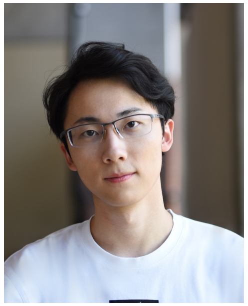
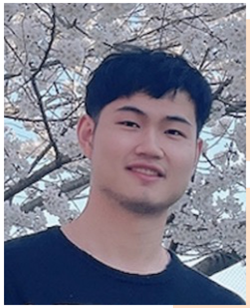
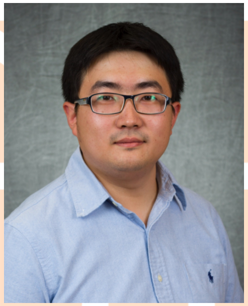
On Structural Explanation of Bias in Graph Neural Networks







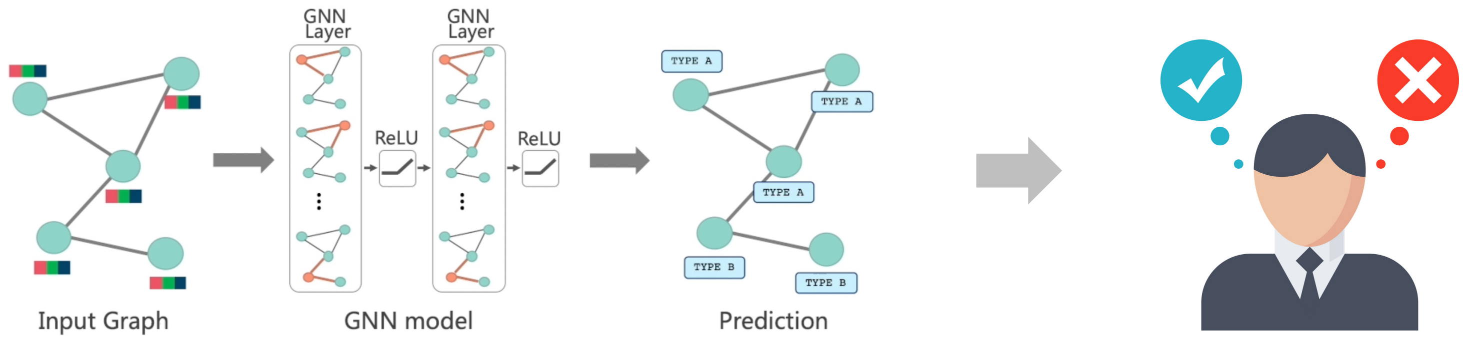
[1]Yushun Dong [1]Song Wang [2]Yu Wang [2]Tyler Derr [1]Jundong Li

[1]University of Virginia

[2]Vanderbilt University

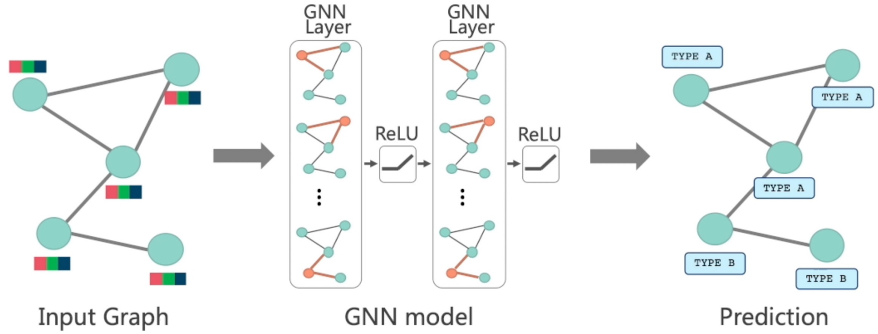
Background Introduction

- Graph Neural Networks (GNNs):
  - handle graph-structured data
  - help decision-making

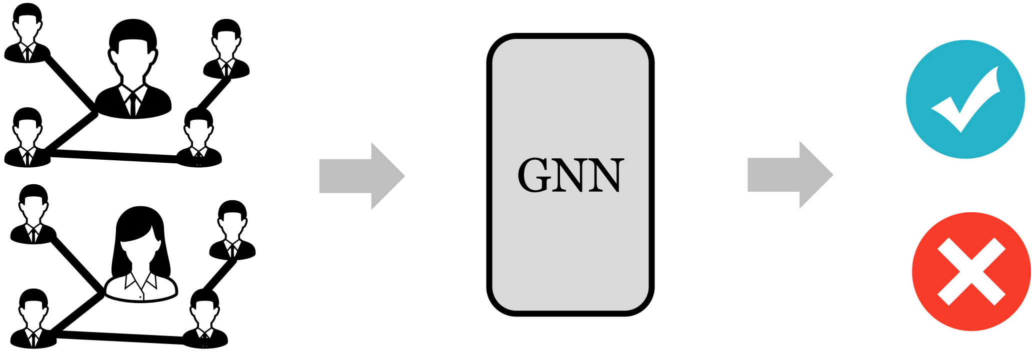


Background Introduction

- Graph Neural Networks (GNNs):
  - handle graph-structured data
  - help decision-making

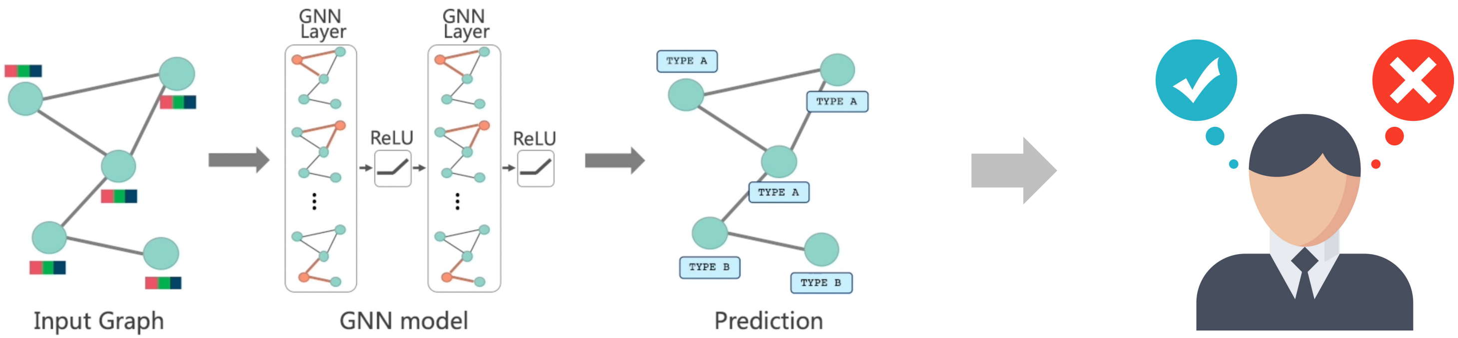


- However: The surrounding structure of a node —Induce bias→ Its corresponding prediction in GNNs

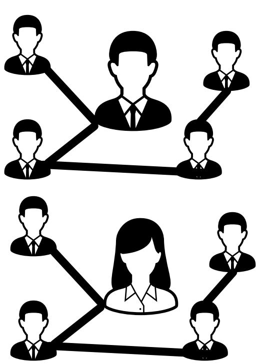


Background Introduction

- Graph Neural Networks (GNNs):
  - handle graph-structured data
  - help decision-making



- However: The surrounding structure of a node → **Induce bias** → Its corresponding prediction in GNNs



"surrounding structure": the edges in a subgraph that centered on this node up to several hops away.

Background Introduction

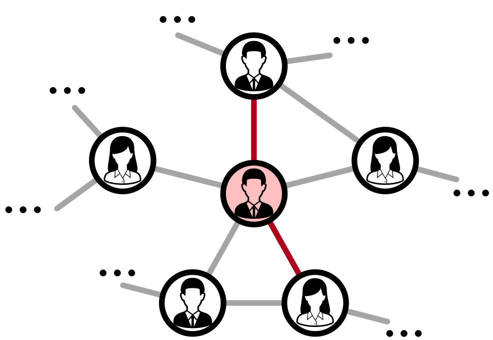
Understanding which edge brings bias is critical.

An example: loan approval decision making in a transaction network.

Background Introduction

Understanding which edge brings bias is critical.

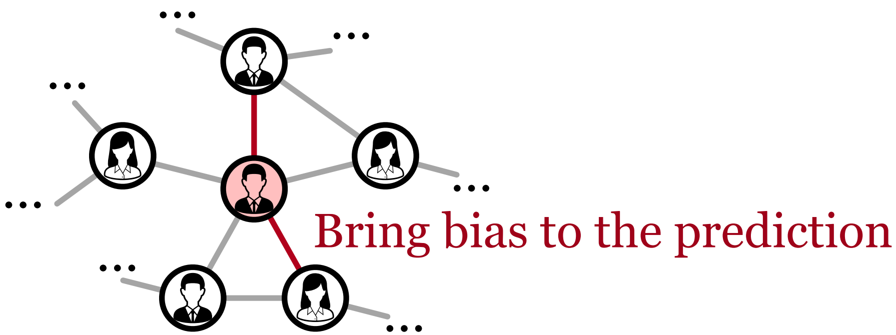
An example: loan approval decision making in a transaction network.



Background Introduction

Understanding which edge brings bias is critical.

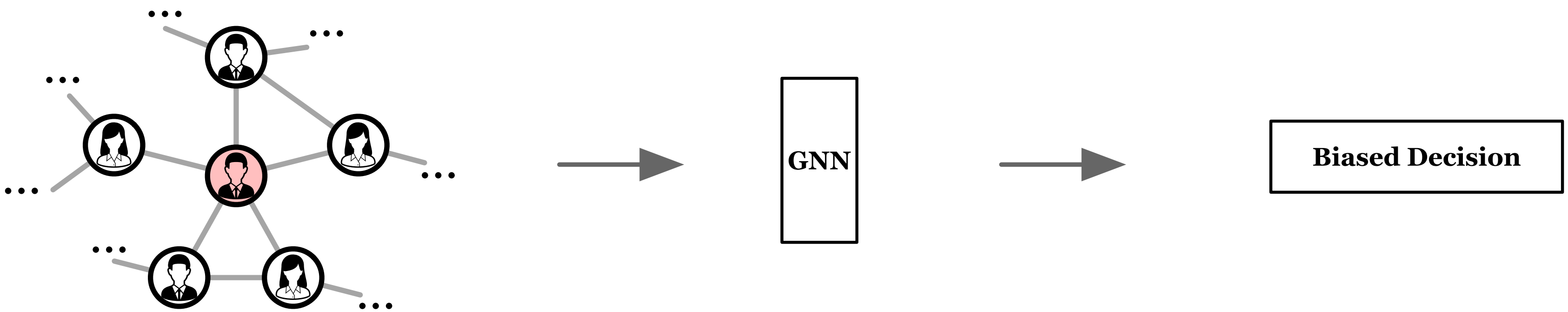
An example: loan approval decision making in a transaction network.



Background Introduction

Understanding which edge brings bias is critical.

