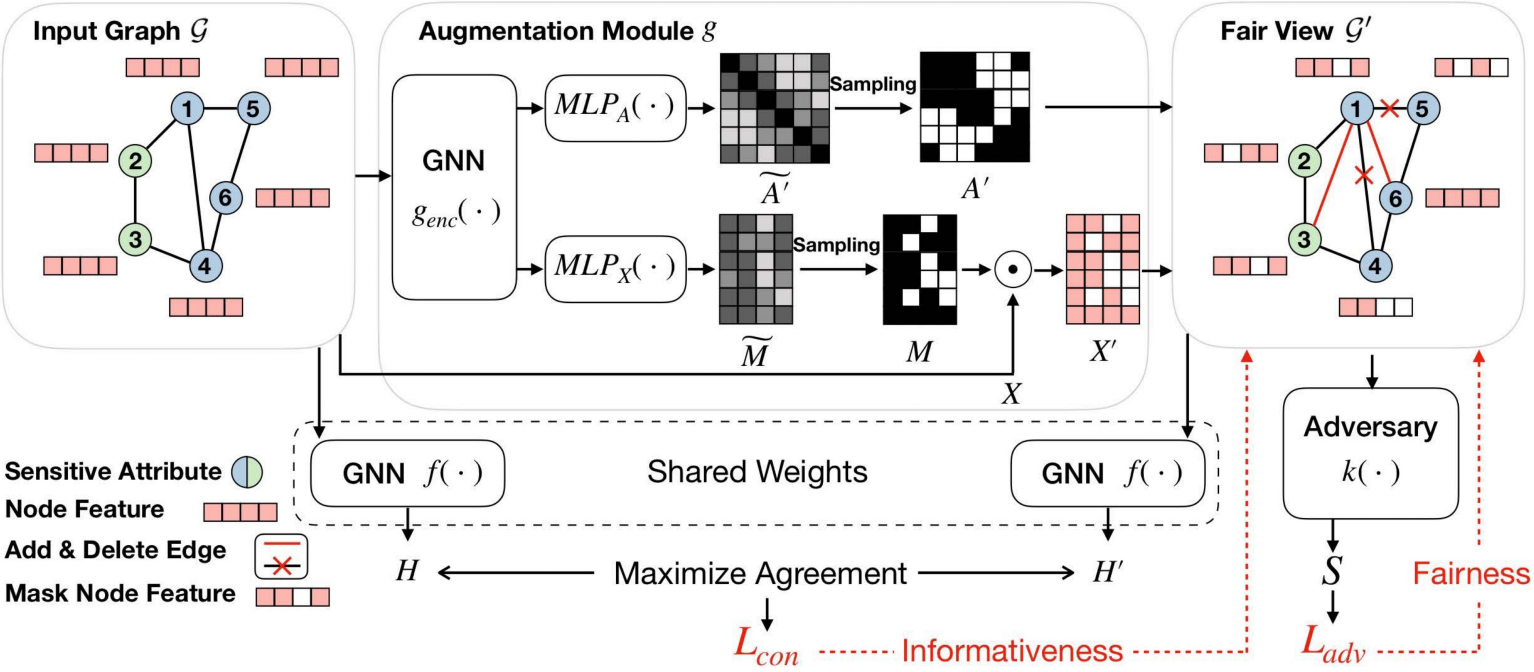
Flipping sensitive attributes

[4] Agarwal, Chirag, Himabindu Lakkaraju, and Marinka Zitnik. "Towards a unified framework for fair and stable graph representation learning." UAI 2021.

Graphair

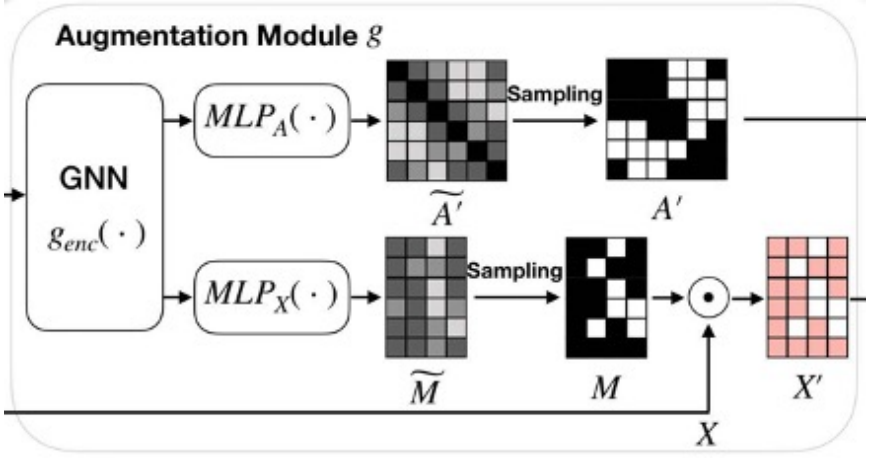
- Graphair is a novel automated graph augmentation method for fair graph representation learning
- Use two types of graph transformations:
  - Edge perturbation: removing existing edges and adding new edges
  - Node feature masking: setting some node features to zeros
- End-to-end training with multiple optimization objectives
  - Fairness: reducing the bias in the generated graphs
  - Informativeness: preserving the most informative components of the input graph in the generated graphs

Overview


Input Graph $\mathcal{G}$
Augmentation Module g
GNN
$g_{enc}(\cdot)$
$MLP_A(\cdot)$
$MLP_X(\cdot)$
Sampling
Sampling
$\widetilde{A}'$
A'
$\widetilde{M}$
M
X'
X
Fair View $\mathcal{G}'$
Sensitive Attribute
Node Feature
Add & Delete Edge
Mask Node Feature
GNN $f(\cdot)$
Shared Weights
GNN $f(\cdot)$
H
Maximize Agreement
H'
L_{con}
Informativeness
Adversary
$k(\cdot)$
S
L_{adv}
Fairness

Augmentation Module

- Graphair generate the fair view by a learnable augmentation module
  - Extract node embeddings by a GNN model
  - Predict probabilities for each graph element (i.e., node features and edge) by MLP models
  - Sample a new graph based on the predicted probabilities



Fairness with Adversarial Training

- Supervised training is impossible as there are no ground truths indicating which graph elements lead to prediction bias and should be modified
- Achieve fairness via adversarial learning
  - Use an adversary model to predict the sensitive attribute from graph generated by augmentation module
  - The adversary is optimized to maximize the prediction accuracy
  - The augmentation module is optimized to mitigate bias, so that it is difficult for the adversary model to identify sensitive attribute information

$$\min_{g}\max_{k} L_{\text{adv}} = \min_{g}\max_{k} \frac{1}{n}\sum_{i=1}^{n}\left[S_i \log \hat{S}_i + (1-S_i)\log\left(1-\hat{S}_i\right)\right]$$

Informativeness via Contrastive Training

- Only using the adversarial training may cause the augmentation module to collapse into trivial solutions
  - Ex. A complete graph with all zero node features
- Achieve informativeness by using contrastive learning
  - Maximizing the similarity between the representations of the same node

$$l(h_i, h_i') = -\log \frac{\exp\left(\text{sim}(h_i, h_i')/\tau\right)}{\sum_{j=1}^{n} \exp\left(\text{sim}(h_i, h_j')/\tau\right) + \sum_{j=1}^{n} \mathbb{1}_{[j \neq i]} \exp\left(\text{sim}(h_i, h_j)/\tau\right)},$$

$$L_{\text{con}} = \frac{1}{2n} \sum_{i=1}^{n} \left[l(h_i, h_i') + l(h_i', h_i) \right].$$

Experiments

- RQ1: Does Graphair outperform state-of-the-art methods in terms of fairness while maintaining comparable accuracy?
- RQ2: Does Graphair has a better trade-off between accuracy and fairness than state-of-the-art methods?
- Three real-world datasets
  - NBA, Pokec-z and Pokec-n

Dataset	NBA	Pokec-z	Pokec-n
# Nodes	403	67,797	66,569
# Node features	39	59	59
# Edges	16,570	882,765	729,129
# Inter-group edges	4,401	39,804	31,515
# Intra-group edges	12,169	842,961	697,614

