Your Neighbor Matters: Towards Fair Decisions Under Networked Interference

Wenjing Yang*
Department of Intelligent Data Science, College of Computer, National University of Defense Technology
Changsha, China
wenjing.yang@nudt.edu.cn

Haotian Wang†
Department of Intelligent Data Science, College of Computer, National University of Defense Technology
Changsha, China
accwht@hotmail.com

Haoxuan Li
Peking University
Beijing, China
hxli@stu.pku.edu.cn

Hao Zou
ZGC laboratory
Beijing, China
ahio@163.com

Ruochun Jin
Department of Intelligent Data Science, College of Computer, National University of Defense Technology
Changsha, China
jinrc@nudt.edu.cn

Kun Kuang‡
Institute of Artificial Intelligence, Zhejiang University
Hangzhou, China
kunkuang@zju.edu.cn

Peng Cui
Tsinghua University
Beijing, China
cuip@tsinghua.edu.cn

Abstract

In the era of big data, decision-making in social networks may introduce bias due to interconnected individuals. For instance, in peer-to-peer loan platforms on the Web, considering an individual's attributes along with those of their interconnected neighbors, including sensitive attributes, is vital for loan approval or rejection downstream. Unfortunately, conventional fairness approaches often assume independent individuals, overlooking the impact of one person's sensitive attribute on others' decisions. To fill this gap, we introduce "Interference-aware Fairness" (IAF) by defining two forms of discrimination as Self-Fairness (SF) and Peer-Fairness (PF), leveraging advances in interference analysis within causal inference. Specifically, SF and PF causally capture and distinguish discrimination stemming from an individual's sensitive attributes (with fixed neighbors' sensitive attributes) and from neighbors' sensitive attributes (with fixed self's sensitive attributes), separately. Hence, a network-informed decision model is fair only when SF and PF are satisfied simultaneously, as interventions in individuals' sensitive attributes or those of their peers both yield equivalent outcomes. To achieve IAF, we develop a deep doubly robust framework to estimate and regularize SF and PF metrics for decision models. Extensive experiments on synthetic and real-world datasets validate our proposed concepts and methods.

CCS Concepts

• **Applied computing** → **Law, social and behavioral sciences**;
• **Computing methodologies** → **Machine learning**.

Keywords

Algorithmic fairness, machine learning, social network

ACM Reference Format:
Wenjing Yang, Haotian Wang, Haoxuan Li, Hao Zou, Ruochun Jin, Kun Kuang, and Peng Cui. 2024. Your Neighbor Matters: Towards Fair Decisions Under Networked Interference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671960

1 Introduction

The widespread use of machine learning (ML) models raises concerns about ethical and legal implications due to potential biases [2, 37]. For example, ML-based credit scoring in loan systems may yield discriminatory results for individuals with similar financial profiles but different races [19]. To ensure fairness, research has developed various fairness metrics [4, 9, 11, 12, 14, 19, 36, 38, 40, 42]. Earlier work focused on statistical independence between ML decisions and sensitive attributes [8, 11]. There has been much recent interest in answering the fairness questions from the perspective

*Wenjing Yang, Haotian Wang and Haoxuan Li contributed equally to this research.
†Haotian Wang is the corresponding author.
‡Kun Kuang is the corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671960

of causality [19, 26, 30], aiming to achieve equity by examining interventions on sensitive attributes, e.g., "How would the decision changes if we intervened with racial attributes?".

Previous fairness criteria, particularly those rooted in causality-based fairness metrics, have often assumed individual independence, i.e., one person's sensitive attributes do not impact others' decisions. This assumption implies that discrimination against individuals, such as the borrower in our P2P loan example, is solely influenced by their features, such as credit status and the sensitive attribute of living location[1]. However, in the era of big data, our living world is interconnected and social relationships play a significant role in decision-making models [7, 20, 31]. For instance, in contemporary online lending platforms, such as peer-to-peer (P2P) loan systems on the Web, decision-making increasingly relies on both an individual's attributes and those of their social network neighbors [20]. In essence, decisions made by many ML systems now incorporate sensitive attributes not only from the individual but also from their neighbors in the social network.

Consequently, when the effects of self-sensitive attributes (self-effect) and neighbors' sensitive attributes (peer-effect) intertwine, a decision model that appears fair under conventional criteria may exhibit unfairness. Keeping the P2P loan example in mind, a loan decision model satisfying conventional fairness criteria [8, 10, 11, 14, 19, 39] entails that the overall correlation/causal effect from sensitive attribute, i.e., location or race, to the loan decision vanishes. However, a loan model unfair for minor-group individuals might seem to be fair with a vanished effect from race to loan decision. The core reason is that scores assigned by the decision model for minor-group individuals will be enhanced by considering their social connections with the major-group neighbors. We term such kinds of unfairness as *interference-specific unfairness* throughout our paper. In general, interference-specific unfairness exists commonly across many scenarios beyond loan systems. For instance, prestigious colleges may admit applicants from the minor group (seemingly fair) with lots of neighbors from the major group, while rejecting applicants with similar academic qualifications from the minor group with few neighbors from the major group (indeed unfair) [3].

In response to such discrimination specific to interference across individuals, we advocate for the establishment of more robust fairness metrics that account for the influence of social relationships on decision-making. Building upon recent advancements in interference-based causality [13, 22, 27], we introduce the concepts of "**self-fairness (SF)**" and "**peer-fairness (PF)**" to causally evaluate equity among individuals with similar self-sensitive attributes and distinct neighbor-sensitive attributes, as well as equity among individuals with dissimilar self-sensitive attributes and analogous neighbor-sensitive attributes, respectively. Subsequently, the simultaneous satisfaction of SF and PF through model regularization can mitigate interference-specific unfairness.

To the best of our knowledge, we are pioneering the formal differentiation and mitigation of discrimination stemming from an individual's sensitive attributes versus those arising from peers' sensitive attributes. We summarize our contributions in below:

1. We contribute the **I**nterference-**a**ware **F**airness (IAF) metric to capture such unfair decisions.
2. To characterize IAF, we introduce causal definitions for Self Fairness (SF) and Peer Fairness (PF) aimed at capturing unfair decisions induced by peer effects.
3. Inspired by networked causal inference [27], we devise a deep doubly robust (DR) framework to regularize unfair decision models in the presence of interference.
4. Our experiments, conducted on one synthetic data and two real-world datasets, yield the following key results (a) SF and PF effectively capture unfair decisions stemming from peer effects, and (b) our designed DR framework successfully eliminates this interference-specific unfairness.

[1]As discussed in the real-world case study by [20].