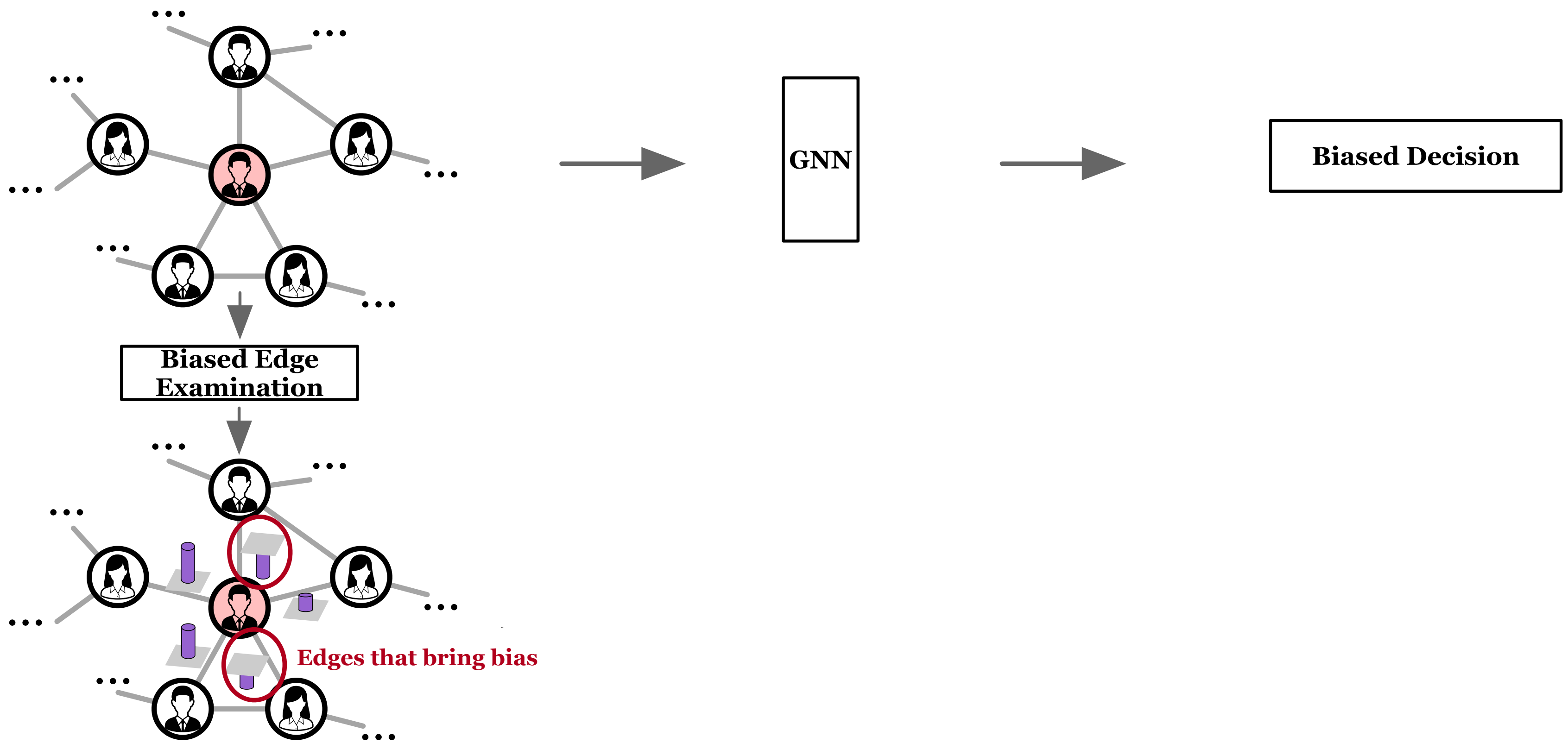
An example: loan approval decision making in a transaction network.



Background Introduction

Understanding which edge brings bias is critical.

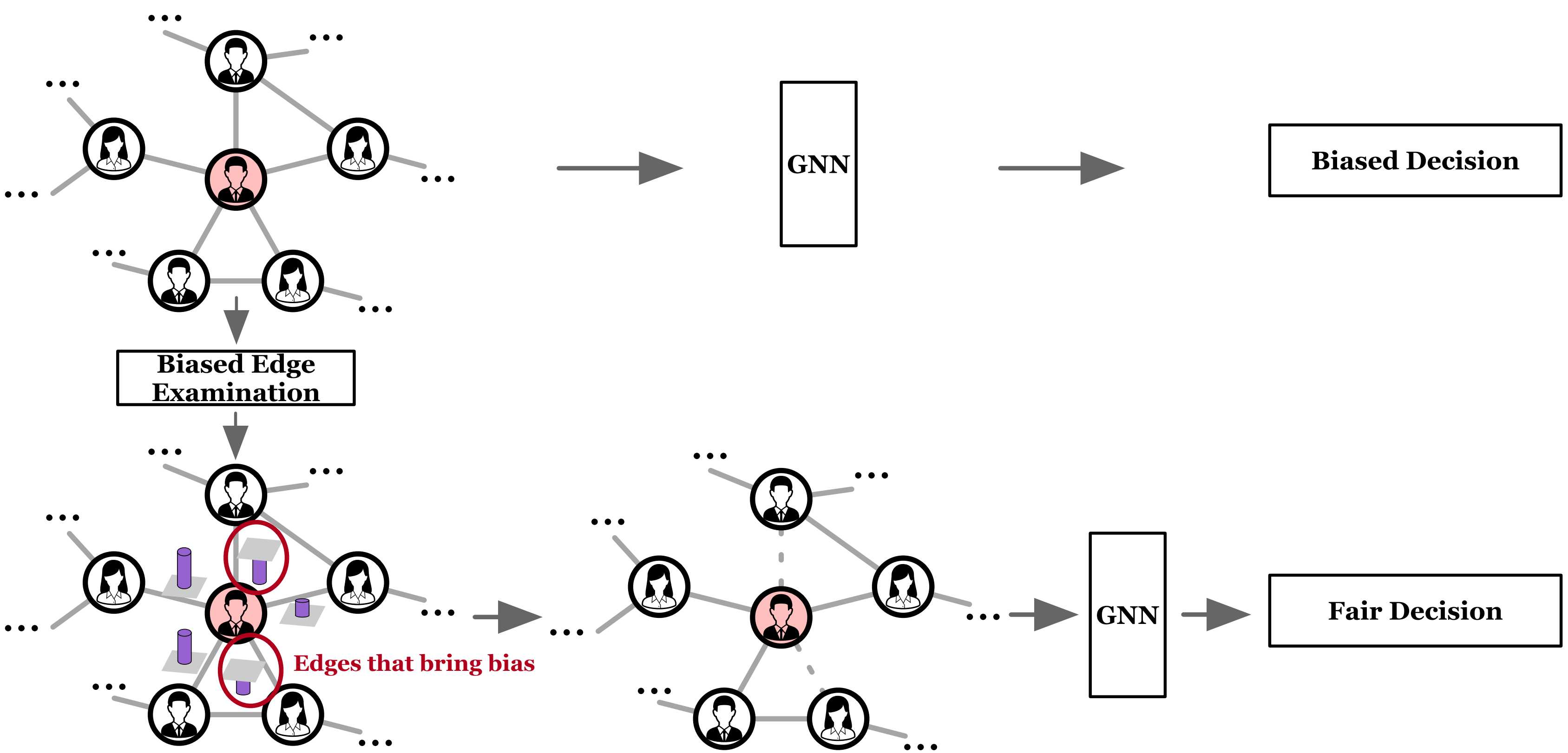
An example: loan approval decision making in a transaction network.



Background Introduction

Understanding which edge brings bias is critical.

An example: loan approval decision making in a transaction network.



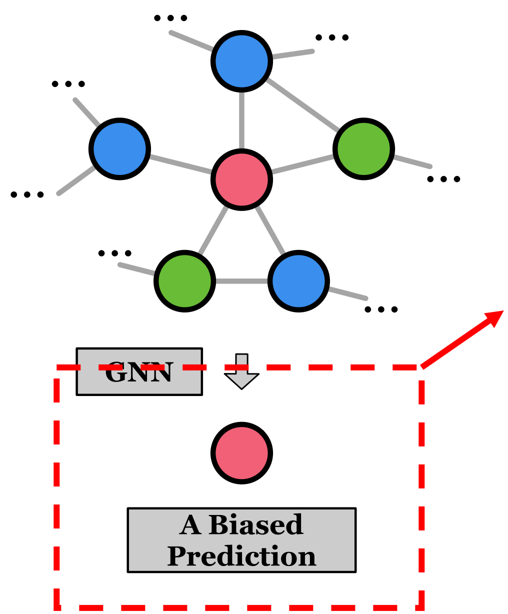
Existing Challenges

Challenges:

Existing Challenges

Challenges:

- (1) Fairness Notion Gap: how to measure the level of bias for the GNN prediction at the instance level?



How to determine the prediction for a specific node as biased?

Existing Challenges

Challenges:

- (1) Fairness Notion Gap: how to measure the level of bias for the GNN prediction at the instance level?
- (2) Usability Gap: to achieve a better understanding, both bias and fairness should be explained.

Existing Challenges

Challenges:

- (1) Fairness Notion Gap: how to measure the level of bias for the GNN prediction at the instance level?
- (2) Usability Gap: to achieve a better understanding, both bias and fairness should be explained.
- (3) Faithfulness Gap: how to obtain bias (fairness) explanations that are faithful to the GNN prediction?

Existing Challenges

Challenges:

- (1) Fairness Notion Gap: how to measure the level of bias for the GNN prediction at the instance level?

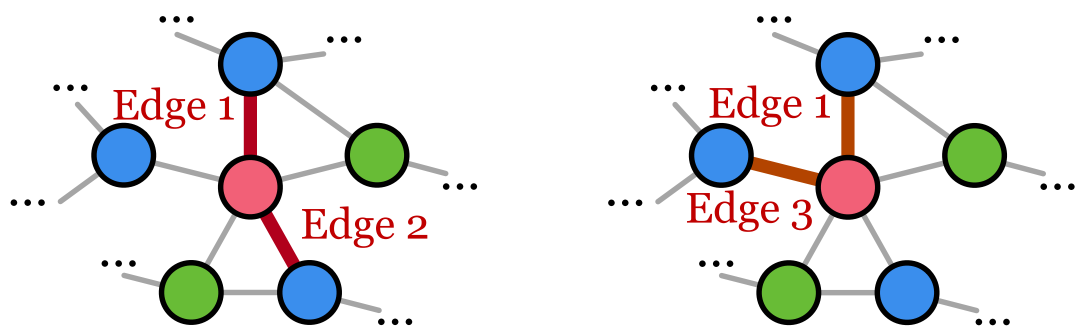
- (2) Usability Gap: to achieve a better understanding, both bias and fairness should be explained.

- (3) Faithfulness Gap: how to obtain bias (fairness) explanations that are faithful to the GNN prediction?

  The obtained explanations should reflect the true reasoning results.