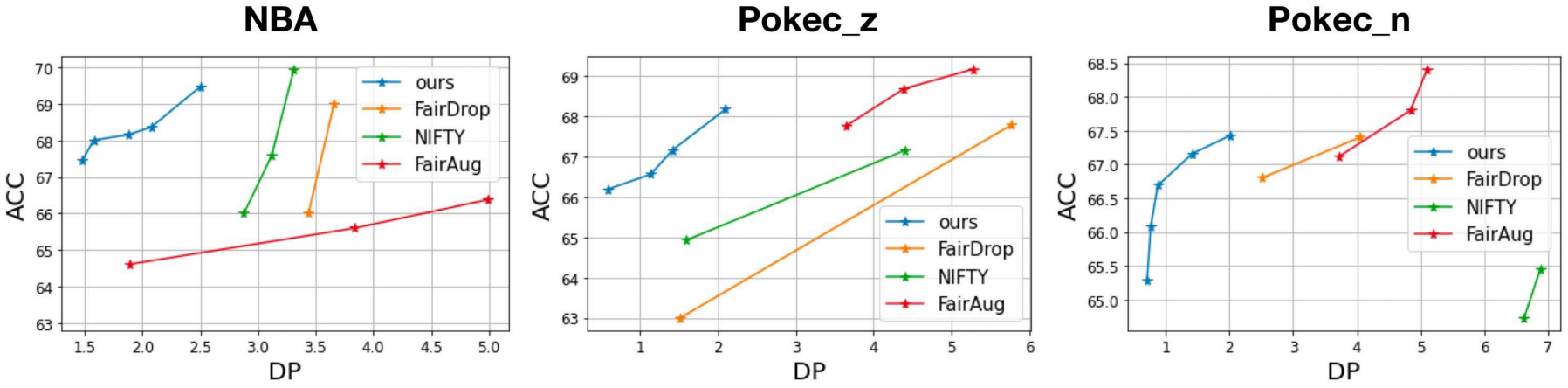
RQ1: Experiments on Fairness Performance

- Our proposed Graphair consistently achieves the best fairness performance in terms of demographic parity and equal opportunity

Method	NBA			Pokec-z			Pokec-n		
	ACC ↑	Δ_{DP} ↓	Δ_{EO} ↓	ACC ↑	Δ_{DP} ↓	Δ_{EO} ↓	ACC ↑	Δ_{DP} ↓	Δ_{EO} ↓
FairWalk	64.54 ± 2.35	3.67 ± 1.28	9.12 ± 7.06	67.07 ± 0.24	7.12 ± 0.74	8.24 ± 0.75	65.23 ± 0.78	4.45 ± 1.25	4.59 ± 0.86
FairWalk+**X**	69.74 ± 1.71	14.61 ± 4.98	12.01 ± 5.38	69.01 ± 0.38	7.59 ± 0.96	9.69 ± 0.09	67.65 ± 0.60	4.46 ± 0.38	6.11 ± 0.54
GRACE	70.14 ± 1.40	7.49 ± 3.78	7.67 ±3.78	68.25 ± 0.99	6.41 ± 0.71	7.38 ± 0.84	67.81 ± 0.41	10.77 ± 0.68	10.69 ± 0.69
GCA	**70.43 ± 1.19**	18.08 ± 4.80	20.04 ± 4.34	**69.34 ± 0.20**	6.07 ± 0.96	7.39 ± 0.82	67.07 ± 0.14	7.90 ± 1.10	8.05 ± 1.07
FairDrop	69.01 ± 1.11	3.66 ± 2.32	7.61 ± 2.21	67.78 ± 0.60	5.77 ± 1.83	5.48 ± 1.32	67.32 ± 0.61	4.05 ± 1.05	3.77 ± 1.00
NIFTY	69.93 ± 0.09	3.31 ± 1.52	4.70 ± 1.04	67.15 ± 0.43	4.40 ± 0.99	3.75 ± 1.04	65.52 ± 0.31	6.51 ± 0.51	5.14 ± 0.68
FairAug	66.38 ± 0.85	4.99 ± 1.02	6.21 ± 1.95	69.17 ± 0.18	5.28 ± 0.49	6.77 ± 0.45	**68.61 ± 0.19**	5.10 ± 0.69	5.22 ± 0.84
Graphair	69.36 ± 0.45	**2.56 ± 0.41**	**4.64 ± 0.17**	68.17 ± 0.08	**2.10 ± 0.17**	**2.76 ± 0.19**	67.43 ± 0.25	**2.02 ± 0.40**	**1.62 ± 0.47**

RQ2: Trade-off between Accuracy and Fairness

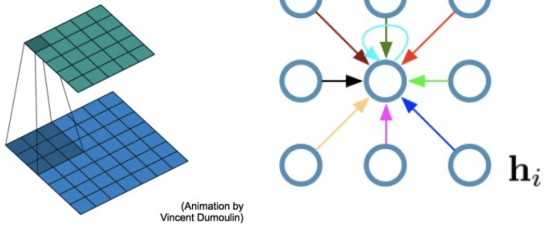
- Graphair achieves the best ACC-DP trade-off
  - The upper-left corner point represents the ideal performance, i.e., highest accuracy and lowest prediction bias.



Fairness Graph Message Passing [C]

- What is Graph Message Passing?

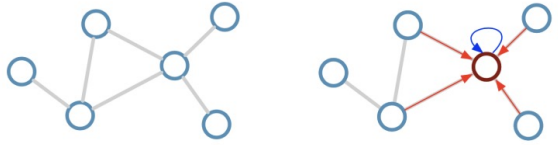
Single CNN layer with 3x3 filter:



CNN

Consider this undirected graph:

Calculate update for node in red:



Update rule: $\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$

Scalability: subsample messages [Hamilton et al., NIPS 2017]

GNN

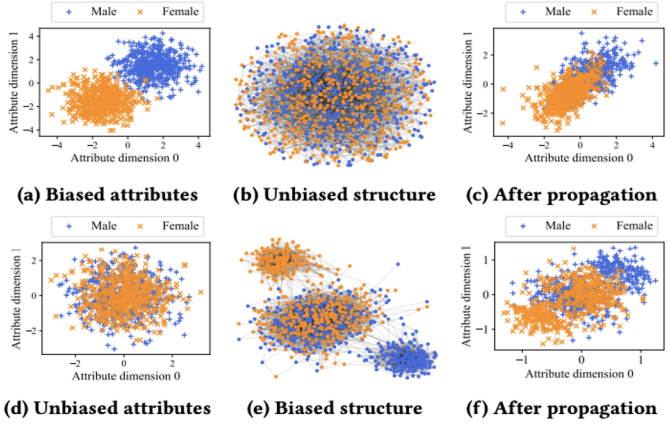
GNN is a general version of CNN!

Empirical Observations

- Aggregations in GNNs amplify bias compared with MLP.
  - GNNs > MLP in terms of prediction bias [5]
  - Representation bias after propagation even with unbias input [6]

Table 2: Results of models w/ and w/o utilizing graph.

Dataset	Metrics	MLP	MLP-e	GCN	GAT
Pokec-z	ACC (%)	65.3 ±0.5	68.6 ±0.3	70.2 ±0.1	70.4 ±0.1
	AUC (%)	71.3 ±0.3	74.8 ±0.3	77.2 ±0.1	76.7 ±0.1
	Δ_{SP} (%)	3.8 ±1.3	6.9 ±1.0	9.9 ±1.1	9.1 ±0.9
	Δ_{EO} (%)	2.2 ±0.7	4.0 ±1.5	9.1 ±0.6	8.4 ±0.6
Pokec-n	ACC (%)	63.1 ±0.4	66.3 ±0.6	70.5 ±0.2	70.3 ±0.1
	AUC (%)	68.2 ±0.3	72.4 ±0.6	75.1 ±0.2	75.1 ±0.2
	Δ_{SP} (%)	3.3 ±0.6	8.7 ±1.0	9.6 ±0.9	9.4 ±0.7
	Δ_{EO} (%)	7.1 ±0.9	9.9 ±0.6	12.8 ±1.3	12.0 ±1.5



[5] Dai, Enyan, et al. ” Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information.“ WSDM, 2021.
[6] Dong, Yushun, et al. “Edits: Modeling and mitigating data bias for graph neural networks.” WWW, 2022.

Challenges

- Message passing is problematic from fairness aspect
  - Try fair message passing!