2 Preliminaries

2.1 Notations

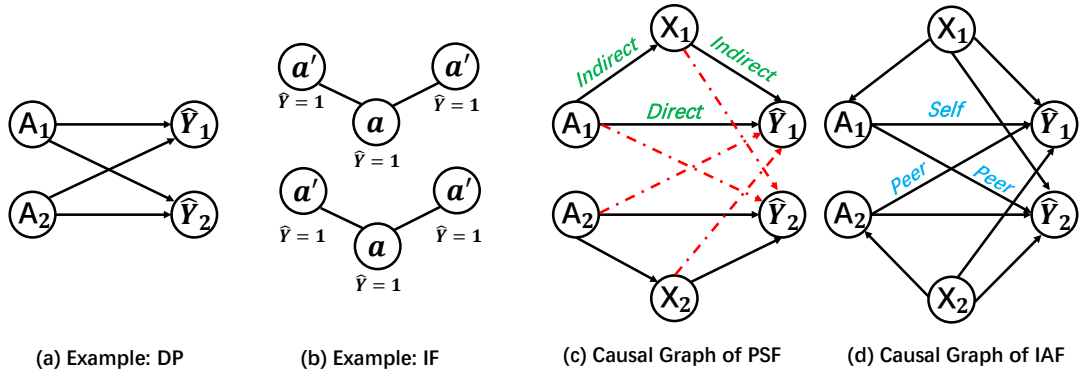
We formalize definitions of fairness in the essence of interference across decision subjects. Throughout this paper, uppercase letters denote the Random variables, e.g., X, and lowercase letters denote their realizations, e.g., x. Let $\{A_i, X_i, Y_i\}_{i=1}^n \sim P$ be the logged dataset with n individuals sampled from the joint distribution $P(A, X, Y)$, where A_i, X_i and Y_i are sensitive attributes, contextual features and outcome to be predicted for individual i, respectively. A decision model M_θ (θ is the parameter of M_θ) is learned on $\{A_i, X_i, Y_i\}_{i=1}^n$ with predictions $\hat{Y}_i$ for individual i. We assume A to be binary throughout this paper, while our discussion framework can be easily generalized to categorical sensitive attributes. We present all causal notions using the language of potential outcome framework [30], i.e., $\hat{Y}(a)$ represents the potential decision if the sensitive attribute A were set to value a. Notably, $(A, X, \hat{Y})$ is the observational data and cannot be intervented arbitrarily, as M_θ attempts to fit $P(A, X, Y)$. Besides, $[n]$ represents the set $\{1, 2, \cdots, n\}$ and $-i$ represents all elements in $[n]$ except for i.

2.2 Correlation-based Fairness

By accounting for the correlations between sensitive attribute A and outcome Y, several popular metrics have been proposed to achieve fairness in the correlation sense [8, 10, 11]. For instance, the Fairness Through Unawareness (FTU) [10] principle proposes to overlook the sensitive attribute A, while the Demographic Parity (DP) [8] criteria enforces M_θ to decisions $\hat{Y}$ independent from A: $A \perp\!\!\!\perp \hat{Y}$. Besides, two important notions, i.e., Equality of Opportunity (EO) [11] and individualized fairness (IF) [9], have generalized the DP metric on some sub-populations/individuals. To be specific, EO requires DP on individuals receiving positive decisions: $A \perp\!\!\!\perp \hat{Y} \mid \hat{Y} = 1$, while IF enforces individuals with similar contextual features to receive similar decisions: $D(\hat{Y}_i, \hat{Y}_j) \leq \epsilon D_X(X_i, X_j)$ (D is some metric and ϵ is some pre-defined threshold).

2.3 Causality-based fairness

While correlation-based fairness notions promote numerous approaches with compelling simplicity [11, 23], they may suffer from bias caused by confounders (explicit (X) or latent factors). We refer to extensive analysis on such phenomena to previous studies [19, 23]. With the aim of quantifying and migrating the causal effect of A on $\hat{Y}$ via controlling third factors, the causality-based fairness approaches have emerged in recent years [14, 19, 25, 26, 39, 43].



(a) Example: DP (b) Example: IF (c) Causal Graph of PSF (d) Causal Graph of IAF

Figure 1: Counterexamples and causal graphs we illustrated in Section 4.2. In (b), we merge the node A and Y for the clarity. In (c), direct (unfair) and indirect (fair) causal paths from A_i to $\hat{Y}_i$ are colored in green, while the red dashed lines represent that PSF do not consider causal effects across individuals. In (d), self and peer causal effects from A_i to $\hat{Y}_i, \hat{Y}_j$ are colored in blue.

The counterfactual parity and conditional counterfactual parity [25] has described population-level causality-based fairness by eliminating average and conditional treatment effect: $E[\hat{Y}(A = 1)] = E[\hat{Y}(A = 0)]$ and $E[\hat{Y}(A = 1) \mid X] = E[\hat{Y}(A = 0) \mid X]$. The principle fairness [14] and counterfactual equalized odds [24] have developed causality-based fairness specific to some important subpopulations. The unit-level causality-based fairness, i.e., the counterfactual fairness [19], has been proposed based on the Structural Causal Model (SCM) [29]: $\hat{Y}_i(A_i = 1) = \hat{Y}_i(A_i = 0)$. The path-specific causality-based fairness [4, 26, 39] has been developed to distinguish fair and unfair causal paths from A to the $\hat{Y}$ with different mediators. For instance, PSF requires that $E[\hat{Y}(A = 1, K(A = 0))] = E[\hat{Y}(A = 0, K(A = 0))]$, where K is the mediator lying in the causal path from A to $\hat{Y}$.

2.4 Fairness on Graphs

Recent research has extensively examined fairness in graph data across various tasks, such as node classification, link prediction, and community detection, through graph representation learning with Graph Neural Networks (GNNs) [1, 6, 7, 17, 28, 31]. Fairness concepts originally designed for tabular data, including DP, EO, IF, and sample perturbations, have been integrated into the context of fair node embedding and classification [7]. Notably, prior work in graph fairness has made significant progress, but none has explicitly addressed or formalized discrimination arising from peer effects.

3 A Real-world Case Study: Unfair Judgment in Peer-to-Peer Loan

Prior research has highlighted the potential for discrimination stemming from peer effects when incorporating social relationships into decision-making processes. As depicted in Fig.2, [20] constructed a real-world social network using the Prosper Loans Network Dataset, encompassing over 1,048,575 Peer-to-peer (P2P) loan data records. Leveraging this data, our analysis addresses two key objectives (a) the necessity of considering social relationships when designing decision-making models, and (b) the identification of fairness concerns arising from the specific modeling of social relationships. In this case study, we designate the Location variable[2] as the sensitive attribute. Therefore, the fairness inquiry centers on whether the loan decision model exhibits bias against borrowers from less favorable areas (see Appendix A for details on implementation.)

The first task is to design decision models to judge the individual's social score based on their social connections, rating, and status of corresponding loan records, etc. Meanwhile, the second task is to design decision models to judge the risk of each loan record based on features including the credit status, and loan rating records of the lender and borrower. For the first task, we design and compare two decision models, i.e., a Graph Neural Network (GNN) model and a multi-layer MLP model. For the second task, we follow [20] and compare the XGBoost classifiers trained with and without social features [20]. This metric quantifies the proportion of negative judgments on creditworthy individuals with connections in unfavorable locations and positive judgments on individuals who lack creditworthiness but have connections in favorable locations[3].

In Fig. 2, we observe a significant degradation in model prediction performance for both tasks in the absence of interference modeling. However, this improved performance comes at the cost of a considerable increase in Unfair Proportion. This suggests that the bias introduced by neighboring relationships is concurrently incorporated into the learning models. Hence, our case study demonstrates that interference across individuals is a double-edged sword, enhancing prediction performance while introducing discrimination stemming from peer effects.

4 Interference-aware Fairness

4.1 Problem Setting

In the context of interference-aware fairness, we introduce a formal problem. Consider a social network comprising n individuals, denoted as decision subjects, with an adjacency matrix $G \in \mathcal{R}^{n*n}$.

[2] A binary variable representing the location of each borrower.

[3] We explicitly encode the feature "creditworthy" for each individual by its social score s_i , i.e., the true label for the first task. $s_i = 2$ refers to creditworthy individuals and $s_i = 0$ refers to individuals who are not creditworthy. In similar, we explicitly encode the feature "favorable" by the value of the variable "Location". When Location= 1, i.e., more low-risk neighbors than high-risk borrowers, the location is "favorable".