Existing Challenges

Challenges:

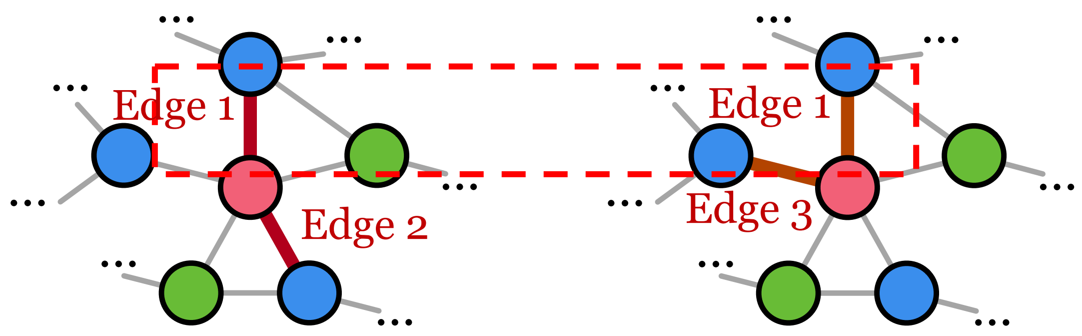


Results would be more biased. Edges the GNN actually relies on.

- (3) Faithfulness Gap: how to obtain bias (fairness) explanations that are faithful to the GNN prediction?

Existing Challenges

Challenges:



Results would be more biased. Edges the GNN actually relies on.

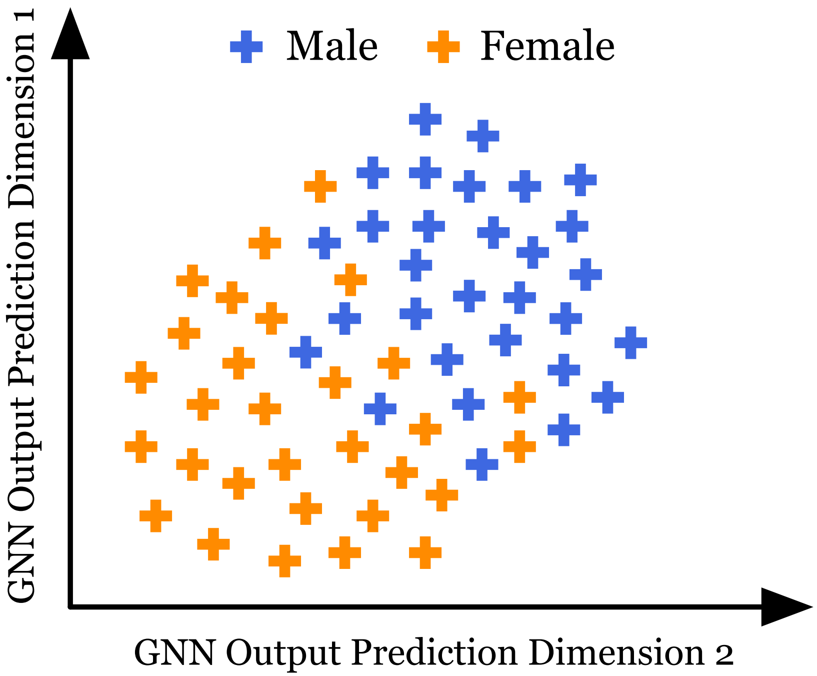
- (3) Faithfulness Gap: how to obtain bias (fairness) explanations that are faithful to the GNN prediction?

Methodology

- Proposed fairness metric: Node-Level Bias in GNNs.

Methodology

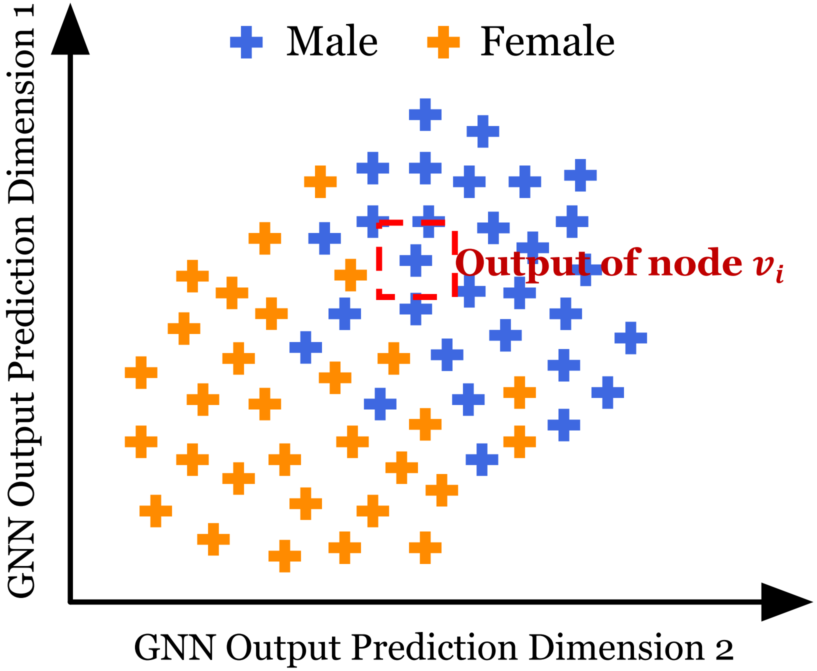
- Proposed fairness metric: Node-Level Bias in GNNs.



Wasserstein Distance ({+}, {+})

Methodology

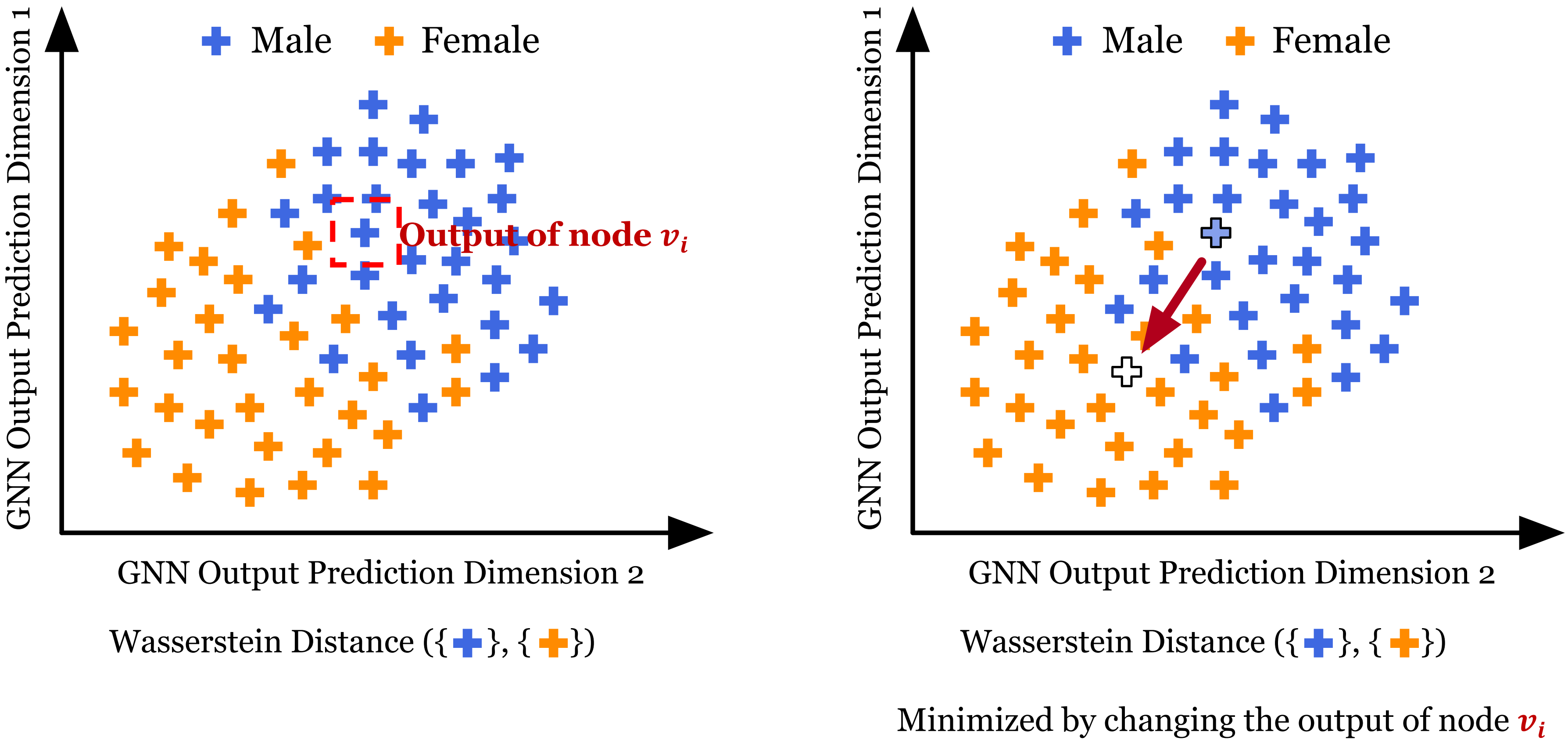
- Proposed fairness metric: Node-Level Bias in GNNs.



Wasserstein Distance ({ }, { })