- Fair message passing scheme is challenging
  - Effective and fair with cross entropy loss and without data pre-processing
  - Compatible with backward progagation training
  - Grounded theoretical support (hand-craft architecture design is usually intuitive)

A Unified Optimization Framework

GNNs are graph signal denoising [7]

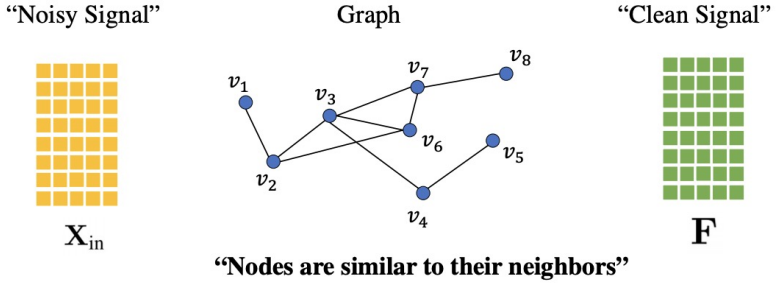
$$\arg\min_{\mathbf{F}} \mathcal{L}(\mathbf{F}) := \|\mathbf{F} - \mathbf{X}_{\text{in}}\|_F^2 + \mathcal{R}(\mathbf{F}, \tilde{\mathbf{L}})$$

Close to the input

Smoothness prior

$$\mathcal{R}(\mathbf{F}, \tilde{\mathbf{L}}) = \lambda \operatorname{tr}(\mathbf{F}^\top \tilde{\mathbf{L}} \mathbf{F}) = \lambda \sum_{(v_i, v_j) \in \mathcal{E}} \| \frac{\mathbf{F}_i}{\sqrt{d_i + 1}} - \frac{\mathbf{F}_j}{\sqrt{d_j + 1}} \|_2^2$$

Define Prior ⇨ Optimization Solver ⇨ Message Passing



- GCN $\quad \mathbf{X}_{\text{out}} = \tilde{\mathbf{A}}\mathbf{X}_{\text{in}}$
- PPNP $\quad \mathbf{X}_{\text{out}} = \alpha\big(\mathbf{I} - (1-\alpha)\tilde{\mathbf{A}}\big)^{-1}\mathbf{X}_{\text{in}}$
- APPNP/GCNII $\quad \mathbf{X}^{(k+1)} = (1-\alpha)\tilde{\mathbf{A}}\mathbf{X}^{(k)} + \alpha\mathbf{X}_{\text{in}}$

[7] Ma, Yao, et al. "A unified view on graph neural networks as graph signal denoising." CIKM 2021

Fair Message Passing

Define Prior ⇒ Optimization Solver ⇒ Message Passing

- Objective design

$$\min_{\mathbf{F}} \underbrace{\frac{\lambda_s}{2} tr(\mathbf{F}^T \tilde{\mathbf{L}} \mathbf{F}) + \frac{1}{2} ||\mathbf{F} - \mathbf{X}_{trans}||_F^2}_{h_s(\mathbf{F})} + \underbrace{\lambda_f ||\mathbf{\Delta}_s SF(\mathbf{F})||_1}_{h_f\left(\mathbf{\Delta}_s SF(\mathbf{F})\right)}.$$

→ Fairness prior

- Optimization solver
  - Avoid L1 norm objective via Fenchel conjugate
  - Proximal Alternating Predictor-Corrector Solver [8]

$$\min_{\mathbf{F}} \max_{\mathbf{u}} h_s(\mathbf{F}) + \langle \mathbf{p}, \mathbf{u} \rangle - h_f^*(\mathbf{u})$$

- Fair Messaage passing

$$\begin{cases} \mathbf{X}_{agg}^{k+1} = \gamma \mathbf{X}_{trans} + (1-\gamma)\tilde{\mathbf{A}}\mathbf{F}^k, & \text{Step ❶} \\ \bar{\mathbf{F}}^{k+1} = \mathbf{X}_{agg}^{k+1} - \gamma \frac{\partial \langle \mathbf{p}, \mathbf{u}^k \rangle}{\partial \mathbf{F}}\Big|_{\mathbf{F}^k}, & \text{Step ❷} \\ \bar{\mathbf{u}}^{k+1} = \mathbf{u}^k + \beta \mathbf{\Delta_s} SF(\bar{\mathbf{F}}^{k+1}), & \text{Step ❸} \\ \mathbf{u}^{k+1} = \min\left(|\bar{\mathbf{u}}^{k+1}|, \lambda_f\right) \cdot sign(\bar{\mathbf{u}}^{k+1}), & \text{Step ❹} \\ \mathbf{F}^{k+1} = \mathbf{X}_{agg}^{k+1} - \gamma \frac{\partial \langle \mathbf{p}, \mathbf{u}^{k+1} \rangle}{\partial \mathbf{F}}\Big|_{\mathbf{F}^k}. & \text{Step ❺} \end{cases}$$

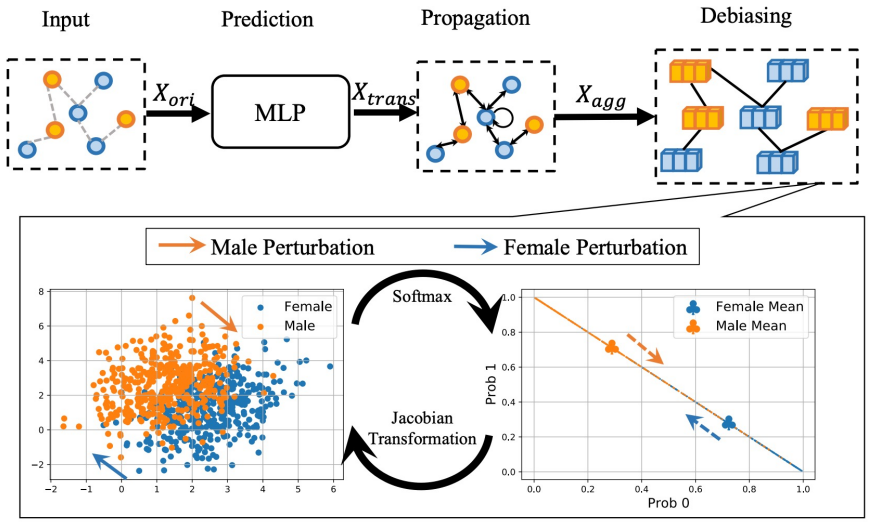
→ Aggregation with skip connection

→ Learn and reshape perturbation vector u

[8] Ignace Loris, et al. On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Problems, 27(12):125007, 2011.

Fair Message Passing

- FMP Interpretation
  - Three stages in FMP
  - Four steps in Debiasing
- Efficiency
  - Negligible additional computation
- White-box sensitive attribute usuage
  - Explicit usuage in FMP
  - Implicit encoding in parameters for fair training



Experiments

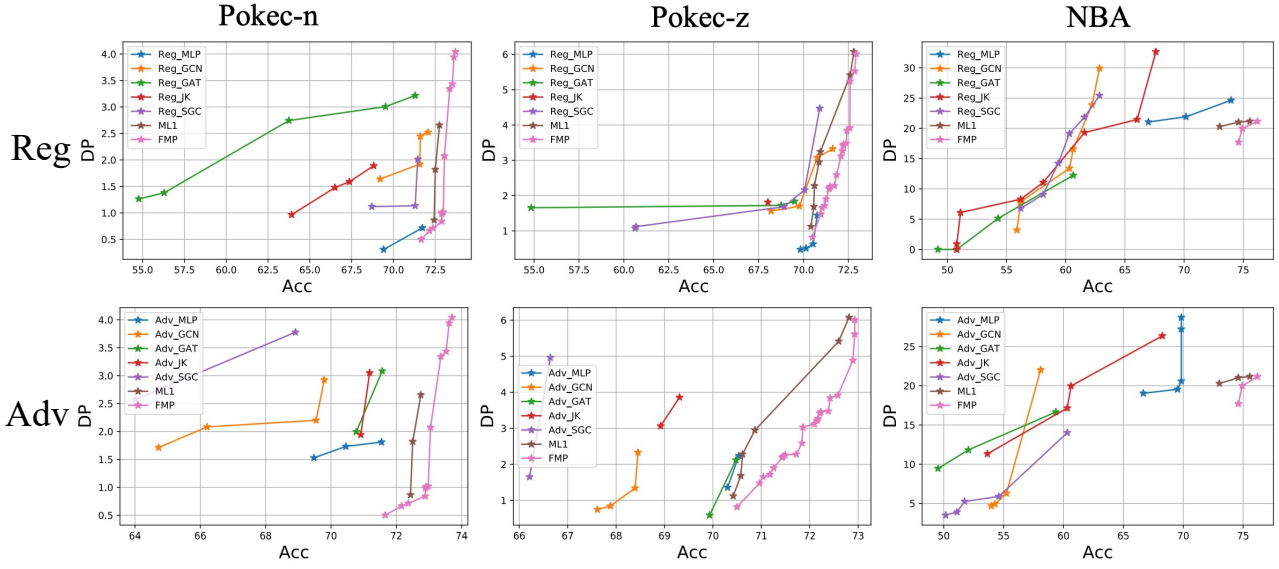
- Various GNNs architectures with cross-entropy loss
  - Lowest DP and EO with comparable accuracy

Table 1: Comparative Results with Baselines on Node Classification.