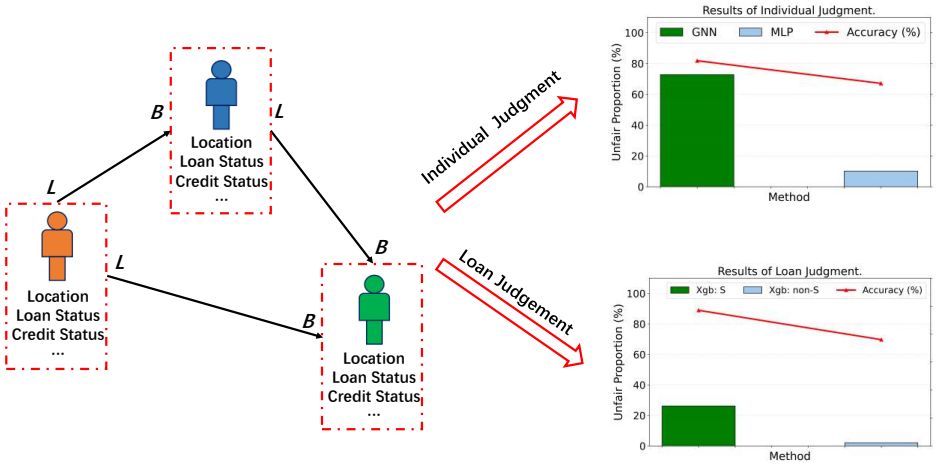


Figure 2: In the case study of P2P loans on the Web, we present our findings and outcomes. Specifically, on the left-hand side, 'L' and 'B' denote the lender and borrower in a loan record, while on the right-hand side, we refer to 'GNN' as a three-layer Graph Neural Network, 'MLP' as a three-layer fully connected network, 'Xgb:S' as the XGB boost classifier incorporating social features, and 'Xgb:nonS' as the XGB boost classifier excluding social features.

Here, $G_{ij} = 1$ signifies a causal relationship between individual A_i and $\hat{Y}_j$, as well as between A_j and $\hat{Y}_i$. Additionally, connections between individuals i and j imply causal effects from X_i to A_j and $\hat{Y}_j$, as well as from X_j to A_i and $\hat{Y}_i$. In summary, when individuals i and j are neighbors, their own covariates influence the sensitive attribute and decision of the other party, and their sensitive attribute also impacts the decision of the other party. See Fig. 1 (c) for an illustrative example of this setting. Consequently, the potential decision of individual i can be expressed as $\hat{Y}(A_i, X_i, A_{-i}, X_{-i})$ with $A_i(X_i, X_{-i})$. To address the data inefficiency challenge arising from the exponential realizations of the high-dimensional treatment vector (A_i, A_{-i}), we commonly adopt the following assumptions in the field of networked interference in prior work [22, 27].

Assumption 1 (Neighborhood interference). *The potential decision for i is only decided by its neighbor's sensitive attributes with covariates: $\hat{Y}(A_i, X_i, A_{-i}, X_{-i}) = \hat{Y}(A_i, X_i, A_{\mathcal{N}_i}, X_{\mathcal{N}_i})$, where $\mathcal{N}_i = \{j \mid G_{ij} = 1, j \neq i, j \in [n]\}$ represents the neighbors of individual i on the network. In similar, $A_i(X_i, X_{-i}) = A_i(X_i, X_{\mathcal{N}_i})$.*

The exposure mapping assumption is further applied to mitigate the exponential combination of $\mathcal{N}_i$ [22]:

Assumption 2 (Exposure Mapping). *There exists Φ which maps $\mathbf{A}_{\mathcal{N}_i}$ to a dense vector $\Phi(\mathbf{A}_{\mathcal{N}_i})$ for any $i \in [n]$ such that for any neighboring treatment vectors $A^1_{\mathcal{N}_i}$ and $A^2_{\mathcal{N}_i}$, we have $\hat{Y}(a, A^1_{\mathcal{N}_i}) = \hat{Y}(a, A^2_{\mathcal{N}_i})$ if $\Phi(A^1_{\mathcal{N}_i}) = \Phi(A^2_{\mathcal{N}_i})$, where the notations $A^1_{\mathcal{N}_i}$ and $A^2_{\mathcal{N}_i}$ are realization values of $A_{\mathcal{N}_i}$.*

Notably, the above assumption states that we can use a function Φ for summarizing the neighbor treatments from a high dimensional vector $A_{\mathcal{N}_i}$ to a dense vector $\Phi(A_{\mathcal{N}_i})$. We denote $\tilde{0}$ in the exposure set $\tilde{A}$ as the no-treatment regime of $\mathcal{N}_i$, serving as the control group for treatment effect without interference. Following prior work [21, 22, 27], we define g_X as the summary function capturing the influence of neighboring covariates, i.e., $A_i = f_A(X_i, g_X(X_{\mathcal{N}_i}), U^A_i)$, where f_A is an unspecified function, and U^A_i represents individual characteristics. Similarly, g_A denotes the summary functions for sensitive attributes, i.e., $\hat{Y}_i = f_Y(X_i, g_X(X_{\mathcal{N}_i}), A_i, g_A(A_{\mathcal{N}_i}), U^Y_i)$, where f_Y and U^Y_i are defined analogously. To ensure fairness in the presence of interference, we introduce the concept of Interference-Aware Fairness (IAF) to differentiate and mitigate discrimination arising from sensitive attributes and network neighbors:

Definition 1. A decision model M_θ satisfies **Self-Fairness (SF)**, i.e., eliminates the discrimination from self's sensitive attribute, if the population-level **direct (self) effect** of A_i on $\hat{Y}_i$ vanishes:

$$\frac{1}{n}\sum_{i\in[n]} E[\hat{Y}_i(A_i = 1, \tilde{a})] - E[\hat{Y}_i(A_i = 0, \tilde{a})] = 0, \quad \forall \tilde{a} \in \tilde{A}. \tag{1}$$

Meanwhile, M_θ satisfies **Peer-Fairness (PF)**, i.e., eliminates the discrimination from neighbor's sensitive attribute, if the population-level **peer effect** of $A_{\mathcal{N}_i}$ on $\hat{Y}_i$ vanishes:

$$\frac{1}{n}\sum_{i\in[n]} E[\hat{Y}_i(A_i = a, \tilde{a})] - E[\hat{Y}_i(A_i = a, \tilde{0})] = 0, \quad \forall a \in \{0, 1\}. \tag{2}$$

Remark. $\tilde{a}$ comes from Φ. By simultaneously satisfying SF and PF through the enforcement of decision model M_θ, we say M_θ satisfies IAF. It is worth noting that conceptualizing SF and PF at the individual lev (counterfactual) will produce much sharper fairness. However, the computational challenges of deriving individual-level IAF from observational data, stemming from the absence of an SCM model, render this approach practically infeasible. Meanwhile, our experiments empirically show that regularizing SF and PF already leads to near-optimal performance. We defer the exploration of individual-level IAF to future research (refer to the Conclusion for details).