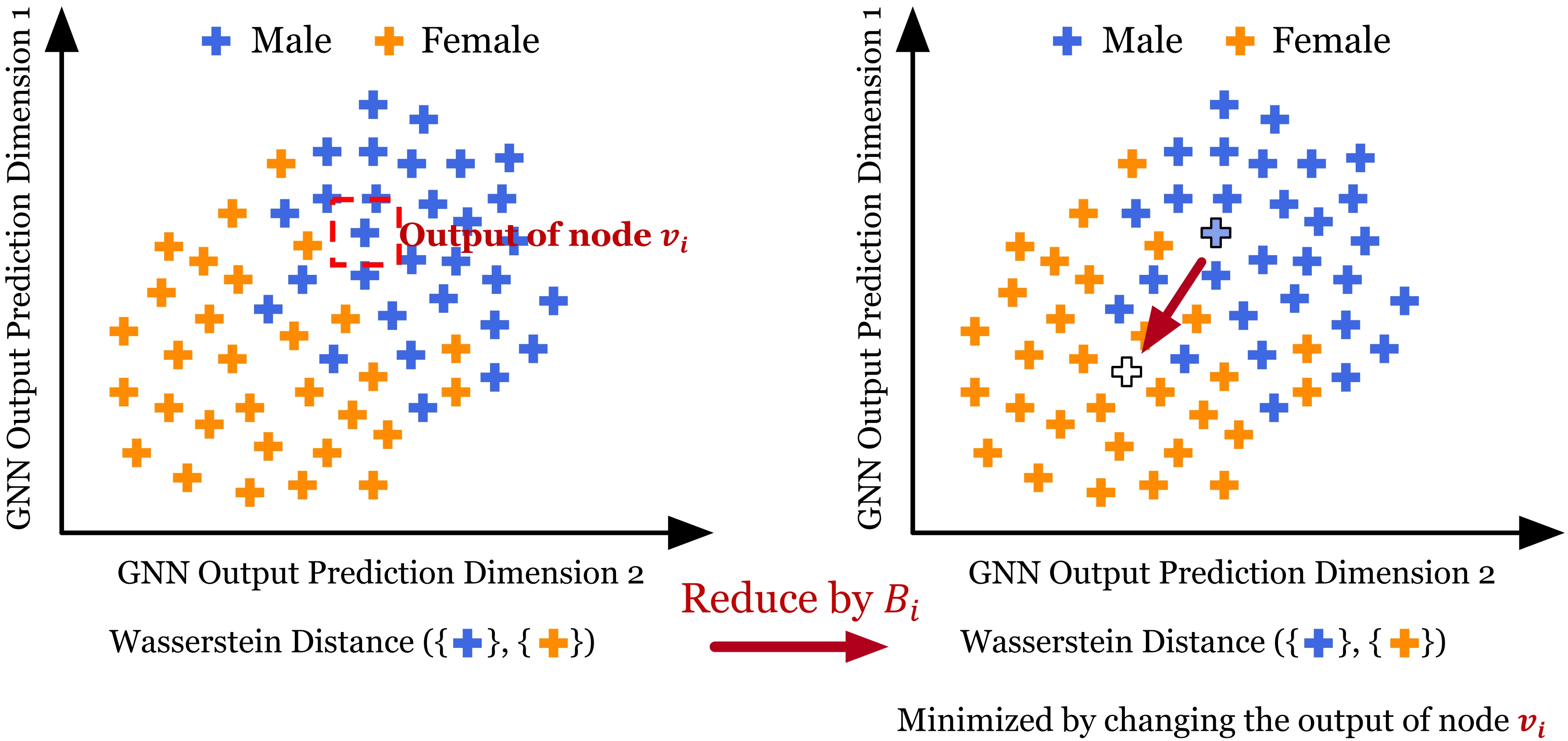
Methodology

- Proposed fairness metric: Node-Level Bias in GNNs.



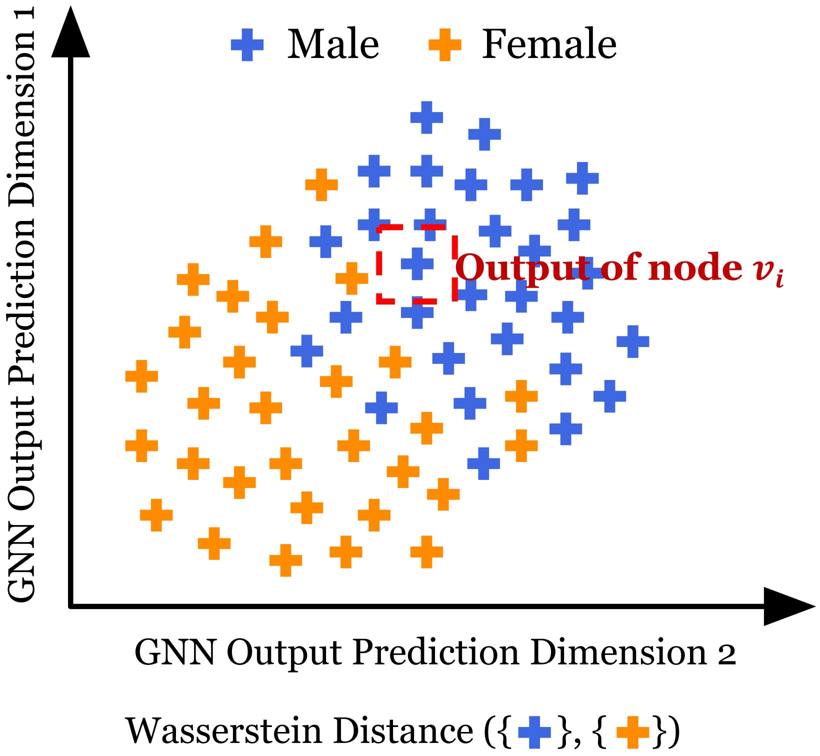
Methodology

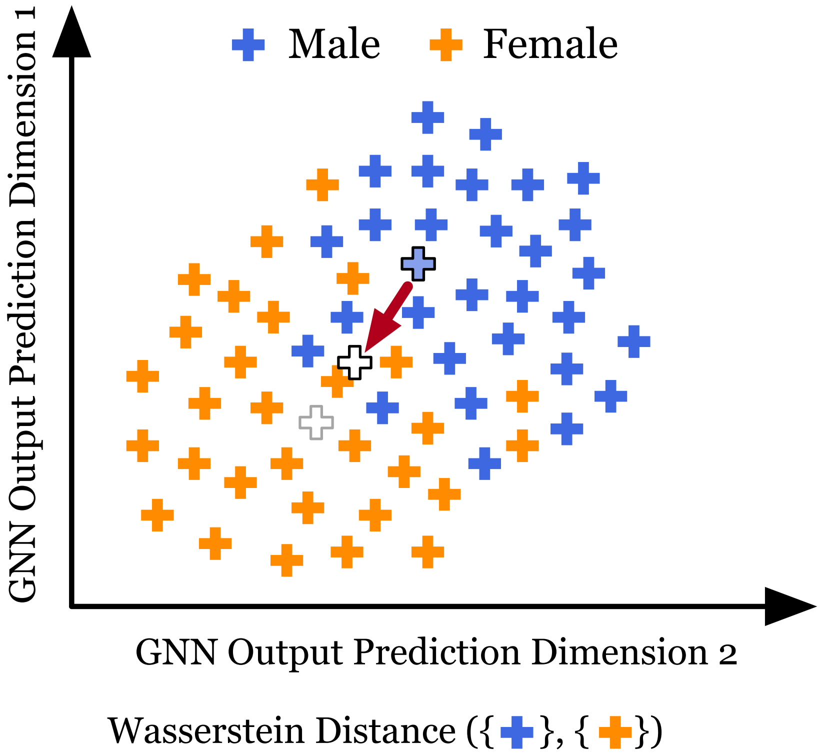
- Proposed fairness metric: Node-Level Bias in GNNs.



Methodology

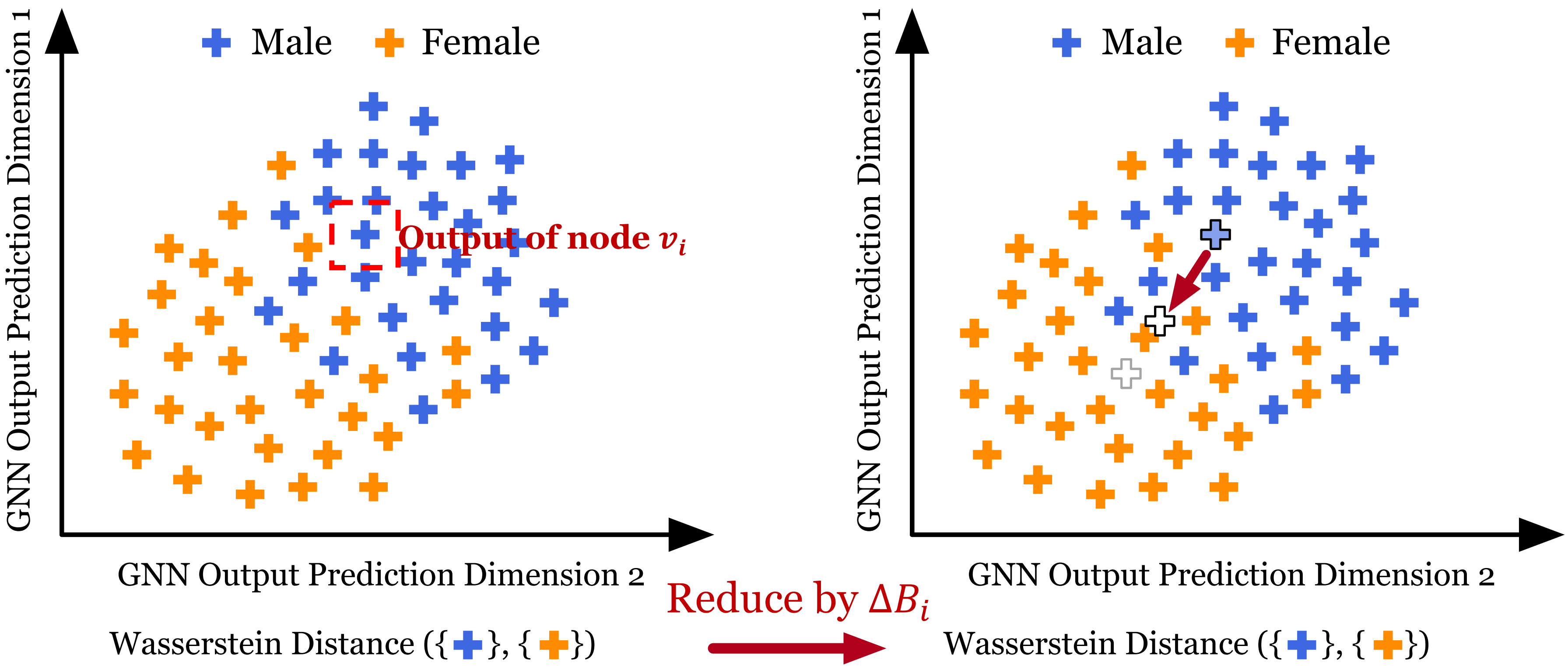
- Proposed fairness metric: Node-Level Bias in GNNs.





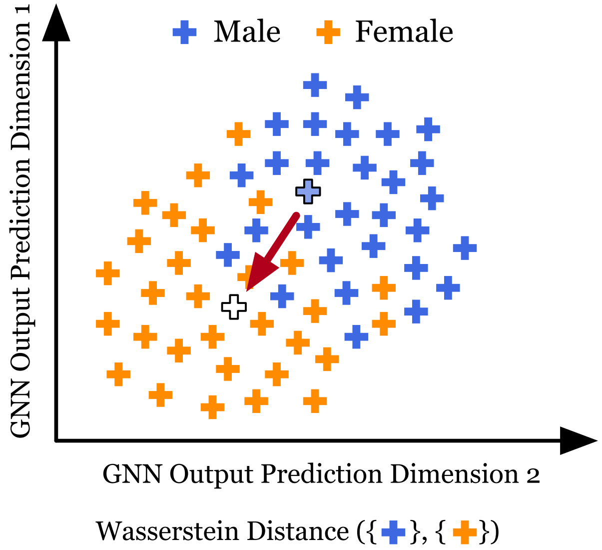
Methodology

- Proposed fairness metric: Node-Level Bias in GNNs.