Models	Pokec-z			Pokec-n			NBA		
	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
MLP	70.48 ± 0.77	1.61 ± 1.29	2.22 ± 1.01	72.48 ± 0.26	1.53 ± 0.89	3.39 ± 2.37	65.56 ± 1.62	22.37 ± 1.87	18.00 ± 3.52
GAT	69.76 ± 1.30	2.39 ± 0.62	2.91 ± 0.97	71.00 ± 0.48	3.71 ± 2.15	7.50 ± 2.88	57.78 ± 10.65	20.12 ± 16.18	13.00 ± 13.37
GCN	**71.78** ± 0.37	3.25 ± 2.35	2.36 ± 2.09	**73.09** ± 0.28	3.48 ± 0.47	5.16 ± 1.38	61.90 ± 1.00	23.70 ± 2.74	17.50 ± 2.63
SGC	71.24 ± 0.46	4.81 ± 0.30	4.79 ± 2.27	71.46 ± 0.41	2.22 ± 0.29	3.85 ± 1.63	63.17 ± 0.63	22.56 ± 3.94	14.33± 2.16
APPNP	66.91 ± 1.46	3.90 ± 0.69	5.71 ± 1.29	69.80 ± 0.89	1.98 ± 1.30	4.01 ± 2.36	63.80 ± 1.19	26.51 ± 3.33	20.00 ± 4.56
JKNet	66.89 ± 3.79	1.28 ±0.96	1.79 ± 0.82	63.59 ± 6.36	1.91 ± 2.14	**0.70** ± 0.92	67.94 ± 2.73	27.80 ± 8.41	20.33 ± 7.52
ML1	70.42 ± 0.40	2.35 ± 0.83	2.00 ± 0.50	72.36 ± 0.26	1.47 ± 1.12	3.03 ± 1.77	72.70 ± 1.19	26.46 ± 4.93	25.50 ± 8.38
FMP	70.50 ± 0.50	**0.81** ± 0.40	**1.73** ± 1.03	72.16 ± 0.33	**0.66** ± 0.40	1.47 ± 0.87	**73.33** ± 1.85	**18.92** ± 2.28	**13.33** ± 5.89

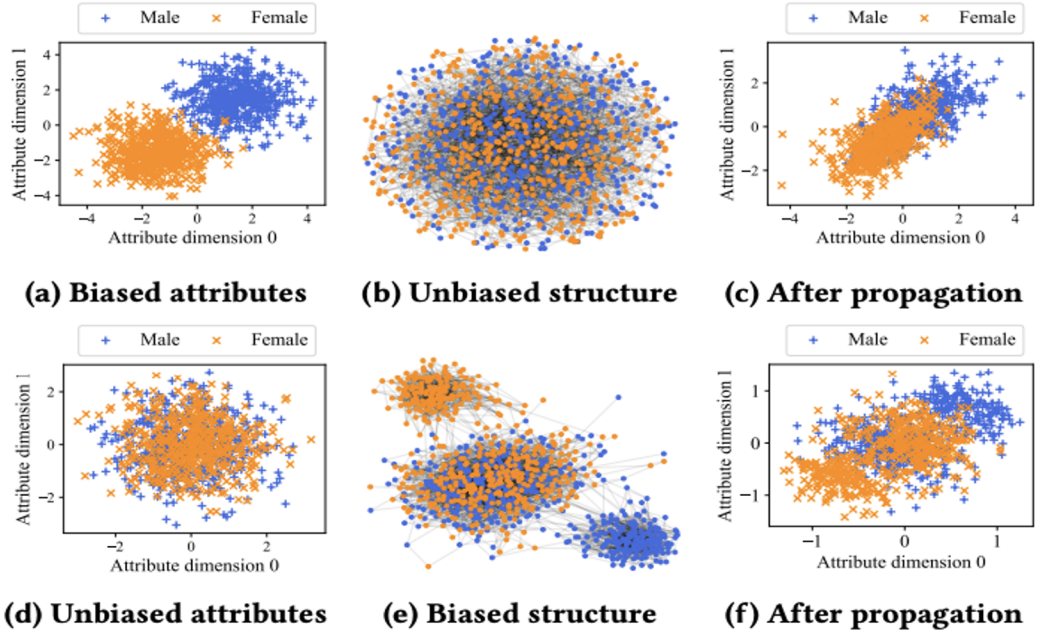
Experiments

- Compared with fair training
  - adding regularization and adversarial debiasing as baselines
  - FMP still achieves better tradeoff performance



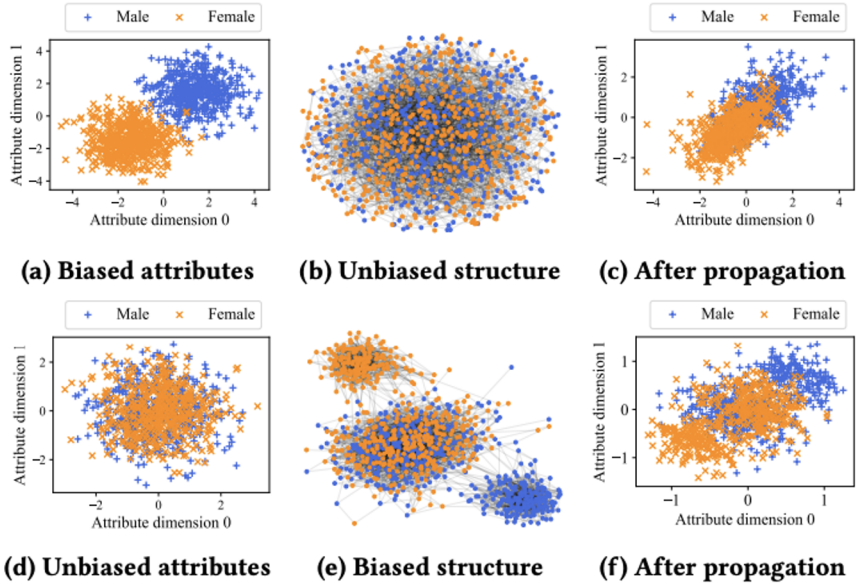
Topology Matters in Fair Graph Learning [D]

- Topology serves as an additional bias source in graph learning.



Understanding Graph Data Bias

- Understanding the bias in graph neural networks (GNNs)
  - GNNs demonstrate empirically higher prediction bias than peer multilayer perception (MLP) [4] but without theoretical understanding.
  - Bias representation after propagation for bias structure even with unbiased attributes
  - When and Why does aggregation enhance the bias?



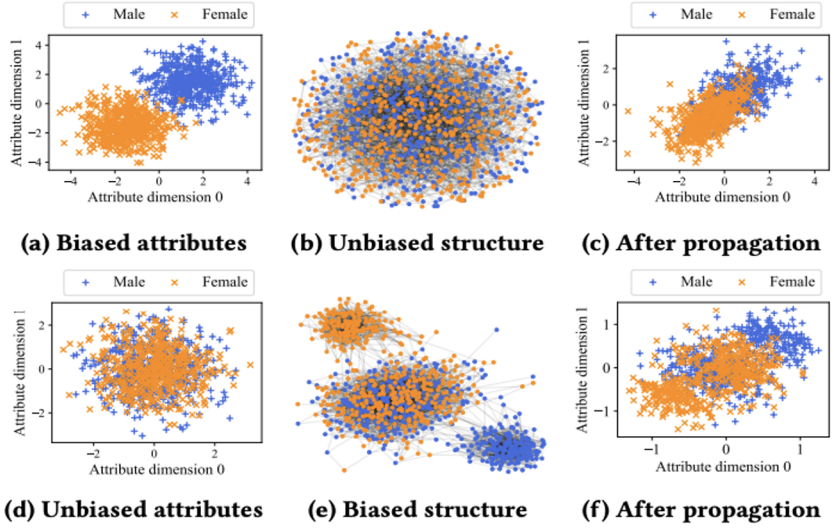
[5] Dai, Enyan, et al. ” Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information.“ WSDM, 2021.
[6] Dong, Yushun, et al. “Edits: Modeling and mitigating data bias for graph neural networks.” WWW, 2022.