Acronym	Full Name	Definition
DP	Demographic Parity	$A \perp \hat{Y}$
FTU	Fairness Through Unawareness	Do not include A to the prediction model
IF	Individualized Fairness	$D\left(\hat{Y}_i, \hat{Y}_j\right) \leq \epsilon D_X\left(X_i, X_j\right)$
EO	Equality of Opportunity	$A \perp \hat{Y} \mid Y = 1$
CP	Counterfactual Parity	$E[\hat{Y}(A=1)] = E[\hat{Y}(A=0)]$
CCP	Conditional Counterfactual Parity	$E[\hat{Y}(A=1)] = E[\hat{Y}(A=0) \mid X=x], \quad \forall x \in \mathcal{X}$
CF	Counterfactual Fairness	$\hat{Y}_i\left(A_i=1\right) = \hat{Y}_i\left(A_i=0\right), \quad \forall i \in [n]$
PSF	Path-specific Fairness	$E[\hat{Y}(A, K(A))] = E\left[\hat{Y}\left(A, K'(A)\right)\right], E[\hat{Y}(A, K(A))] = E\left[\hat{Y}\left(A', K(A)\right)\right]$
IACP	Interference-aware Counterfactual Parity	$\frac{1}{n}\sum_{i\in[n]} E\left[\hat{Y}_i\left(A_i=1, \tilde{a}\right) - \hat{Y}\left(A_i=0, 0\right)\right] = 0, \quad \forall \tilde{a} \in \tilde{A}$
IACCP	Interference-aware Conditional Counterfactual Parity	$\frac{1}{n}\sum_{i\in[n]} E\left[\hat{Y}_i\left(A_i=1, \tilde{a}\right) - \hat{Y}\left(A_i=0, 0\right) \mid X=x\right] = 0, \quad \forall x \in \mathcal{X}, \tilde{a} \in \tilde{A}$
IACF	Interference-aware Counterefactual Fairness	$\hat{Y}_i\left(A_i=1, \tilde{a}\right) = \hat{Y}\left(A_i=0, 0\right), \quad \forall i \in [n], \tilde{a} \in \tilde{A}$
SF, PF	Self-fairness, Peer-fairness	Def. 1

Table 1: Summary of various fairness notions with their capabilities, including whether well-defined, whether can identify discrimination stemmed from self-effect and peer effects, in our IAF setting.

4.2 Comparison to Other Fairness Notions

In our IAF framework, we assert that prior fairness concepts, such as correlation-based, causality-based, and graph fairness, fall short in attaining SF and PF.

Correlation-based Fairness We investigate the limitations of demographic parity (DP) [8], equal opportunity (EO) [11], and individual fairness (IF) [9] in achieving statistical fairness (SF) and personalized fairness (PF) through counterexamples. In the context of the P2P loan case, we modify it by introducing a sensitive attribute, the borrower's race (A_i), while keeping other factors constant. We consider empty contextual features, i.e., $X = \emptyset$, with the ability to generalize to non-empty X. In our scenario, two borrowers apply for a loan, and the decision $\hat{Y}_1, \hat{Y}_2$ depends on both A_1, A_2. We distinguish between minor and major race groups denoted as a and a'. We set $P(A_2 = a') = P(A_2 = a) = 0.5$, $P(A_1 = a \mid A_2 = a') = 0.99$, and $P(A_1 = a \mid A_2 = a) = 0.01$. We construct an unfair binary decision model M_θ with $P(\hat{Y} = 1 \mid A_1 = a, A_2 = a) = 0.1$, $P(\hat{Y} = 1 \mid A_1 = a', A_2 = a') = 0.9$, $P(\hat{Y} = 1 \mid A_1 = a', A_2 = a) = P(\hat{Y} = 1 \mid A_1 = a, A_2 = a') = 0.5$. Consequently, we calculate that $P(\hat{Y}_1 = 1 \mid A_1 = a') = 0.5 \approx 0.496 = P(\hat{Y}_1 = 1 \mid A_1 = a)$ and $P(\hat{Y}_2 = 1 \mid A_2 = a') = 0.504 \approx 0.505 = P(\hat{Y}_2 = 1 \mid A_2 = a)$ (see Appendix B for details). Therefore, both population-level and individual-level DP fail to identify unfairness in this setting. Similarly, EO, which is a finer-grained version of DP, also cannot detect unfairness in the presence of interference. In the context of the IF metric, consider the loan system example in Fig.1(b) with 6 individuals: 4 in the major group (a') and 2 in the minor group (a). The loan decision model initially exhibits unfairness, rejecting the minor group and approving the major group with one neighbor. However, when an applicant from the minor group has two neighbors from the major group, decisions become positive. Upon utilizing IF for decision analysis, all subjects (regardless of the same or different A) receive identical decisions. This remains true even when incorporating network structure as features through encoding techniques such as social score in [20] or motifs in [41].

Causality-based Fairness Previous causality-based fairness approaches, including CP, CCP, CF, and PSF, rely on the Stable Unit Treatment Value Assumption (SUTVA) [15], which assumes no interference between an individual's sensitive attribute and neighbors' decisions. To facilitate a fair comparison, we extend CP, CCP, and CF to address self-discrimination and peer discrimination within our proposed IAF problem, demonstrating their limitations:

Definition 2. A decision model M_θ satisfies Interference-aware CP (IACP), Interference-aware CCP (IACCP) and Interference-aware CF (IACF), if the following criteria are satisfied:

$$\begin{cases} \textbf{IACP}: \frac{1}{n}\sum_{i\in[n]} E[\hat{Y}_i(A_i=1,\tilde{a}) - \hat{Y}(A_i=0,\tilde{0})] = 0 \quad \forall \tilde{a} \in \tilde{A}. \\ \textbf{IACCP}: \frac{1}{n}\sum_{i\in[n]} E[\hat{Y}_i(A_i=1,\tilde{a}) - \hat{Y}(A_i=0,\tilde{0}) \mid X] = 0 \quad \forall \tilde{a} \in \tilde{A}. \\ \textbf{IACF}: \hat{Y}_i(A_i=1,\tilde{a}) = \hat{Y}(A_i=0,\tilde{0}) \quad \forall i \in [n], \tilde{a} \in \tilde{A}. \end{cases}$$

In light of the extended concepts mentioned earlier, it becomes evident that IACP, IACCP, and IACF fail to distinguish between SF and PF as defined in Def.1. When $\tilde{a}$ is held fixed in SF and $a = 0$ in PF, we observe that $IACP = PF + SF$, implying that a vanished IACP, representing the complete absence of the combined effects of A_i and $A_{\mathcal{N}_i}$, can be decomposed into a negative PF (direct effect of A_i) and a positive SF (peer effect of $A_{\mathcal{N}_i}$). Consequently, a decision model deemed fair under IACP may exhibit unfairness when viewed through the lenses of SF and PF. Similarly, decisions considered fair under IACCP and IACF may appear unfair from the SF and PF perspectives. In contrast, a fair decision model adhering to SF and PF criteria is guaranteed to satisfy IACP, as indicated by the decomposition mentioned above. Finally, we compare PSF and our IAF in Fig.1 (c) and (d). Specifically, PSF quantifies both direct and indirect causal effects for the same individual via different causal pathways from A_i to $\hat{Y}_i$, such as $A_i \rightarrow \hat{Y}_i$ and $A_i \rightarrow M_i \rightarrow \hat{Y}_i$, where some causal pathways are deemed fair while others are deemed unfair. For example, a loan rejection directly caused by an individual's race is considered unfair, while a rejection indirectly caused by inadequate education is considered fair. In contrast, our IAF focuses on distinguishing and identifying unfair decisions resulting from self-effect and peer-effects of sensitive attributes.

Graph Fairness. In our analysis, we find that graph fairness methods, including DP, EO, IF, and graph representation perturbations [7, 17, 28, 31], lack the capability to achieve IAF. These methods (a) typically apply fairness regularization inherited from tabular data directly to graph models and (b) fail to distinguish between discrimination originating from an individual and their

neighbors, leading to the same limitations we have discussed in the context of correlation-based fairness comparisons.

Remark. We leave a summary of the boundary capabilities of various fairness metric in Tab. 1. We observe the potential to extend PSF's scope to encompass a broader definition. This involves analyzing the impact of variable A through $\hat{Y}$ across various paths, both at the feature and individual levels.

Remark. We also provide a summary of different fairness notions and their boundary capabilities in our IAF problem in Tab. 1.