Methodology

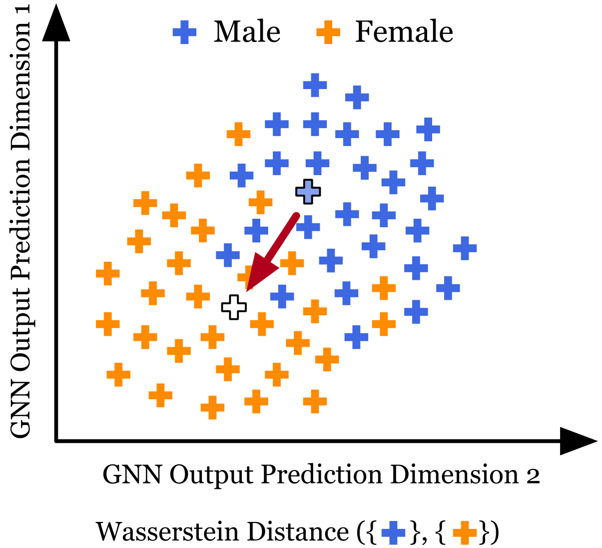
- Proposed fairness metric: Node-Level Bias in GNNs.

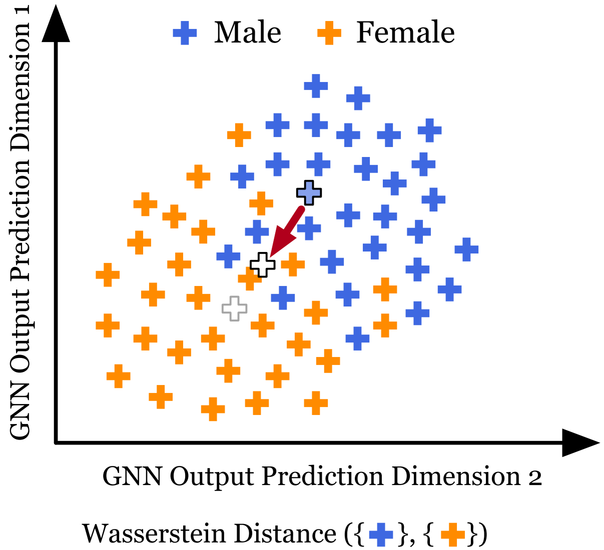


B_i: the potential Wasserstein distance reduction given by node v_i.

Methodology

- Proposed fairness metric: Node-Level Bias in GNNs.



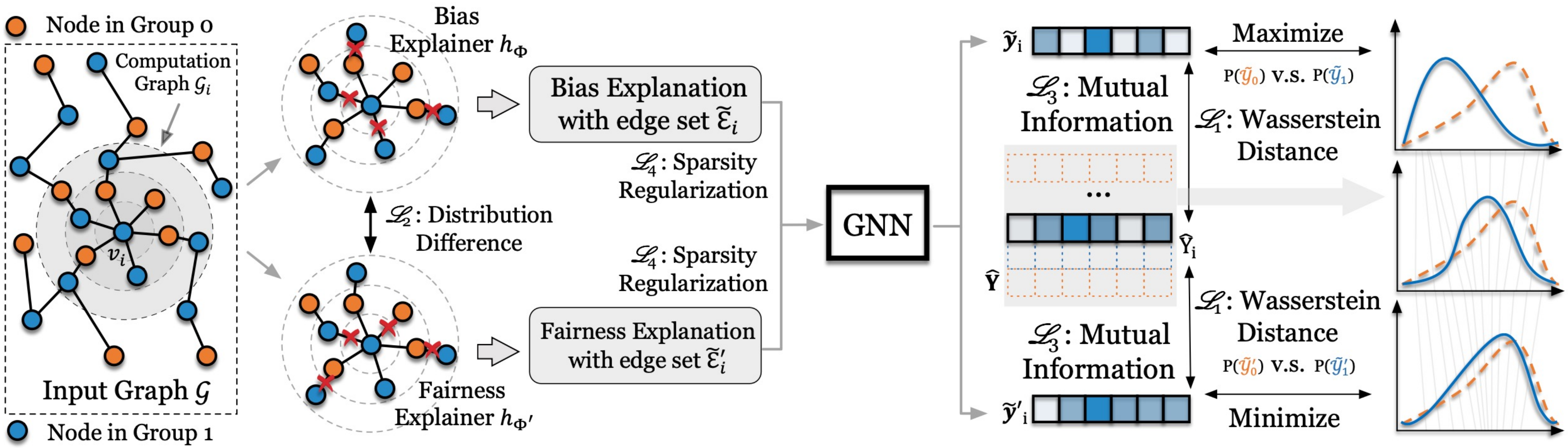


B_{i}: the potential Wasserstein distance reduction given by node v_{i}.

ΔB_{i}: the Wasserstein distance reduction when prediction of node v_{i} is changed.

Methodology

- Proposed framework REFEREE.

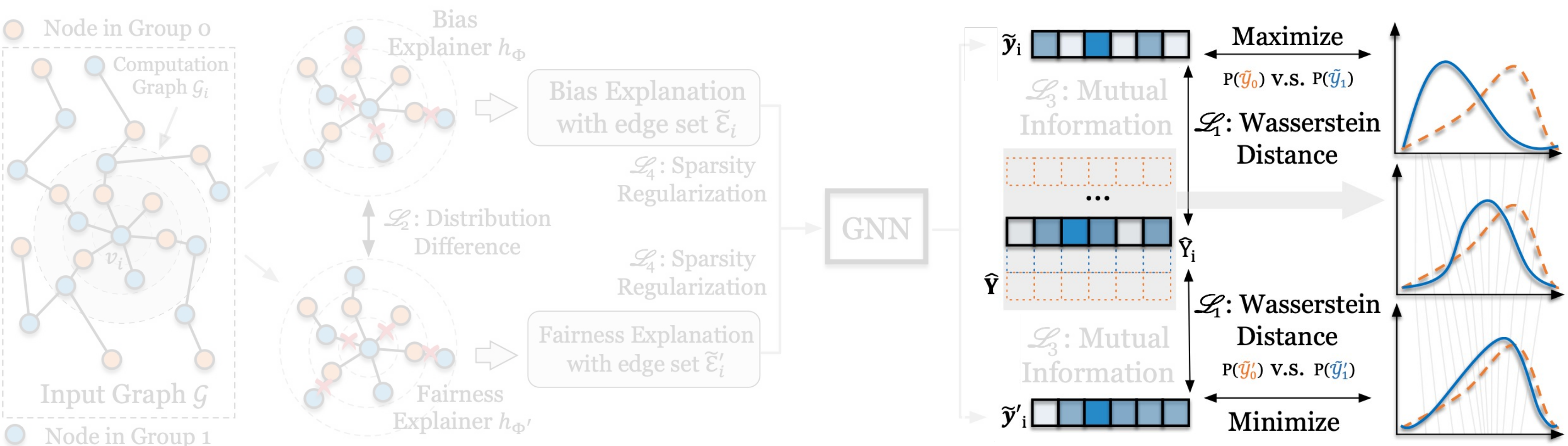


REFEREE includes:
(1) Bias Explainer and (2) Fairness Explainer.

Explainer backbone model:
any differentiable GNN explanation model that identifies edge sets.

Methodology

- Proposed framework REFEREE.

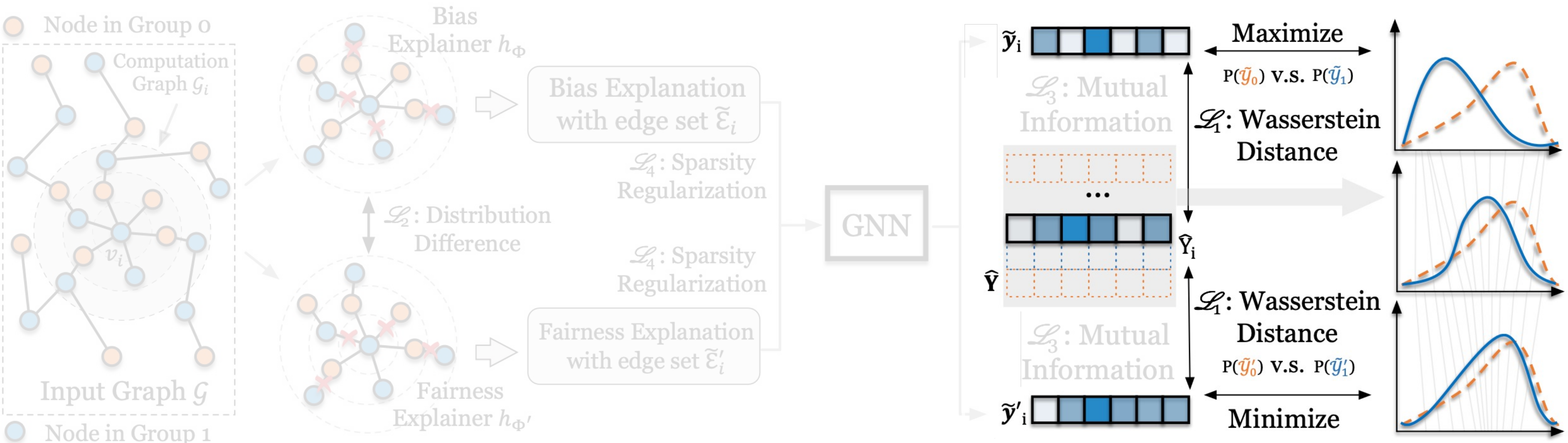


First goal: Explaining Bias and Fairness.

Overall loss for the first goal: $\mathscr{L}_1(\Phi, \Phi') = \mathrm{W}_1(P(\tilde{\mathcal{Y}}'_0), P(\tilde{\mathcal{Y}}'_1)) - \mathrm{W}_1(P(\tilde{\mathcal{Y}}_0), P(\tilde{\mathcal{Y}}_1))$

Methodology

- Proposed framework REFEREE.



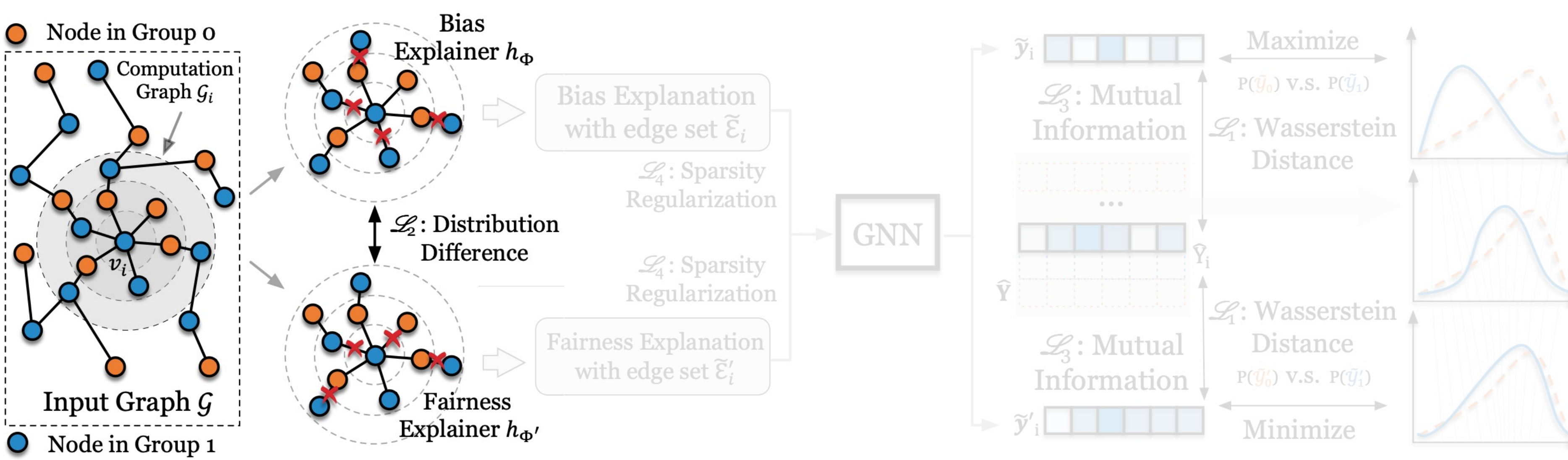
First goal: Explaining Bias and Fairness.