Empirical Observations

- Aggregations in GNNs amplify bias compared with MLP.
  - GNNs > MLP in terms of prediction bias [5]
  - Representation bias after propagation even with unbias input [6]

Table 2: Results of models w/ and w/o utilizing graph.

Dataset	Metrics	MLP	MLP-e	GCN	GAT
Pokec-z	ACC (%)	65.3 ±0.5	68.6 ±0.3	70.2 ±0.1	70.4 ±0.1
	AUC (%)	71.3 ±0.3	74.8 ±0.3	77.2 ±0.1	76.7 ±0.1
	Δ_{SP} (%)	3.8 ±1.3	6.9 ±1.0	9.9 ±1.1	9.1 ±0.9
	Δ_{EO} (%)	2.2 ±0.7	4.0 ±1.5	9.1 ±0.6	8.4 ±0.6
Pokec-n	ACC (%)	63.1 ±0.4	66.3 ±0.6	70.5 ±0.2	70.3 ±0.1
	AUC (%)	68.2 ±0.3	72.4 ±0.6	75.1 ±0.2	75.1 ±0.2
	Δ_{SP} (%)	3.3 ±0.6	8.7 ±1.0	9.6 ±0.9	9.4 ±0.7
	Δ_{EO} (%)	7.1 ±0.9	9.9 ±0.6	12.8 ±1.3	12.0 ±1.5

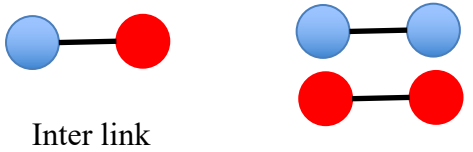


[5] Dai, Enyan, et al. " Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information." WSDM, 2021.
[6] Dong, Yushun, et al. "Edits: Modeling and mitigating data bias for graph neural networks." WWW, 2022.

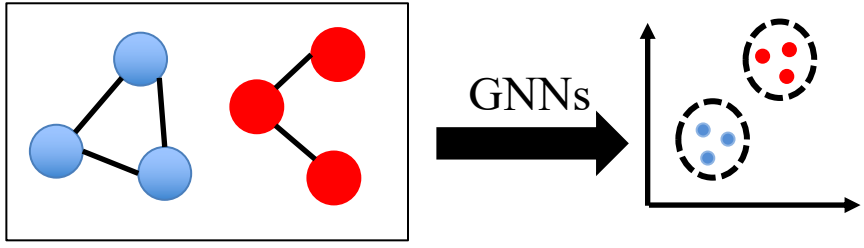
Why does Aggregation suffer?

Intuition

- Graph topology with high sensitive homophily coefficient
  - Definition: #sensitive homo links / # links
  - E.g., 95.30% for Pokec-n dataset
  - Higher than label homophily coefficient
- Graph concentration (over-smoothing)
  - More similar representation within the demographic group
  - Conditionally happens: no bias for fully over-smoothing



$$\text{Sensitive Homophily} = \frac{\#\ Intra\ links}{\#\ all\ links}$$



How can we theoretically understanding such GNNs behavior?

A Pilot Theoretical Study

Goal: find a sufficient condition of bias enhancement after aggregation

- Synthetic graph data: contexture stochastic block model
  - Topology with intra/inter-connect probability
  - Features with Gaussian Mixture Model
- GCN-like Aggregation
- Bias difference before/after aggregation

When does bias enhancement happen?

- large sensitive homophily coefficient & node number & connection density
- Balanced demographic size

Topology matters in fair graph learning!

Fair Graph Rewiring

Preprocessing: rewire graph topology to achieve graph fairness

- Large label homophily coefficient
- Low sensitive homophily coefficient
- Low topology perturbation

$$L(\hat{A}|s,y,A) = \frac{\|H(ss^T)\odot\hat{A}\|_1}{\|\hat{A}\|_1} - \alpha\frac{\|H(yy^T)\odot\hat{A}\|_1}{\|\hat{A}\|_1} + \beta\|\hat{A}-A\|_1$$

Sensitive Homophily

Label Homophily

Topology Perturbation

Synthetic Experiments

Observations:

1) Bias enhancement happens conditionally
2) The bias enhancement tendency is consistent with our theory

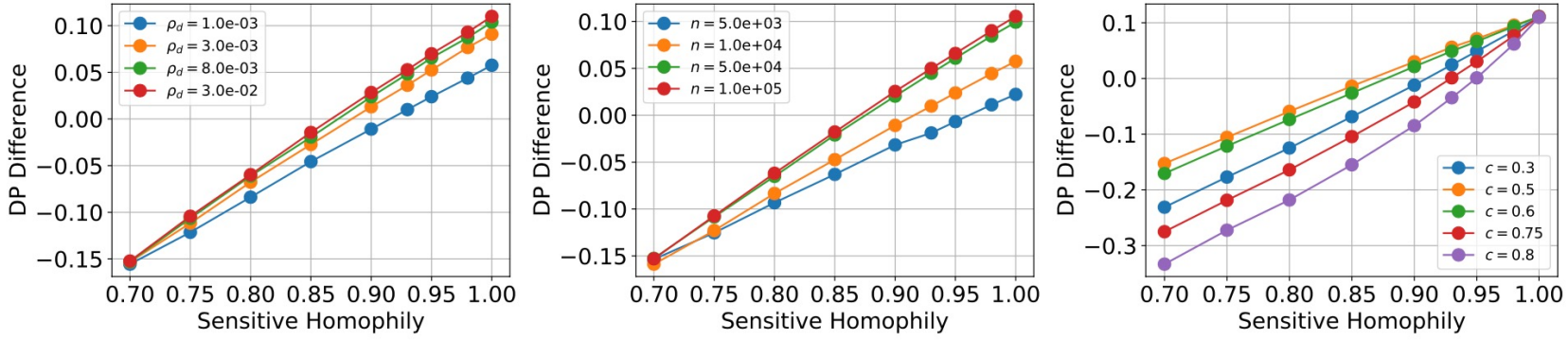


Figure 2: The difference of demographic parity for message passing. **Left:** DP difference for different graph connection density ρ_d with sensitive attribute ratio $c = 0.5$ and number of nodes $n = 10^4$; **Middle:** DP difference for different number of nodes n with sensitive attribute ratio $c = 0.5$ and graph connection density $\rho_d = 10^{-3}$; **Right:** DP difference for different sensitive attribute ratio c with graph connection density $\rho_d = 10^{-3}$ and number of nodes $n = 10^4$;

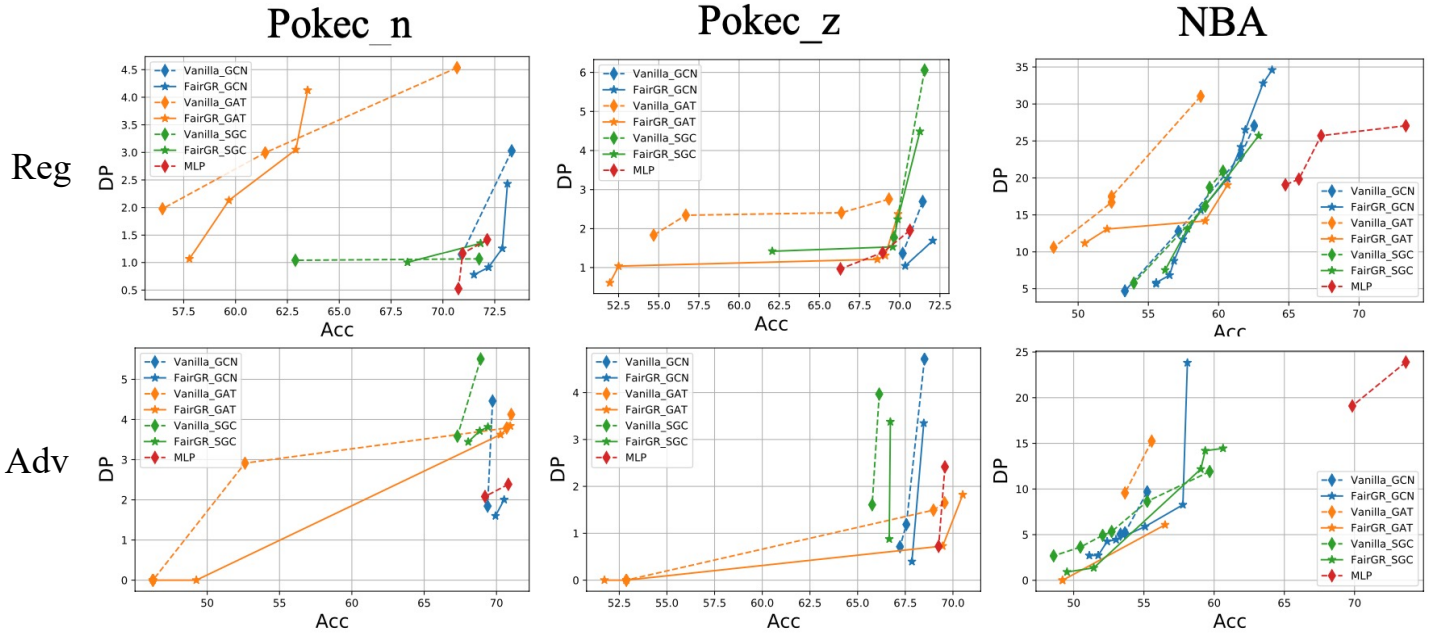
Experiments on Real Data

- Graph rewiring facilitates fairness in vanilla GNNs
  - Lower demographic parity and equal opportunities for various GNNs architectures