5 A Doubly Robust Debiasing Framework

To achieve IAF, we begin by estimating direct and peer effects from historical data, following Def.1. We then regularize the ML model M_θ to mitigate bias, drawing inspiration from the DragonNet paradigm [32]. We formulate a doubly robust (DR) framework, named **"IAF+DR"**, aligning with the optimization objective and influence curve of the average causal effect under interference [27]. Formally, we aim to constrain the causal effect of the joint variables A_i and $A_{\mathcal{N}_i}$ on $\hat{Y}$ across the population:

$$\phi(a^*, \tilde{a}^*) = \frac{1}{n} \sum_{i \in [n]} E[\hat{Y}_i(A_i = a^*, A_{\mathcal{N}_i} = \tilde{a}^*)], \tag{3}$$

where we use $a^*, \tilde{a}^*$ to highlight the interventional values of A_i and $A_{\mathcal{N}_i}$ and distinguish from observational values. We adapt the well-known overlap assumption and unconfounded assumption into our problem:

Assumption 3 (Overlap). *For all x in the support of X, for all $i \in [n]$ and for all $\tilde{a} \in \tilde{A}$, a in the support of A, we have:*

$$P\left(A_i = a, A_{\mathcal{N}_i} = \tilde{a} \mid g_X(x)\right) > 0$$

Assumption 4 (Unconfoundness). *All the variables affecting both the treatments and the potential outcome are observed:*

$$\hat{Y}_i(A_i, A_{\mathcal{N}_i}) \perp\!\!\!\perp A_i, A_{\mathcal{N}_i} \mid g_X(X)$$

We then have the following identification result:

Theorem 1. *The target estimand, i.e., $\phi(a^*, \tilde{a}^*)$ for intervened values $a^*, \tilde{a}^*$, can be identified as follows:*

$$\phi(a^*, \tilde{a}^*) = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbf{x}} E[\hat{Y}_i \mid a^*, \tilde{a}^*, g_X(x)] p_x(\mathbf{x}) dx.$$

Proof. We have the following derivation:

$$\begin{aligned}
E\left[\phi(a, \tilde{a})\right] &= \frac{1}{n} \sum_{i=1}^{n} E\left[\hat{Y}_i(a, \tilde{a})\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbf{x}} E\left[\hat{Y}_i(a, \tilde{a}) \mid X = x\right] p_X(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbf{x}} E\left[\hat{Y}_i(a, \tilde{a}) \mid a, \tilde{a}, X = x\right] p_X(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbf{x}} E\left[\hat{Y}_i \mid a, \tilde{a}, X = x\right] p_X(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbf{x}} E\left[\hat{Y}_i \mid a, \tilde{a}, g_X(X)\right] p_X(\mathbf{x}).
\end{aligned}$$

□

We begin by parameterizing the model for regression estimation, denoted as $E[\hat{Y}_i \mid a, \tilde{a}, g_X(x)]$, using a deep model, $\hat{Y}^{nn}$, which also serves as the decision model. To model the aggregation of $g_X(x)$ from neighboring elements, we opt for a GNN model within $\hat{Y}^{nn}$.

Prior research [32] has demonstrated the benefit of targeted regularization for causal inference without interference [27, 34]. To improve the robustness and finite-sample efficiency of estimating $\phi(a, \tilde{a})$, we incorporate targeted regularization with interference. Specifically, we introduce a scoring head $\hat{A}^{nn}$ to the GNN representation output for capturing the propensity score, denoted as $P\left(A_i = a, A_{\mathcal{N}_i} = \tilde{a} \mid g_X(x)\right)$. We then apply targeted regularization to ensure that the estimated ϕ and $(\hat{A}^{nn}, \hat{Y}^{nn})$ satisfy the estimation equation, i.e., $\varphi(Y, A, \tilde{A}, X; \hat{Y}^{nn}, \hat{A}^{nn}, \phi)$, defined as follows:

$$\hat{Y}^{nn}(a^*, \tilde{a}^*, g_X(x)) - \phi + \frac{\hat{A}^{nn}(a^*, \tilde{a}^* \mid g_X(x))}{\hat{A}^{nn}(a, \tilde{a} \mid g_X(x))} [\,] y_i - \hat{Y}^{nn}(a, \tilde{a}, g_X(x))], \tag{4}$$

By regulating that $\frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i, A_i, \tilde{A}_i, X_i; \hat{Y}_i^{nn}, \hat{A}_i^{nn}, \phi_i) = 0$, we designed targeted-regularized outcome estimation as follows:

$$\hat{Y}^{reg}(a^*, \tilde{a}^*, g_X(x)) = \hat{Y}^{nn}(a^*, \tilde{a}^*, g_X(x)) + \epsilon * \frac{\hat{A}^{nn}(a^*, \tilde{a}^* \mid g_X(x))}{\hat{A}^{nn}(a, \tilde{a} \mid g_X(x))} \tag{5}$$

Hence, the regularization objective is formularized as

$$\mathcal{L}_{reg} = \frac{1}{n} \sum_{i=1}^{n} \sum_{a^* \in A, \tilde{a}^* \in \tilde{A}} (Y_i - \hat{Y}_i^{reg}(a_i^*, \tilde{a}_i^*, g_X(x_i)))^2$$

and the estimated ϕ as $\phi^{reg}(a^*, \tilde{a}^*) = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i^{reg}(a^*, \tilde{a}^*, g_X(x))$. Based on the above identification of $\phi(a^*, \tilde{a}^*)$, we are now capable of designing fairness penalty loss, i.e., $\mathcal{L}_f$, to remove bias conveyed in the decision model $\hat{Y}^{nn}$:

$$\mathcal{L}_f = \sum_{a^* \in A} \mathrm{Var}_{\tilde{a}^* \in \tilde{A}}(\phi^{reg}(a^*, \tilde{a}^*)) + \sum_{\tilde{a}^* \in \tilde{A}} \mathrm{Var}_{a^* \in A}(\phi^{reg}(a^*, \tilde{a}^*)), \tag{6}$$

where $\mathrm{Var}_{\tilde{a}^* \in \tilde{A}}$ refers to the variance of $\phi^{reg}(a^*, \tilde{a}^*)$ with different $\tilde{a}$. This objective can lead to the disappearance of SF and PF, and attainment of IAF under optimal optimization conditions. Additionally, two standard objectives are employed for optimizing $\hat{Y}^{nn}$ and $\hat{A}^{nn}$: $\mathcal{L}_Y = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, \hat{Y}_i^{nn}(a_i^*, \tilde{a}_i^*, g_X(x_i)))^2$ and $\mathcal{L}_A = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(A_i, \hat{A}^{nn}(a, \tilde{a} \mid g_X(x_i)))$ ($\mathcal{L}$ refers to the classification loss, i.e., cross-entropy loss). To summarize, the overall objective of our debiasing framework can be formulated as follows:

$$\mathcal{L}_{sum} = \mathcal{L}_Y + \mathcal{L}_A + \alpha \mathcal{L}_f + \epsilon \mathcal{L}_{reg},$$

where α is the parameter to control the fairness penalty. One observation hints that minimizing targeted regularization term, e.g., $\mathcal{L}_{reg}$, force $(\phi^{reg}, \hat{A}^{nn}, \hat{Y}^{reg})$ to satisfy the estimating equation in (4):

$$0 = \partial_\epsilon (\mathcal{L}_{sum})|_\epsilon = \epsilon \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i, A_i, \tilde{A}_i, X_i; \hat{Y}_i^{reg}, \hat{A}_i^{nn}, \phi_i^{reg}).$$

Remark. As our method is formulated using the semi-parametric analysis, it naturally satisfies the doubly-robustness property. For ease to understand, we summarize the training & inference procedure as follows:

Step 1. Model training. Learn $\hat{Y}^{nn}$ as the regression model, g_X as the GNN model to capture the aggregation of neighbors'

covariates, and $\hat{A}^{nn}$ as the propensity model by minimizing the training loss:

$$L_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a^*\in A,\tilde{a}^*\in\tilde{A}}\left(Y_i - \hat{Y}_i^{\text{reg}}\left(a_i^*,\tilde{a}_i^*,g_X\left(x_i\right)\right)\right)^2,$$

where

$$\hat{Y}^{reg}\left(a^*,\tilde{a}^*,g_X(x)\mid a,\tilde{a}\right)=\hat{Y}^{nn}\left(a^*,\tilde{a}^*,g_X(x)\right)+\epsilon\frac{A^{nn}\left(a^*,\tilde{a}^*\mid g_X(x)\right)}{\hat{A}^{nn}\left(a,\tilde{a}\mid g_X(x)\right)};$$

Step 2. Counterfactual outcomes inference. For testing samples $\left\{\left(x_i^{te},a_i^{te},\tilde{a}_i^{te}\right)\mid i=n+1,\ldots,n+m\right\}$, the potential outcome $\hat{Y}^{te}\left(a^*,\tilde{a}^*,x^{te}\right)$ is estimated by:

$$\hat{Y}^{te}\left(a^*,\tilde{a}^*,x^{te}\right)=\hat{Y}^{nn}\left(a^*,\tilde{a}^*,g_X\left(x^{te}\right)\right)+\epsilon\frac{\hat{A}^{nn}\left(a^*,\tilde{a}^*\mid g_X\left(x^{te}\right)\right)}{\hat{A}^{nn}\left(a^{te},\tilde{a}^{te}\mid g_X\left(x^{te}\right)\right)}$$

Step 3. Calculate SF and PF. Then the SF for $\tilde{A}=\tilde{a}$ can be calculated as:

$$\text{SF}=\frac{1}{m}\sum_{i=n+1}^{n+m}\hat{Y}_i^{te}\left(A_i=1,\tilde{a}\right)-\frac{1}{m}\sum_{i=n+1}^{n+m}\hat{Y}_i^{te}\left(A_i=0,\tilde{a}\right).$$

We note that the PF can be calculated from a similar argument.

6 Experiments

In this section, we conduct extensive experiments to answer the following questions:

Q1. Does our PF and SF models effectively identify peer-induced unfair decisions in reality?

Q2. Does our IAF+DR framework mitigate bias while preserving high prediction accuracy?

Q3. Does debiasing with IAF impact other fairness metrics?

Q4. What is the impact of the fairness penalty on the trade-off between prediction accuracy and fairness?

Q5. Can our IAF+DR framework accurately estimate self-effects and peer-effects?

6.1 Dataset, Baselines, and Metric

Dataset. In this study, we perform experiments on three datasets: one synthetic and two real-world datasets. The synthetic dataset, created based on simulations from [5], models hiring decisions for physically demanding jobs. It comprises one binary sensitive attribute with three covariates and a binary outcome generated using a predefined SCM. Additionally, we evaluate our methods on two real-world datasets, namely, the NBA dataset and the Credit Default Dataset, which include constructed social networkss [1, 7].

Baselines. We compare our **IAF+DR** framework with the following baselines including (a) two *vanilla decision models* on graphs: (1) Graph Convolution Network (**GCN**) model [18] and (2) the Graph Attention Network (**GAT**) model [35]; (b) *Correlation-based fair GNN methods*: (3) **CrossWalk** [17] achieves fairness by biasing random walks to cross group boundaries, (4) **FairGNN** [7] achieves fairness by incorporating fairness regularization to ensure equitable treatment of different groups, (5) **NIFTY** [1] aims to improve counterfactual fairness and stability of node representations by sample perturbations; (6) **InFo_GNN** [16] adapts the individual fairness by considering the interconnectedness of nodes on the network, (7) **GEAR** [21] proposes graph augmentation by sample perturbations. (c) *Causality-based Fairness Methods*: In our study, we employ the GNN with CP regularization, denoted as **IACP** (adapted from Def. 2). We do not consider path-specific fairness or counterfactual fairness methods, as they rely on prior causal knowledge and are not well-suited for interference scenarios. Adapting these methods to our IAF setting would require independent research.

Metric. In prediction, we assess testing accuracy (ACC) using AUC for each method. Fairness is quantified following established protocols [1, 7] through reporting Demographic Parity (DP) and Equal Opportunity (EO) as: DP = $|P(\hat{Y}=1\mid A=1)-P(\hat{Y}=1\mid A=0)|$ and EO = $|P(\hat{Y}=1\mid Y=1,A=1)-P(\hat{Y}=1\mid Y=1,A=0)|$. Additionally, we estimate the proposed SF and PF metrics using IAF+DR on each dataset by reporting the direct effect and peer effect defined in Assumption. 1. However, recognizing that SF and PF lack ground truth in real-world data, we introduce a complementary metric for other baselines to ensure a fair comparison. We extend the concept of "unfair proportion" used in our case study by employing matching to mitigate confounding bias from covariates. We term this metric "Interference-aware Unfairness from Neighbors" (IUFN) as the quantity measuring the extent of SF and PF. To be spcific, we first define the variable to reflect the neighboring effect, i.e., N. We let $N_i=1$ if more than half of the individual's neighbors have positive sensitive attributes, and $N_i=0$ otherwise. Supposing that the test samples are $\{(X_i,Y_i,N_i,A_i)\}_{i=1}^n$, we then define the overall metric as the quantity measuring the extent of the unfairness arising from interference in the network as follows:

$$\begin{aligned}\text{IUFN}=&\frac{1}{n}\sum_{i\in B_{000}}\sum_{j\in B_{101}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)+\frac{1}{n}\sum_{i\in B_{000}}\sum_{j\in B_{011}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)\\&+\frac{1}{n}\sum_{i\in B_{010}}\sum_{j\in B_{111}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)+\frac{1}{n}\sum_{i\in B_{100}}\sum_{j\in B_{111}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)\\&-\frac{1}{n}\sum_{i\in B_{001}}\sum_{j\in B_{100}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)-\frac{1}{n}\sum_{i\in B_{001}}\sum_{j\in B_{010}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)\\&-\frac{1}{n}\sum_{i\in B_{011}}\sum_{j\in B_{110}}1\left(D\left(X_i,X_j\right)\le\epsilon\right)-\frac{1}{n}\sum_{i\in B_{101}}\sum_{j\in B_{110}}1\left(D\left(X_i,X_j\right)\le\epsilon\right),\end{aligned}$$

where $B_{lmk}=\{i\in[n]\mid A_i=l,N_i=m,Y_i=k\}$, 1 is the indicator function, and $D(X_i,X_j)$ is the L2 distance between X_i and X_j, and the first term is the proportion of units whose outcome would change from 0 to 1, if we keep $N_i=0$ the same but change A_i from 0 to 1. The rest terms follow a similar argument. We refer to such metric as "Interference-aware Unfairness from Neighbors" (**IUFN**), as IUFN is similar to the matching method [33] to account for causal peer effects.