Overall loss for the first goal: $\mathscr{L}_1(\Phi, \Phi') = \boxed{\mathrm{W}_1(P(\tilde{\mathcal{Y}}'_0), P(\tilde{\mathcal{Y}}'_1))} - \boxed{\mathrm{W}_1(P(\tilde{\mathcal{Y}}_0), P(\tilde{\mathcal{Y}}_1))}$

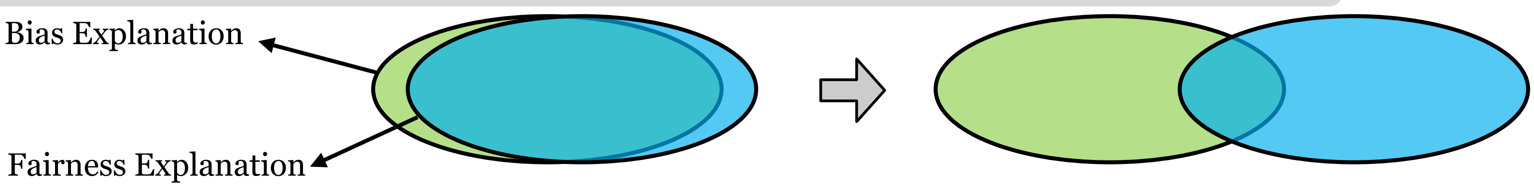
Bias Explainer: $\max_{\tilde{\mathcal{E}}_i} \mathrm{W}_1(P(\tilde{\mathcal{Y}}_0), P(\tilde{\mathcal{Y}}_1))$ Fairness Explainer: $\min_{\tilde{\mathcal{E}}'_i} \mathrm{W}_1(P(\tilde{\mathcal{Y}}'_0), P(\tilde{\mathcal{Y}}'_1))$

Methodology

- Proposed framework REFEREE.

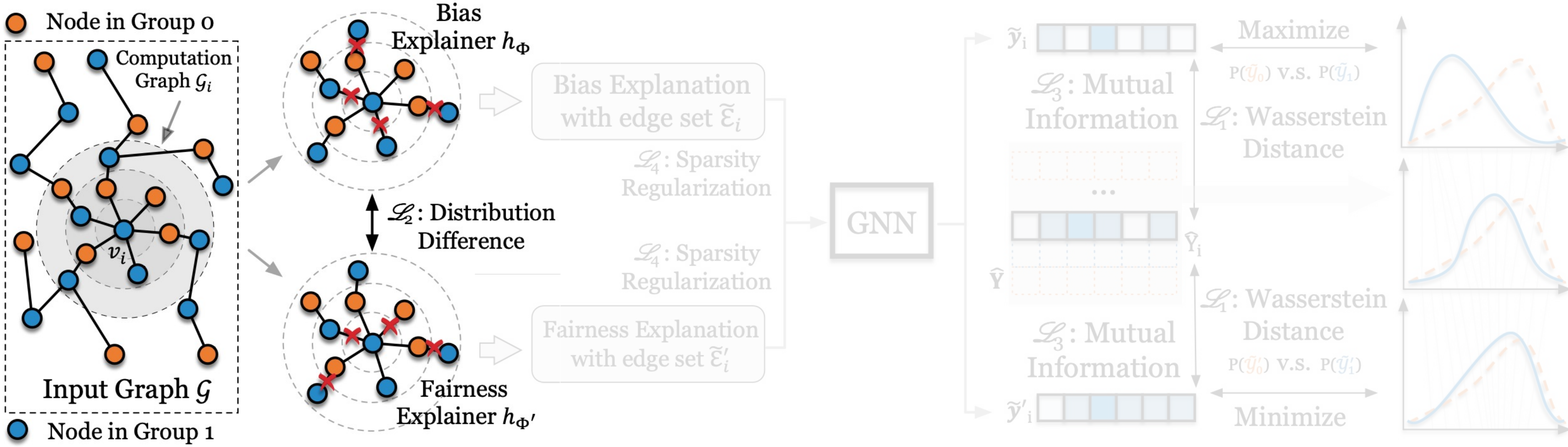


Second goal: Enforcing difference to enhance stability.

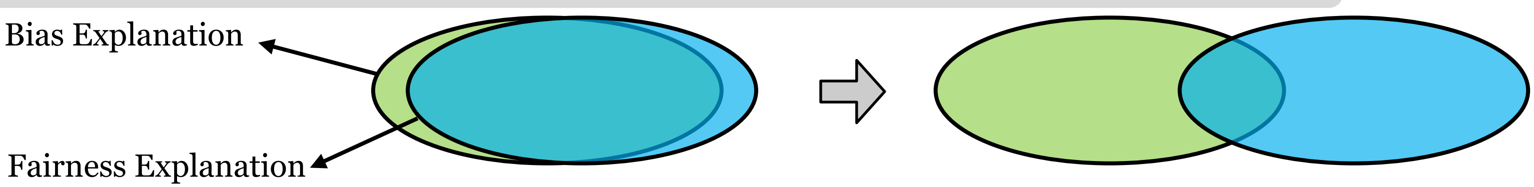


Methodology

- Proposed framework REFEREE.



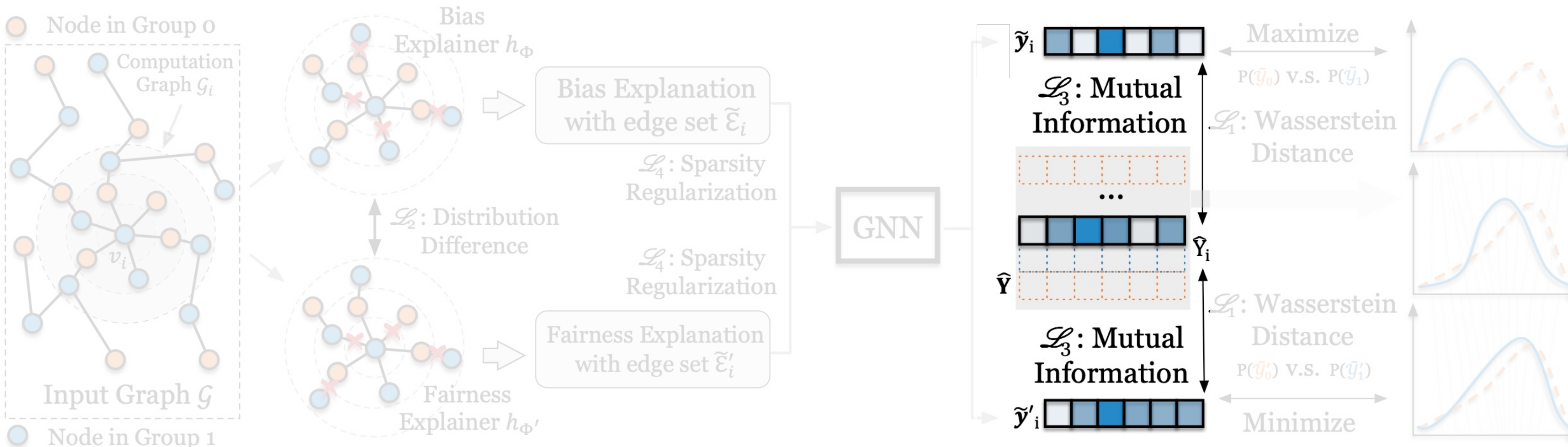
Second goal: Enforcing difference to enhance stability.



Overall loss for goal 2: $\mathscr{L}_2(\Phi, \Phi') = -\text{Dist\_Diff}(P_{\Phi'}(\tilde{\mathcal{E}}'_i|\mathcal{E}_i)||P_{\Phi}(\tilde{\mathcal{E}}_i|\mathcal{E}_i))$

Methodology

- Proposed framework REFEREE.

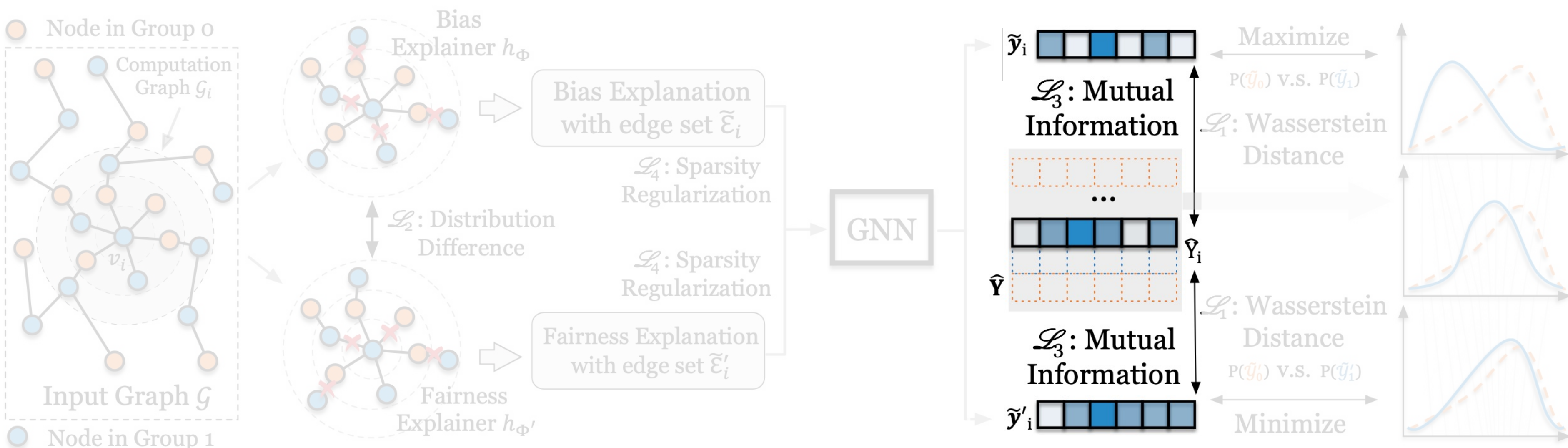


Third goal: Enforcing Fidelity.