Models	Pokec-z			Pokec-n			NBA		
	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
MLP	70.48 ± 0.77	1.61 ± 1.29	2.22 ± 1.01	72.48 ± 0.26	1.53 ± 0.89	3.39 ± 2.37	65.56 ± 1.62	22.37 ± 1.87	18.00 ± 3.52
GAT	69.76 ± 1.30	2.39 ± 0.62	2.91 ± 0.97	71.00 ± 0.48	3.71 ± 2.15	7.50 ± 2.88	57.78 ± 10.65	20.12 ± 16.18	13.00 ± 13.37
GAT-GR	56.75 ± 6.32	**1.04** ± 0.80	**1.14** ± 1.02	61.27 ± 9.34	**0.54** ± 0.51	**2.27** ± 1.55	53.65 ± 10.31	**4.16** ± 5.13	**3.67** ± 3.23
GCN	71.78 ± 0.37	3.25 ± 2.35	2.36 ± 2.09	73.09 ± 0.28	3.48 ± 0.47	5.16 ± 1.38	61.90 ± 1.00	23.70 ± 2.74	17.50 ± 2.63
GCN-GR	71.68 ± 0.58	1.94 ± 1.59	**1.27** ± 0.71	72.68 ± 0.44	**0.47** ± 0.39	**0.82** ± 0.78	61.59 ± 1.85	**20.24** ± 4.41	**9.50** ± 2.77
SGC	71.24 ± 0.46	4.81 ± 0.30	4.79 ± 2.27	71.46 ± 0.41	2.22 ± 0.29	3.85 ± 1.63	63.17 ± 0.63	22.56 ± 3.94	14.33± 2.16
SGC-GR	70.95 ± 0.91	3.32 ± 1.31	3.20 ± 1.90	71.91 ± 0.52	**0.71** ± 0.65	**2.39** ± 0.69	62.54 ± 1.62	**18.56** ± 2.81	**2.50** ± 1.66

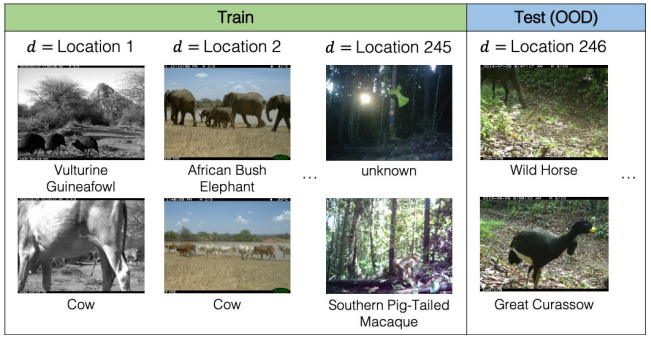
Experiments on Real Data

- Graph rewiring is orthogonal with other fair learning methods
  - E.g., adding regularization, adversarial debiasing, and Fair Mixup
  - Better fairness-prediction tradeoff performance with GR

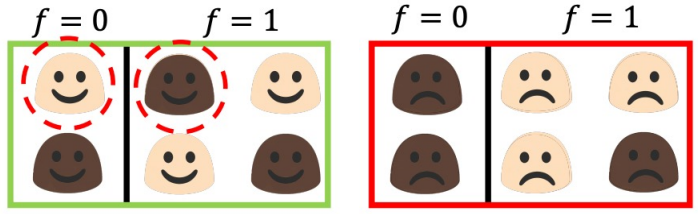


Chasing Fairness under Distribution Shift [E]

- When do distribution shifts happen [9]?
  - Different locations/hospitals;
  - Different experiments;
  - Different time periods;
  - Different devices.



- Robust Fairness under Distribution Shift
  - Many metrics (acc, fairness) under distribution shifts
  - Fairness metric is more vulnerable under distribution shifts [10]
    - Statistical rate (SR) v.s. sensitive attribute perturbation



[9] Koh, Pang Wei, et al. "Wilds: A benchmark of in-the-wild distribution shifts." *International Conference on Machine Learning*. PMLR, 2021.
[10] Celis, L. Elisa, et al. "Fair classification with adversarial perturbations." Advances in Neural Information Processing Systems. (2021).

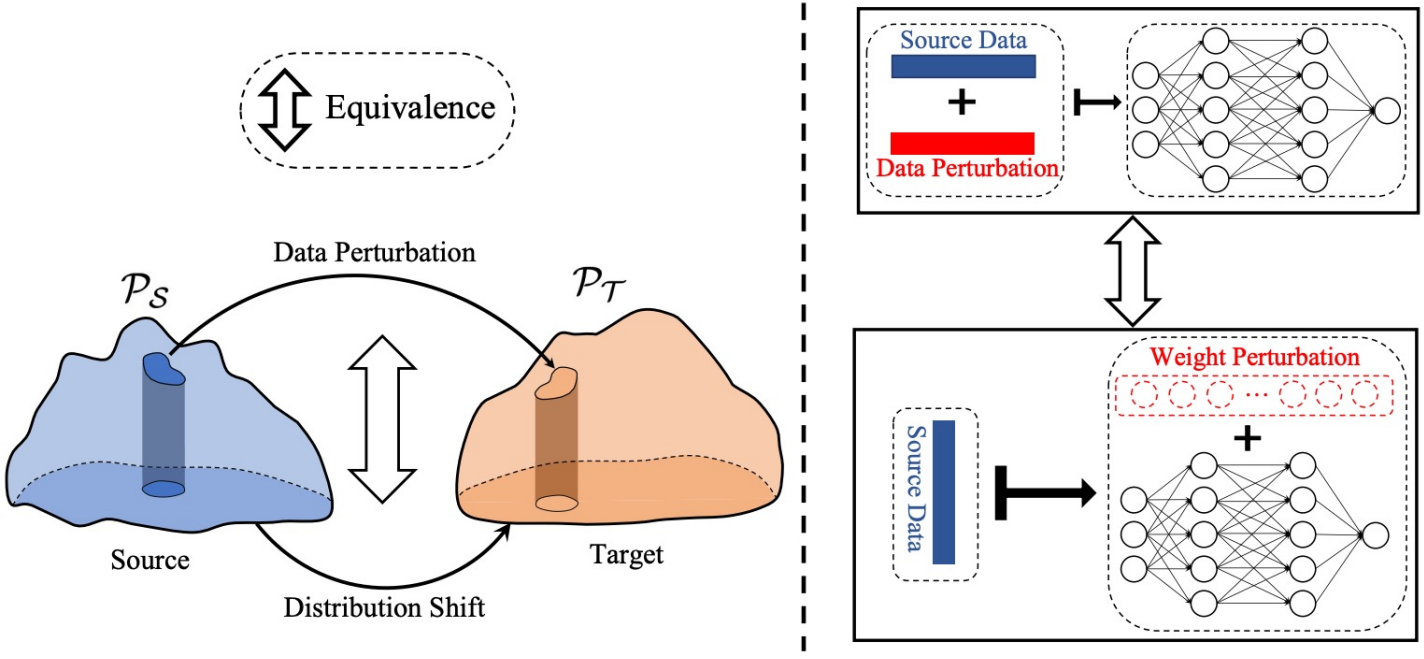
Challenges

- The distribution shifts are unknown during training
  - Can not access feature distributions in the test dataset

- Fairness is vulnerable under distribution shifts
  - Many methods are originally designed for performance
  - What can we do from model perspective to tackle data distribution shifts problem?