Implementations. Our approach, IAF+DR, utilizes a 3-layer GCN as its core. Specifically, we employ GCN as the embedding model to capture interference across individuals with respect to covariates X. We assume a prior knowledge of the neighborhood exposure mapping, denoted as $A_{\mathcal{N}_i}\in\mathcal{R}^{|\mathcal{N}_i|}$, which transforms into $\tilde{A}\in\mathcal{R}$. Notably, while this assumption may seem restrictive for general tasks such as estimating causal effects with interference, we argue that it is justifiable for fairness-related tasks. In fairness tasks, historical data typically consists of past decision records or decisions made by previous decision-makers. It is reasonable for the current decision maker, who is training the model, to possess some prior knowledge of past decision-making rules. Therefore, following established protocols in causal inference [27, 34], we define the

Table 2: Comparisons of our proposed IAF+DR with the baselines on ACC, AUC, DP, EO, and IUFN. All experiments are repeated and averaged with 5 independent random seeds. The best performance is marked in bold.

Dataset	Metrics	GCN	GAT	CrossWalk	FairGNN	NIFTY	InFoRM_GNN	Gear	IACP	IAF+DR
Synthetic	ACC (%)	70.1±0.2	71.3±0.2	65.5±0.6	69.6±0.3	65.3±0.3	69.1±0.2	**74.5±0.1**	71.1±0.2	73.2±0.1
	AUC (%)	62.7±0.2	65.8±0.1	61.7±0.3	60.6±0.2	50.0±0.4	69.1±0.1	72.5±0.1	68.0±0.1	**73.1±0.2**
	DP	0.06±0.03	0.06±0.01	0.04±0.01	0.05±0.01	**0.03±0.01**	3.68±0.4	0.94±0.1	0.16±0.03	0.08±0.02
	EO	0.05±0.02	0.05±0.01	0.06±0.01	**0.03±0.01**	0.03±0.02	1.11±0.03	0.95±0.08	0.13±0.03	0.09±0.02
	IUFN (%)	17.5±0.3	19.6±0.4	16.3±0.6	18.3±0.6	18.5±0.2	19.5±0.3	18.5±0.4	18.0±0.2	**0.5±0.2**
NBA	ACC (%)	65.9±0.2	66.8±0.1	46.8±0.4	72.3±0.3	56.3±0.6	68.1±0.5	68.2±0.4	70.1±0.2	**72.3±0.1**
	AUC (%)	65.9±0.2	67.2±0.1	46.5±0.6	**72.3±0.4**	56.3±0.6	68.8±0.4	65.9±0.3	70.1±0.3	71.4±0.2
	DP	0.31±0.06	0.41±0.10	10.71±1.25	0.14±0.06	**0.04±0.01**	2.97±0.41	0.19±0.08	0.06±0.03	0.05±0.02
	EO	0.26±0.05	0.31±0.14	28.69±1.68	0.22±0.07	0.13±0.03	1.20±0.32	0.21±0.11	0.26±0.03	**0.03±0.01**
	IUFN (%)	16.60±0.2	22.20±0.4	21.1±0.3	11.1±0.3	17.1±0.4	12.2±0.2	18.8±0.2	10.0±0.1	**3.3±0.1**
Credit	ACC (%)	69.6±0.2	70.4±0.1	**72.7±0.6**	66.8±0.4	67.6±0.3	69.3±0.1	66.2±0.2	68.7±0.2	70.9±0.2
	AUC (%)	64.7±0.2	66.2±0.1	54.3±0.5	63.1±0.3	64.2±0.4	**68.1±0.1**	65.4±0.3	67.5±0.2	67.9±0.2
	DP	0.13±0.01	0.11±0.02	**0.03±0.01**	0.32±0.10	0.19±0.03	1.49±0.69	0.11±0.01	0.06±0.03	0.05±0.01
	EO	0.12±0.02	0.12±0.04	0.03±0.02	0.30±0.11	0.19±0.02	6.35±1.05	0.11±0.06	0.03±0.02	**0.03±0.01**
	IUFN (%)	5.8±0.2	6.7±0.2	3.6±0.7	4.0±0.6	4.0±0.7	3.8±0.1	4.6±0.2	5.2±0.2	**0.8±0.2**

neighborhood exposure mapping from $A_{\mathcal{N}_i}$ to $\tilde{A}$ as binary-valued: $\tilde{A} = 1$ when half of $A_{\mathcal{N}_i}$ is positive, and $\tilde{A} = 0$ otherwise (see Appendix D.2 for details on baselines).

6.2 Performance Comparison

SF and PF captures unfair decisions raised from peer effects (Q1). In Fig. 3a, 3b, and 3c, we present the IFUN metric on synthetic data and two real-world datasets prior to applying our IAF+DR debiasing method. As previously indicated, the IFUN metric estimates SF and PF by cross-referencing diverse groups of individuals. Statistical findings regarding IFUN, particularly with NBA and credit data, reveal that real-world data often exhibits bias attributable to self-effect or peer effects of sensitive attributes. For instance, over 20% of NBA data records exhibit nationality-based discrimination originating from either an individual's own nationality or that of their neighbors. This reaffirms our stance: decision-making in an interconnected world is susceptible to bias stemming from self or neighbor-related sensitive attributes, emphasizing the importance of identifying and mitigating these forms of discrimination.

Our IAF+DR migrates unfair decisions while maintains prediction performance (Q2). In Table 2, we present the mean and standard deviations of metrics across baseline models for three datasets. Our observations are as follows:

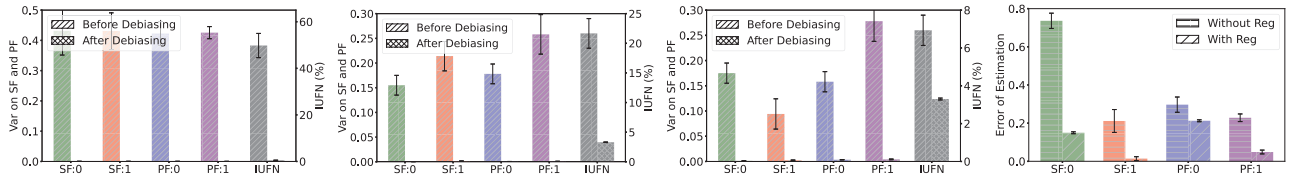
- Previous debiasing approaches, including CrossWalk, FairGNN, NIFTY, InFoRM_GNN, and Gear, achieve satisfactory prediction performance and reduce bias on standard fairness metrics like DP and EO. However, their performance on the interference-fairness metric (IUFN) indicates that they still retain discrimination stemming from neighboring relationships. Notably, the IUFN of CrossWalk and Gear exceeds that of GCN by nearly 20% on the NBA dataset.
- Our compared baseline, IACP, falls short in mitigating interference-aware discrimination across all datasets. These results confirm our theoretical analysis in Section 4.2, which asserts that the vanished total effect cannot be assumed to eliminate both self-effect and peer effects.
- In contrast, our proposed IAF+DR method effectively reduces the IUFN metric across all datasets when compared to other baselines, while maintaining competitive prediction accuracy.

Debias on SF and PF will not conflict with conventional fairness metric (Q3). As depicted in Table 2, our IAF+DR approach effectively mitigates bias in IUFN while maintaining low bias in DP and DP across various datasets. These results support our decomposition, TF=SF+PF, where DP and EO represent correlated versions of TF for the overall population and specific sub-populations. We contend that addressing bias in SF and PF aligns with established fairness criteria.

6.3 In-depth Analysis

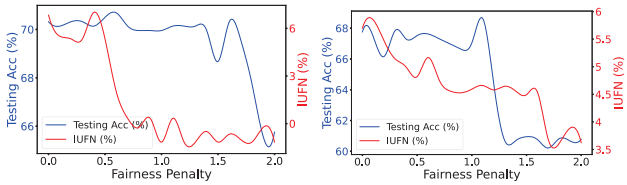
Impact of fairness penalty to prediction and fairness (Q4). In Fig.4, we observe that as the penalty parameter (α) increases, the test accuracy of our IAF+DR stabilizes in the range $[0.0, 1.5]$ while the IUFN metric sharply declines for $\alpha > 1.25$. As α approaches 2.0, both ACC and IUFN reach and maintain lower values. This phenomenon highlights the successful debiasing capability of our proposed IAF+DR while maintaining high prediction accuracy within the fairness penalty range of $\alpha \in [1, 1.5]$. It's worth noting that the reduction in prediction performance with increasing fairness penalty is a common occurrence due to historical decision data bias, as discussed in Q1. The fairness task seeks to strike a balance between acceptable prediction accuracy and minimizing decision bias.