Overall loss for the third goal: $\mathscr{L}_3(\Phi, \Phi') = -\mathbb{E}_{\hat{Y}_i|\tilde{\mathcal{G}}_i}[\log P_\Theta(\hat{Y}_i|\tilde{\mathcal{G}}_i)] - \mathbb{E}_{\hat{Y}_i|\tilde{\mathcal{G}}'_i}[\log P_\Theta(\hat{Y}_i|\tilde{\mathcal{G}}'_i)]$

Methodology

- Proposed framework REFEREE.



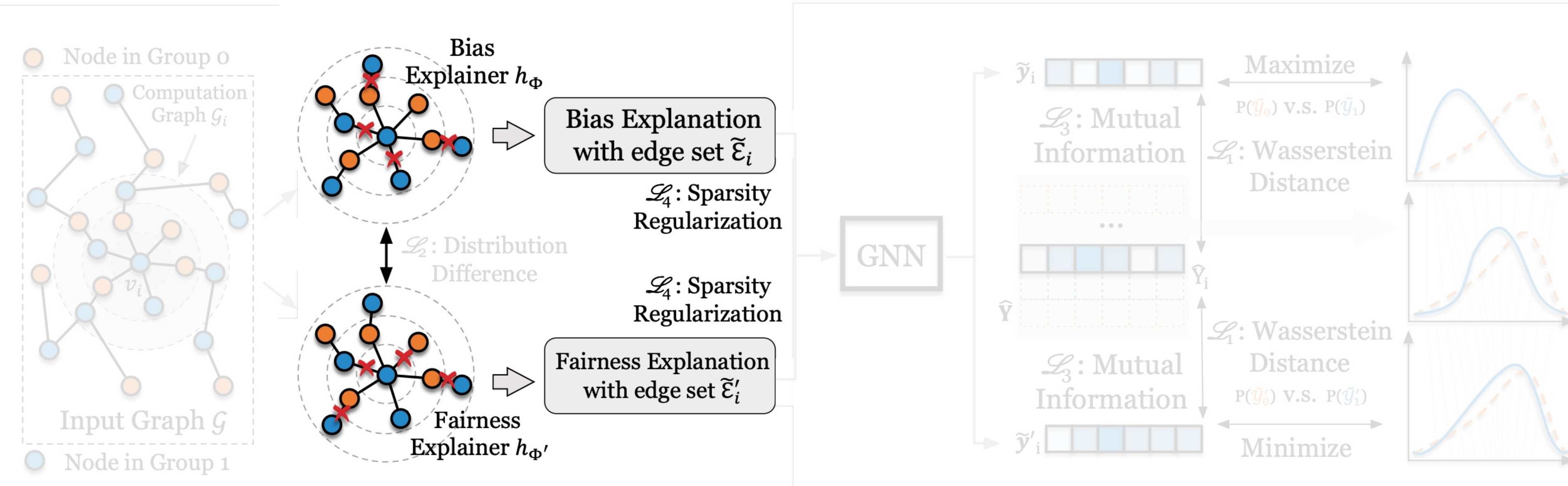
Third goal: Enforcing Fidelity.

Overall loss for the third goal: $\mathscr{L}_3(\Phi, \Phi') = \boxed{-\mathbb{E}_{\hat{Y}_i|\tilde{\mathcal{G}}_i}[\log P_\Theta(\hat{Y}_i|\tilde{\mathcal{G}}_i)]} \boxed{-\mathbb{E}_{\hat{Y}_i|\tilde{\mathcal{G}}'_i}[\log P_\Theta(\hat{Y}_i|\tilde{\mathcal{G}}'_i)]}$

Bias Explainer: $\max_{\tilde{\mathcal{G}}_i} \mathrm{MI}(\hat{Y}_i, \tilde{\mathcal{G}}_i)$ Fairness Explainer: $\max_{\tilde{\mathcal{G}}'_i} \mathrm{MI}(\hat{Y}_i, \tilde{\mathcal{G}}'_i)$

Methodology

- Proposed framework REFEREE.



Fourth goal: Refining Explanation.

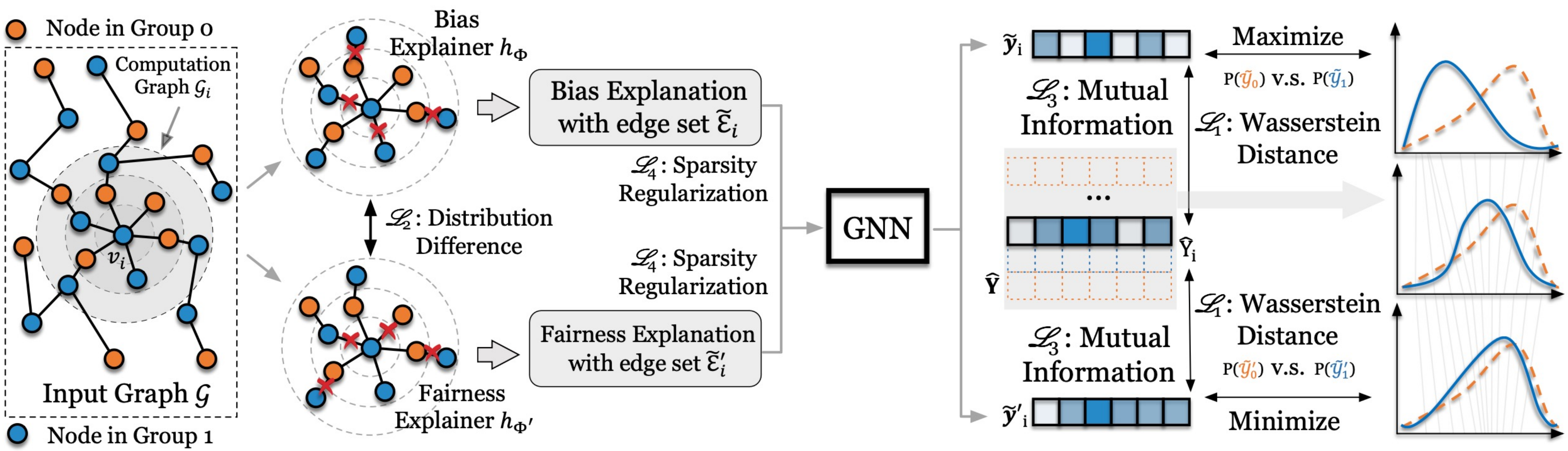
Overall loss for the fourth goal:

$$\mathscr{L}_4(\Phi, \Phi') = \|\mathrm{M}\|_1 + \|\mathrm{M}'\|_1$$

$$\mathscr{L}_4(\Phi, \Phi', T, T') = \mathrm{ReLU}(\|\mathrm{M}\|_1 - T) + \mathrm{ReLU}(\|\mathrm{M}'\|_1 - T')$$

Methodology

- Proposed framework REFEREE.

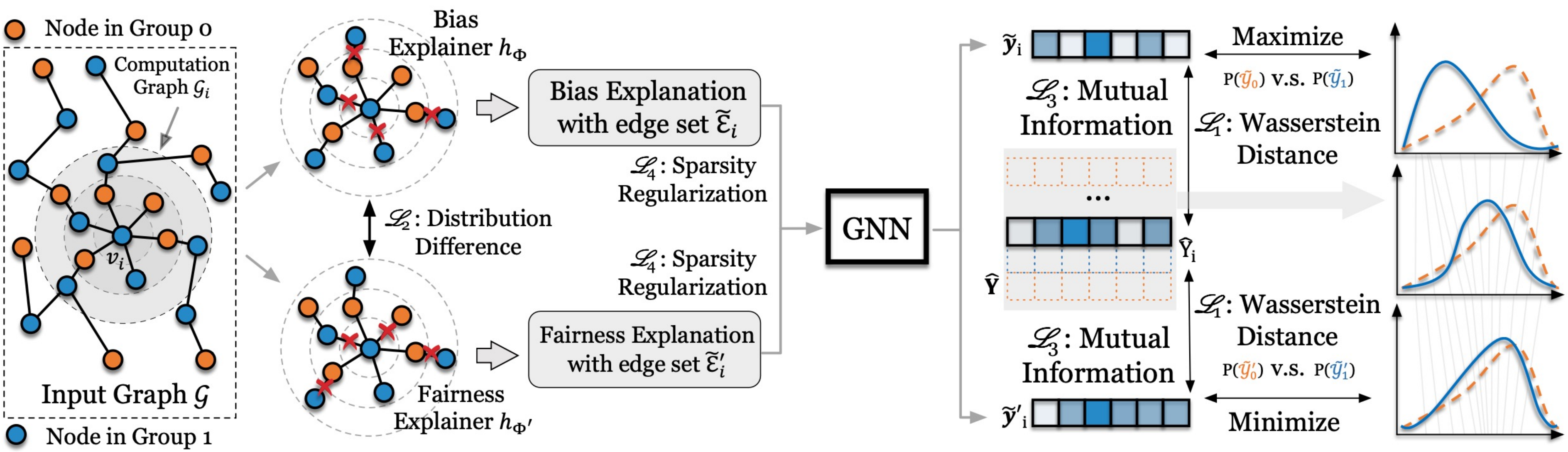


Overall loss for all goals above:

$$\mathscr{L} = \mathscr{L}_1 + \alpha\mathscr{L}_2 + \beta\mathscr{L}_3 + \gamma\mathscr{L}_4$$

Methodology

- Proposed framework REFEREE.



Overall loss for all goals above:

$$\mathscr{L} = \mathscr{L}_1 + \alpha\mathscr{L}_2 + \beta\mathscr{L}_3 + \gamma\mathscr{L}_4$$

Hyperparameters:

Controlling: (1) difference between explanations; (2) fidelity; (3) refinement.

Experiments

- Downstream Task: Node classification.
- 3 Real-world Datasets: German [1], Recidivism [2], and Credit [3].
- 2 Explanation Framework Backbones: GNN Explainer [4] and PGExplainer [5].
- 4 Baselines: Attention-based explanation [4], gradient-based explanation [4], vanilla GNN Explainer [4], vanilla PGExplainer [5].
- 4 Evaluation Metrics: the proposed ΔB (introduced in previous slides), Fidelity– score [6], Δ_{SP} [7], and Δ_{EO} [8];

Dataset	**German Credit**	**Recidivism**	**Credit Defaulter**
# Nodes	1,000	18,876	30,000
# Edges	22,242	321,308	1,436,858
# Attributes	27	18	13
Avg. degree	44.5	34.0	95.8
Sens.	Gender	Race	Age
Label	Good / Bad	Bail / No Bail	Default / No Default

Experiments

- Effectiveness of Explaining Bias (Fairness).

To enable the adopted baselines identify bias/fairness explanations, we also add the loss term explaining bias and fairness on them.