Rethinking Distribution Shifts

Distribution shifts ⟷ Data Perturbation ⟷ Model Weight Perturbation



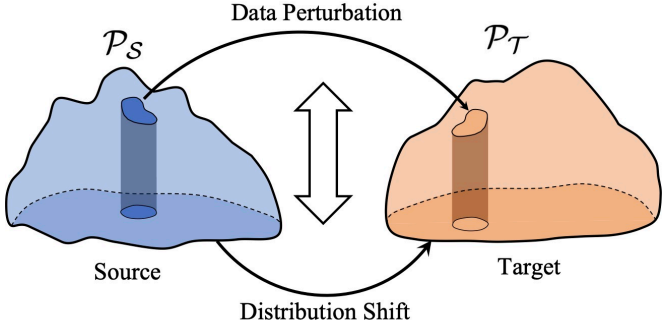
Fairness under Distribution Shifts

- Distribution shifts are data perturbation
  - Holds for any loss function and model architectures

Corollary 2.2. *Given source and target datasets with probability distribution $\mathcal{P}_S$ and $\mathcal{P}_T$, there exists data perturbation δ so that the training loss of any neural network $f_\theta(\cdot)$ for target distribution equals that for source distribution with data perturbation δ, i.e.,*

$$\mathbb{E}_{(X,Y)\sim\mathcal{P}_T}[l(f_\theta(X),Y)] = \mathbb{E}_{\delta_X,\delta_Y}\mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_\theta(X+\delta_X),Y+\delta_Y)]. \quad (3)$$

Data Perturbation



Fairness under Distribution Shifts

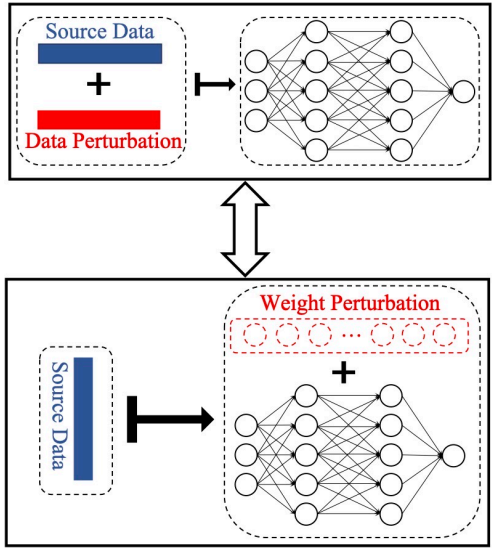
- Data perturbation equals model weight perturbation
  - Holds for any loss and model architectures

Theorem 2.3. *Considering the source dataset with distribution $\mathcal{P}_S$, suppose the source dataset is perturbed with data perturbation δ, and the neural network is given by $f_\theta(\cdot)$, there exists model weight perturbation $\Delta\theta$ so that the training loss on perturbed source dataset is the same with that for model weight perturbation $\Delta\theta$ on source distribution:*

$$\mathbb{E}_{\delta_X,\delta_Y}\mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_\theta(X+\delta_X),Y+\delta_Y)] = \mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_{\theta+\Delta\theta}(X),Y)]. \quad (4)$$

Data Perturbation

Model Weight Perturbation



Fairness under Distribution Shifts

- Fairness under distribution shifts
  - What are the conditions guarantee such robust fairness?

- Take demographic parity (DP) as an example
  - Low DP at source dataset
  - Low average prediction gap between source/target dataset at the same sensitive group

$$DP_{\mathcal{T}} \overset{(a)}{\leq} DP_{\mathcal{S}} + \Big| |\mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]| - |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]| \Big|$$

$$\overset{(b)}{\leq} DP_{\mathcal{S}} + |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})]| + |\mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]|$$

Δ_0 Δ_1

Loss function-agnostic

Fairness under Distribution Shifts

- How can we achieve low prediction gap for each demographic group?
  - Model weight perturbation: bi-level optimization problem
    - The worst case within the model weight perturbation ball for each sensitive attribute group
  - Can be accelerated with two forward-backward propagation

$$
\begin{aligned}
\Delta_0 &\leq \max_{\|\epsilon_0\|_p \leq \rho} \left| \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] \right| \\
&\approx \max_{\|\epsilon_0\|_p \leq \rho} \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] \\
&\triangleq \mathcal{L}_{RFR,\mathcal{S}_0},
\end{aligned}
$$

Model Weight Perturbation

Fairness under Distribution Shifts

- Robust Fairness Regularization (RFR)

$$\mathcal{L}_{all} = \mathcal{L}_{CLF} + \lambda \cdot (\mathcal{L}_{DP} + \mathcal{L}_{RFR}),$$

Classification loss

Low DP on source

Low prediction gap

Experiments

Distribution shifts type

- Synthetic distribution shifts
  - Use different sample selection prior distribution across training and test dataset [11]
  - Dataset: Adult; ACS-I; ACS-E
- Real-world distribution shifts
  - New Adult dataset [12]
  - Distribution shift across states
  - Distribution shift across times

Evaluation metrics: Accuracy & DP tradeoff

Baselines:

- Regularization & Adversarial debiasing
- Fair consistency regularization (FCR)

[11] Ashkan Rezaei, et al. Robust fairness under 375 covariate shift. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 9419–9427, 2021.
[12] Frances Ding, et al. Retiring adult: New datasets for fair machine learning. NeurIPS, 2021.

Experiments

Synthetic distribution shifts

- Low prediction bias at low-intensity distribution shifts
- Comparable at high-intensity distribution shifts

Table 1: Performance Comparison with Baselines on Synthetic Dataset. (α, β) control distribution shift intensity, and $(0, 1)$ represents no distribution shift. The best/second-best results are highlighted in **boldface**/underlined, respectively.

(α, β)	Methods	Adult			ACS-I			ACS-E		
		Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
(1.0, 2.0)	MLP	82.09±0.05	15.11±0.04	14.33±0.05	77.95±0.52	3.51±0.59	3.77±0.55	80.95±0.10	1.10±0.06	1.43±0.06
	REG	80.60±0.05	3.79±0.06	3.27±0.08	77.77±0.09	2.28±0.32	2.59±0.23	80.44±0.07	0.86±0.09	1.05±0.10
	ADV	78.80±0.68	0.83±0.26	0.79±0.14	75.72±0.63	1.96±0.38	2.00±0.35	79.39±0.15	1.09±0.26	0.95±0.26
	FCR	79.06±0.09	9.98±0.06	9.47±0.07	76.99±0.47	2.94±0.34	2.95±0.28	79.74±0.11	0.97±0.21	1.00±0.22
	RFR	78.84±0.09	**0.44**±0.05	**0.12**±0.06	74.15±0.81	**1.84**±0.27	**1.60**±0.33	80.08±0.08	**0.71**±0.10	**0.06**±0.11
(1.5, 3.0)	MLP	82.05±0.05	15.16±0.09	14.33±0.09	77.85±0.25	3.73±0.53	3.70±0.56	80.42±0.10	1.14±0.07	1.10±0.07
	REG	80.64±0.08	3.74±0.11	3.23±0.10	77.87±0.18	2.25±0.28	2.37±0.27	80.21±0.13	**0.72**±0.04	0.75±0.03
	ADV	78.71±0.41	1.07±0.87	0.87±0.96	75.79±0.68	2.22±0.53	2.44±0.48	79.58±0.13	1.07±0.19	1.26±0.18
	FCR	79.05±0.12	10.01±0.07	9.51±0.06	77.06±0.68	3.39±0.33	3.10±0.36	79.59±0.26	1.17±0.24	1.08±0.23
	RFR	78.91±0.03	**0.46**±0.10	**0.16**±0.09	74.19±0.58	**1.82**±0.29	**2.17**±0.32	80.47±0.03	**0.72**±0.04	**0.71**±0.05
(3.0, 6.0)	MLP	82.07±0.05	15.23±0.14	14.45±0.15	77.89±0.45	3.35±0.36	3.47±0.41	80.30±0.04	1.17±0.04	1.13±0.04
	REG	80.62±0.07	3.72±0.05	3.21±0.04	78.19±0.12	**1.60**±0.48	**1.84**±0.44	80.36±0.09	**0.70**±0.09	0.68±0.11
	ADV	78.97±0.49	**1.28**±0.74	**1.09**±0.50	75.71±0.68	2.28±0.39	2.24±0.41	79.66±0.16	1.34±0.14	1.16±0.13
	FCR	79.03±0.13	10.00±0.05	9.50±0.05	76.71±0.39	2.97±0.34	3.28±0.31	79.89±0.22	1.06±0.14	1.14±0.18
	RFR	80.15±0.07	1.75±0.15	1.30±0.14	74.22±0.56	1.80±0.26	1.89±0.24	80.28±0.12	0.74±0.04	**0.51**±0.04

Experiments

Real distribution shifts

- Low(comparable) prediction bias under temporal(spatial) distribution shift

Table 2: Performance comparison with baselines on real temporal (the year 2016 to the year 2018) and spatial (Michigan State to California State) distribution shift. The best and second-best results are highlighted with **hold** and underline, respectively.

