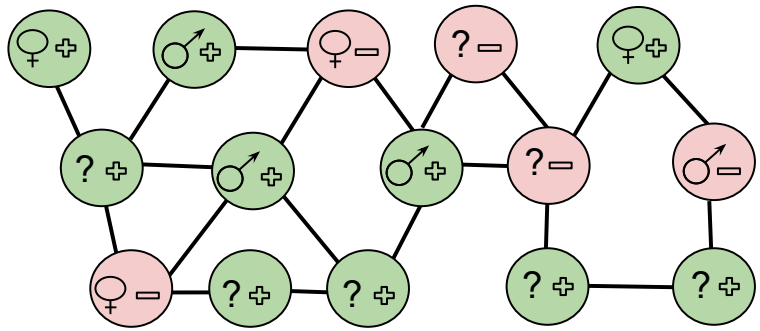
Fair Graph Machine Learning under Adversarial Missingness Processes

International Conference on Learning Representations, 2026

Debolina Halder Lina, Arlei Silva

Rice University

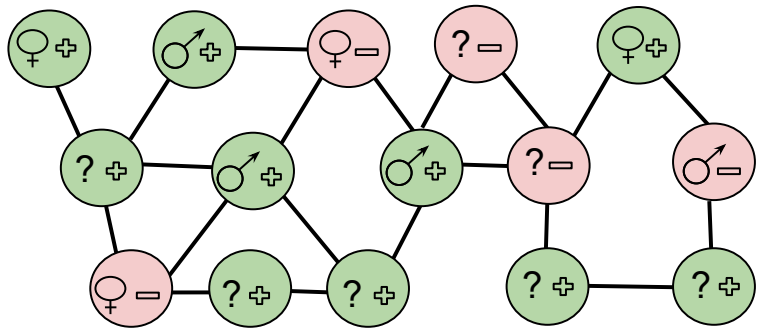
Fair Classifiers and Sensitive Information



Fair classification requires sensitive information

Protected group membership (e.g., race, age, gender)

Fair Classifiers and Sensitive Information



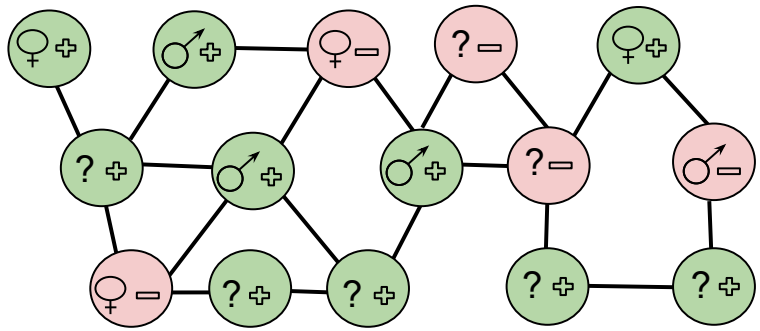
Fair classification requires sensitive information

Protected group membership (e.g., race, age, gender)

Current literature on fair classifiers

1. Assumes complete access to sensitive attributes, or
2. Applies independent imputation (random missingness).

Fair Classifiers and Sensitive Information



Fair classification requires sensitive information

Protected group membership (e.g., race, age, gender)

Current literature on fair classifiers

1. Assumes complete access to sensitive attributes, or
2. Applies independent imputation (random missingness).

Real-world scenario

- Some groups are reluctant to share sensitive information.
- Missingness process is not random.

Adversarially Missing Sensitive Information

Adversarially missing sensitive information

- Adversary partially hides sensitive value.
- Missing values are imputed independently.
- A learned fair model learned based on imputed values can still exhibit unfair behavior on the complete data.

Adversarially Missing Sensitive Information

Adversarially missing sensitive information

- Adversary partially hides sensitive value.
- Missing values are imputed independently.
- A learned fair model learned based on imputed values can still exhibit unfair behavior on the complete data.

Problems with independent imputation:

- Can potentially underestimate the true bias in the dataset.
- Imputed training set might appear unbiased when it is not.
- Model might fail to minimize the true bias.

Adversarially Missing Sensitive Information

Adversarially missing sensitive information

- Adversary partially hides sensitive value.
- Missing values are imputed independently.
- A learned fair model learned based on imputed values can still exhibit unfair behavior on the complete data.

Problems with independent imputation:

- Can potentially underestimate the true bias in the dataset.
- Imputed training set might appear unbiased when it is not.
- Model might fail to minimize the true bias.

How to achieve fairness under adversarially missing sensitive information?

Proposed Model

The fair classifier optimizes:

$$\theta^*_{class} = \arg\min_{\theta_{class}} \mathcal{L}_{class} + \alpha \mathbb{E}_{s \sim \mathcal{P}_s}[\mathcal{L}_{bias}]$$

where $\mathcal{P}_s$ is the sensitive attribute distribution.