	German		Recidivism		Credit	
	ΔB_i (Promoted)	ΔB_i (Reduced)	ΔB_i (Promoted)	ΔB_i (Reduced)	ΔB_i (Promoted)	ΔB_i (Reduced)
Att.	6.11 ± 2.51	7.84 ± 3.48	4.58 ± 1.67	7.18 ± 2.24	6.72 ± 0.75	8.48 ± 3.29
Grad.	4.27 ± 0.98	5.60 ± 1.85	3.59 ± 2.02	4.42 ± 2.01	5.97 ± 1.07	9.79 ± 1.78
GNN Explainer	5.17 ± 1.20	3.37 ± 1.53	1.74 ± 0.72	3.55 ± 2.08	7.41 ± 1.75	9.24 ± 2.66
PGExplainer	8.73 ± 0.74	9.37 ± 1.87	6.36 ± 2.39	8.66 ± 1.82	7.48 ± 2.70	10.54 ± 3.22
GE-REFEREE	14.29 ± 2.73	**14.45 ± 2.29**	**13.94 ± 3.74**	12.05 ± 2.79	10.30 ± 2.64	**15.07 ± 3.35**
PGE-REFEREE	**15.72 ± 2.31**	11.97 ± 2.62	10.39 ± 4.08	**12.57 ± 3.12**	**11.57 ± 2.91**	14.67 ± 3.49

Observations: REFEREE achieves the best performance over alternatives on identifying

(1) edges that account for the exhibited bias;

(2) edges that are helpful to fulfill fairness;

Experiments

- Explanation Fidelity.

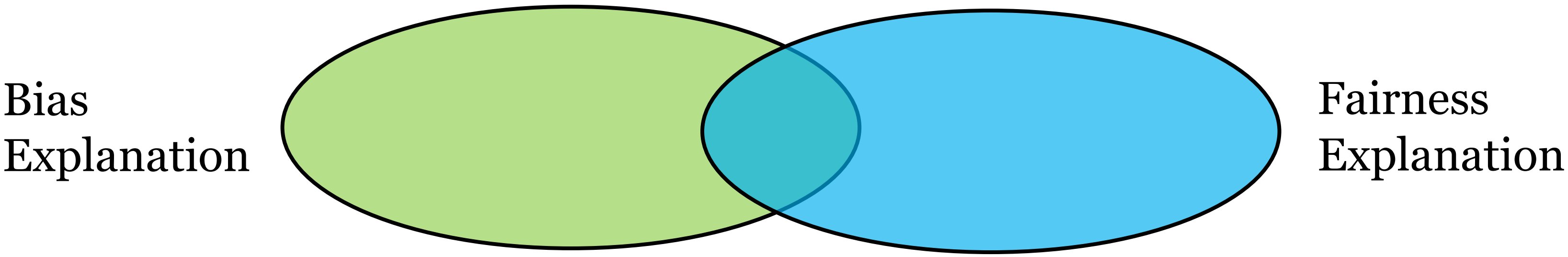
Here, Fidelity– generally measures to what extent the GNN model maintains the vanilla predictions based on the identified explanations.

		German	Recidivism	Credit
GCN	Vanilla	88.02 ± 1.48	90.04 ± 1.43	85.26 ± 1.67
	B. Explainer	**92.20 ± 1.39**	90.26 ± 3.24	87.60 ± 2.79
	F. Explainer	89.17 ± 0.85	**92.08** ± 2.44	**89.41** ± 4.08
GAT	Vanilla	83.65 ± 3.02	87.91 ± 2.04	**88.64 ± 3.41**
	B. Explainer	**85.71 ± 2.31**	90.51 ± 4.58	86.09 ± 2.07
	F. Explainer	84.40 ± 1.57	**91.98 ± 3.95**	87.04 ± 3.10
GIN	Vanilla	88.58 ± 2.50	**91.77 ± 1.42**	87.62 ± 2.60
	B. Explainer	88.11 ± 1.78	90.26 ± 4.13	86.47 ± 2.13
	F. Explainer	**89.67 ± 2.23**	91.45 ± 1.78	**88.17 ± 2.98**

Observations:
Both Bias Explainer and Fairness Explainer achieve comparable performance on fidelity with the vanilla GNN Explainer across different datasets and GNNs.

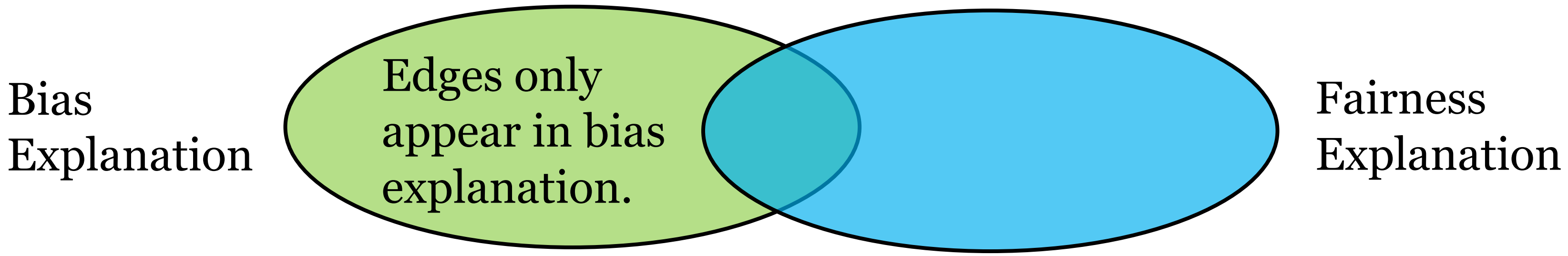
Experiments

- Debiasing GNNs with Explanations.



Experiments

- Debiasing GNNs with Explanations.

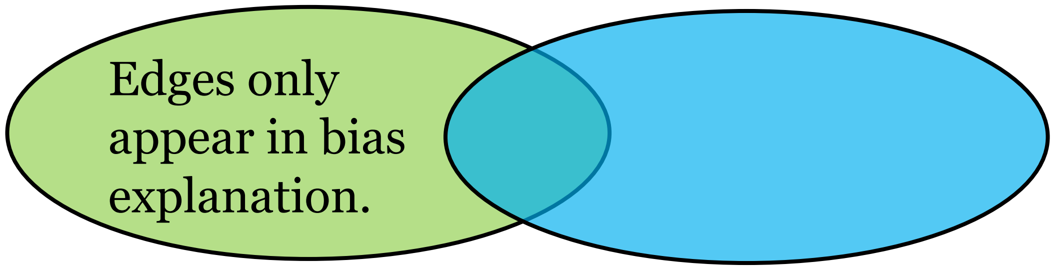


A strategy: deleting edges only appear in bias explanation (but not in fairness explanation) to help achieve a balance between GNN debiasing and utility.

Experiments

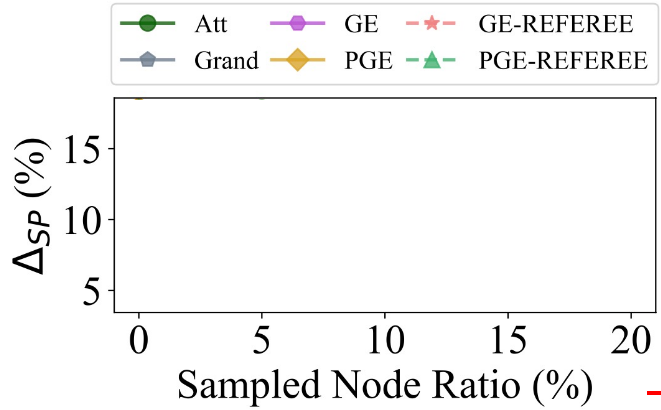
- Debiasing GNNs with Explanations.

Bias Explanation



Fairness Explanation

A strategy: deleting edges only appear in bias explanation (but not in fairness explanation) to help achieve a balance between GNN debiasing and utility.

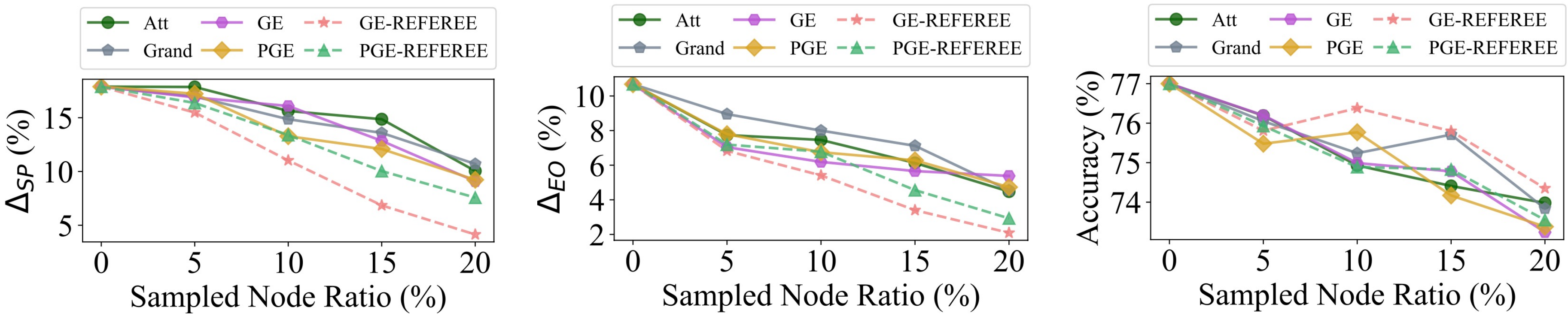


The value changes of adopted metrics (Δ_{SP} in this example).

Edges only appear in bias explanations are deleted (for sampled nodes).

Experiments

- Debiasing GNNs with Explanations.



Observations:
(1) With more edges that only appear in the bias explanations being removed, both Δ_{SP} and Δ_{EO} reduce significantly.
(2) Removing the edges that only appear in the bias explanations generally reduces the GNN prediction accuracy.
(3) REFEREE leads to limited accuracy reduction but achieves a more significant reduction on Δ_{SP} and Δ_{EO}.

Conclusions

(1) A novel explanation problem is studied: how does the surrounding structure of a node influences the bias level of its corresponding GNN prediction?

(2) A novel explanation framework is proposed: stRuctural Explanation oF biasEs in gRaph nEural nEtworks (REFEREE);

(3) Extensive experiments corroborate the effectiveness of REFEREE on rendering effective explanations and helping GNN debiasing.

Acknowledgement

This material is supported by the National Science Foundation (NSF) under grants No. 2006844 and the Cisco Faculty Research Award.

Reference

[1] Dheeru Dua and Casey Graff. UCI machine learning reposi- tory, 2017.
[2] Jordan, K. L., & Freiburger, T. L. (2015). The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. Journal of Ethnicity in Criminal Justice, 13(3), 179-196.
[3] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert systems with applications, 36(2), 2473-2480.
[4] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnn explainer: A tool for post-hoc explanation of graph neural networks. arXiv preprint arXiv:1903.03894.
[5] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. Advances in neural information processing systems, 33, 19620-19631.
[6] Yuan, H., Yu, H., Gui, S., & Ji, S. (2020). Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv:2012.15445.
[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).
[8] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

The End

VIRGINIA DATA SCIENCE AND ANALYTICS (VAST) LAB

Thanks for listening!