Real	Methods	ACS-I			ACS-E		
		Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
2016 → 2018	MLP	77.75±0.44	3.26±0.38	3.48±0.41	80.46±0.05	1.07±0.10	1.02±0.10
	REG	77.74±0.62	2.09±0.21	2.27±0.24	80.37±0.12	0.77±0.08	0.74±0.08
	ADV	75.94±0.40	2.41±0.49	2.53±0.55	79.62±0.14	1.17±0.14	1.10±0.14
	FCR	76.40±0.45	2.81±0.30	2.96±0.30	79.59±0.38	0.95±0.42	0.91±0.34
	RFR	77.49±0.32	**1.36**±0.17	**1.49**±0.17	80.36±0.05	**0.61**±0.11	**0.58**±0.10
MI → CA	MLP	75.62±0.80	5.22±0.86	3.60±0.34	79.02±0.20	0.73±0.07	0.94±0.05
	REG	75.52±0.78	2.88±0.44	2.17±0.22	75.34±1.11	**0.42**±0.09	**0.61**±0.11
	ADV	73.38±1.07	**1.04**±0.58	**0.54**±0.38	77.56±0.41	0.61±0.18	0.80±0.13
	FCR	74.28±0.35	5.06±0.62	3.67±0.51	77.96±0.22	0.44±0.14	0.67±0.38
	RFR	74.63±0.45	1.35±0.39	1.30±0.24	78.84±0.21	0.44±0.09	0.65±0.07

Experiments

Pareto tradeoff performance for synthetic distribution shifts

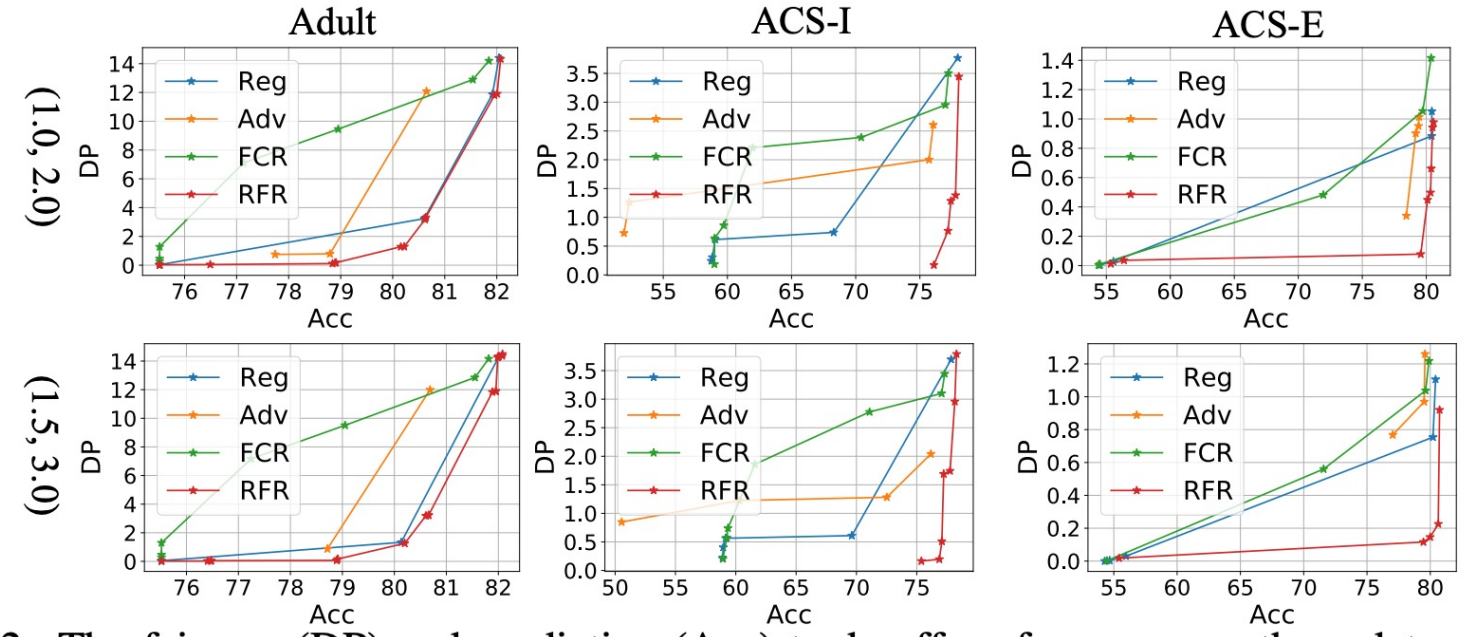


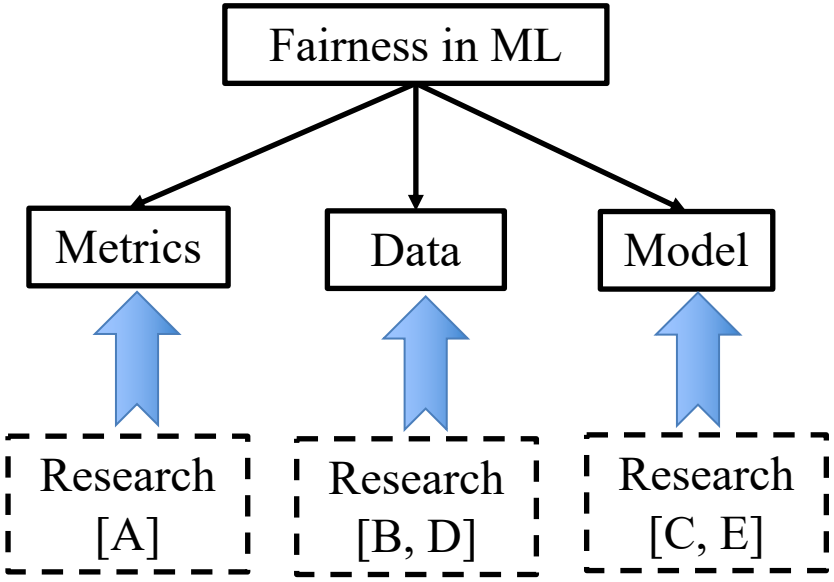
Figure 2: The fairness (DP) and prediction (Acc) trade-off performance on three datasets with different synthetic distribution shifts. The units for x- and y-axis are percentages (%).

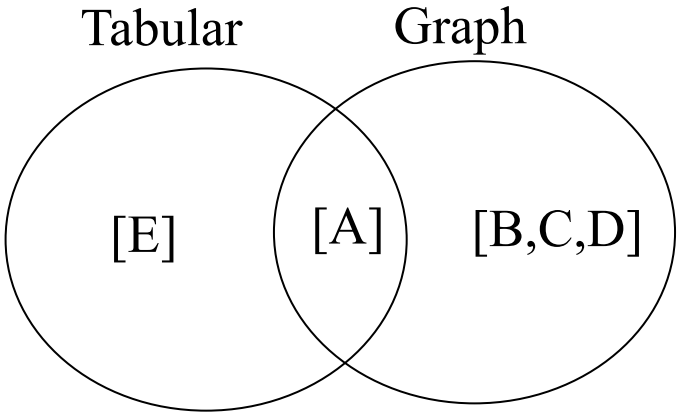
Summary

- Algorithmic fairness is considered from **three** aspects: evaluation metrics, data, and model.

- Propose *Generalized Demographic Parity (GDP)* to broad evaluation metrics for continuous sensitive attributes with tractable computation.
- Propose *Graphair* to conduct automatic graph data augmentation via learnable feature masking and adjacency matrix.
- Propose *Fair Message Passing (FMP)* to achieve graph fairness from model architecture perspective.
- Understand the role of topology in fair graph learning and propose *Fair Graph Rewiring (FairGR)* to mitigate bias from the data perspective.
- Rethink the distribution shifts problem and propose *Robust Fairness Regularization (RFR)* to achieve fairness under distribution shifts.

Summary


Fairness in ML
Metrics
Data
Model
Research
[A]
Research
[B, D]
Research
[C, E]


Tabular
Graph
[E]
[A]
[B,C,D]

Q & A