Proposed Model

The fair classifier optimizes:

$$\theta^*_{class} = \arg\min_{\theta_{class}} \mathcal{L}_{class} + \alpha \mathbb{E}_{s \sim \mathcal{P}_s}[\mathcal{L}_{bias}]$$

where $\mathcal{P}_s$ is the sensitive attribute distribution.

True $\mathcal{P}_s$ is not recoverable under adversarial missingness.

Proposed Model

The fair classifier optimizes:

$$\theta^*_{class} = \arg\min_{\theta_{class}} \mathcal{L}_{class} + \alpha \mathbb{E}_{s\sim\mathcal{P}_s}[\mathcal{L}_{bias}]$$

where $\mathcal{P}_s$ is the sensitive attribute distribution.

True $\mathcal{P}_s$ is not recoverable under adversarial missingness.

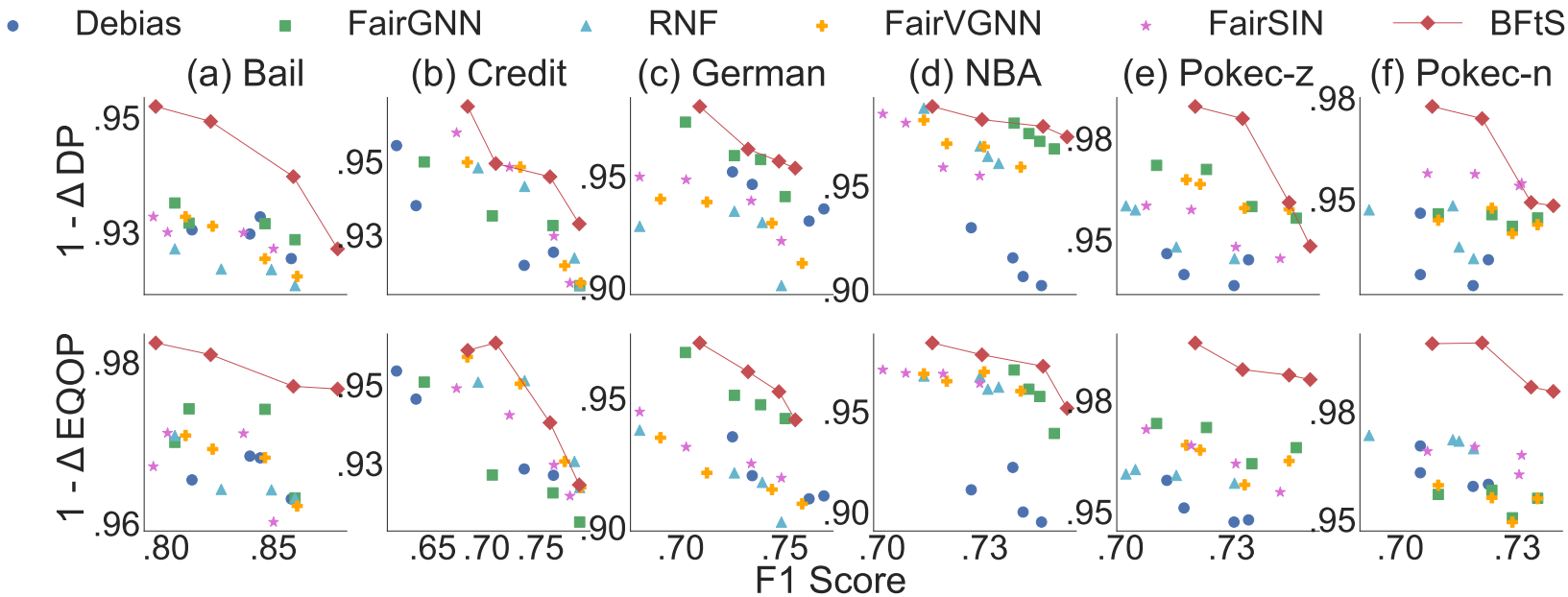
Our approach: define uncertainty set $\mathcal{U}$ of plausible distributions; minimize worst-case bias

$$\theta^*_{class} = \arg\min_{\theta_{class}} \mathcal{L}_{class} + \alpha \max_{u\in\mathcal{U}} \mathbb{E}_{s\sim u}[\mathcal{L}_{bias}]$$

We propose: Better Fair than Sorry(BFtS)

- A three-player adversarial learning framework
- Assumes the worst fairness scenario in training data and produces challenging values for fairness.

Results and Analysis



Fairness vs. accuracy for all the methods. The top right corner of the plot represents a high F1 with high fairness.

BFtS achieves the best fairness vs. accuracy trade-off.

Conclusions

We analyze how adversarial missingness can bias fair GNNs.

We introduce BFtS, a 3-player adversarial framework that imputes worst-case sensitive attributes.

We establish theoretical fairness guarantees for BFtS.

We prove that BFtS is robust to convergence issues in adversarial learning.

We demonstrate that BFtS achieves a stronger fairness–accuracy trade-off than baseline methods.