Targeted regularization is crucial for estimations of SF and PF (Q5). We evaluated SF and PF estimation with and without our targeted regularization term $\mathcal{L}_{reg}$ (see Fig. 3d). Specifically, "without $\mathcal{L}_{reg}$" denotes the optimization of only $\mathcal{L}Y$ and $\mathcal{L}A$, with decision outcomes derived from $\hat{Y}^{nn}$ rather than $\hat{Y}^{reg}$ Since only synthetic data possesses ground truth for SF and PF, measuring estimation error on NBA and Credit is infeasible, necessitating the introduction of the IUFN metric. Notably, the absence of $\mathcal{L}_{reg}$ (left bars) results in a significant increase in estimation error compared to its presence (right bars).



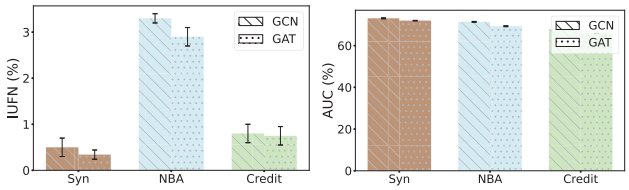
(a) Debiasing on Synthetic (b) Debiasing on NBA (c) Debiasing on Credit (d) Estimation on Synthetic

Figure 3: Results on: (a,b,c): Debiasing performance of SF and PF across datasets; (d) Estimation performance on SF and PF for synthetic data. All experiments are repeated and averaged with 5 independent random seeds.



(a) Trade-off on NBA Data (b) Trade-off on Credit Data

Figure 4: Trade-off between fairness and prediction accuracy on NBA and Credit data by tuning the fairness penalty α ranging in $[0.0, 0.01, 0.02, \ldots, 2.00]$.



(a) AUC (b) IUFN

Figure 5: Impact of substitution on the backbone across three datasets. All experiments are repeated and averaged with 5 independent random seeds.

Impact of different backbones. Questions may arise regarding the criticality of this choice and the adaptability of our proposed IAF+DR framework to various backbones. To address these concerns, we present a comparative analysis of AUC and IUFN results for our IAF+DR model across three datasets using both GCN and GAT as backbones (see Fig. 5). Our results demonstrate that the choice of backbone, whether GCN or GAT, does not significantly impact the prediction performance or debiasing capabilities of our IAF+DR framework.

Impact of parameter ϵ. In Fig.6, we observe that as ϵ increases, the debias capability of our IAF+DR, i.e., IUFN, and the estimation error decreases sharply. When $\epsilon \geq 1$, all metrics are stabilized. Such phenomenon shows that the DR regularization effectively estimates SF and PF, and efficiently migrates bias.

7 Conclusion

In this paper, we introduce Interference-Aware Fairness (IAF), a novel concept addressing discrimination within interconnected individuals on social networks. We establish that achieving IAF is tantamount to achieving two proposed fairness metrics: Self-Fairness (SF) and Peer-Fairness (PF). Consequently, we present a doubly robust framework for end-to-end estimation and mitigation of SF and PF in model decisions.

We propose several avenues for future research. Firstly, an extension of IAF to individual-level fairness, specifically addressing counterfactual causal inference, warrants attention. While individual-level IAF offers sharper fairness compared to population-level IAF, its identification in the context of SF and PF (individual-level) remains an open challenge. An alternative approach gaining popularity involves establishing a computationally feasible upper bound for SF and PF based on observational data.

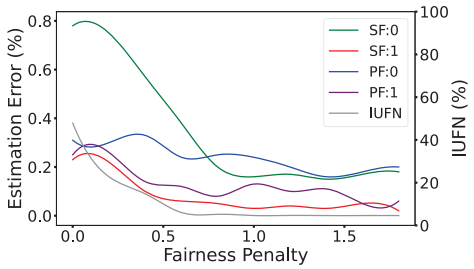


Figure 6: Impact of the parameter for our DR framework. i.e., ϵ, on estimation and fairness. Results are reported by tuning ϵ ranging in $[0.0, 0.01, 0.02, \ldots, 2.00]$.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (623B2002,62141607,62376243,62441605,62302503,KY0402052402), Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), and NUDT Youth Independent Innovation Science Fund Project (Grant No. ZK23-15).

References

[1] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*. PMLR, 2114–2124.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. (2016).

[3] Hal Berghel. 2020. A Critical Look at the 2019 College Admissions Scandal? *Computer* 53, 1 (2020), 72–77.

[4] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7801–7808.

[5] Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. 2021. Learning individually fair classifier with path-specific causal-effect constraint. In *International conference on artificial intelligence and statistics*. PMLR, 145–153.

[6] Manvi Choudhary, Charlotte Laclau, and Christine Largeron. 2022. A survey on fairness for machine learning on graphs. *arXiv preprint arXiv:2205.05396* (2022).

[7] Enyan Dai and Suhang Wang. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 680–688.

[8] Richard B Darlington. 1971. Another look at "cultural fairness" 1. *Journal of educational measurement* 8, 2 (1971), 71–82.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[10] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. Barcelona, Spain, 11.

[11] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[12] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fair multiple decision making through soft interventions. *Advances in Neural Information Processing Systems* 33 (2020), 17965–17975.

[13] Michael G Hudgens and M Elizabeth Halloran. 2008. Toward causal inference with interference. *J. Amer. Statist. Assoc.* 103, 482 (2008), 832–842.

[14] Kosuke Imai and Zhichao Jiang. 2023. Principal fairness for human and algorithmic decision-making. *Statist. Sci.* 38, 2 (2023), 317–328.

[15] Guido W Imbens and Donald B Rubin. 2010. Rubin causal model. In *Microeconometrics*. Springer, 229–241.

[16] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. 2020. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 379–389.

[17] Ahmad Khajehnejad, Moein Khajehnejad, Mahmoudreza Babaei, Krishna P Gummadi, Adrian Weller, and Baharan Mirzasoleiman. 2022. Crosswalk: Fairness-enhanced node representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11963–11970.

[18] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[19] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[20] Yanying Li, Yue Ning, Rong Liu, Ying Wu, and Wendy Hui Wang. 2020. Fairness of classification using users' social relationships in online peer-to-peer lending. In *Companion Proceedings of the Web Conference 2020*. 733–742.

[21] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 695–703.

[22] Yunpu Ma and Volker Tresp. 2021. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3700–3708.

[23] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553* (2020).

[24] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 386–400.

[25] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[26] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. 2022. Causal inference for social network data. *J. Amer. Statist. Assoc.* (2022), 1–15.

[28] John Palowitch and Bryan Perozzi. 2019. Monet: Debiasing graph embeddings via the metadata-orthogonal training unit. *arXiv preprint arXiv:1909.11793* (2019).

[29] Judea Pearl. 2009. *Causality*. Cambridge university press.

[30] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[31] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. 2022. Fairsna: Algorithmic fairness in social network analysis. *arXiv preprint arXiv:2209.01678* (2022).

[32] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* 32 (2019).

[33] Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

[34] Mark J Van der Laan. 2014. Causal inference for a population of causally connected units. *Journal of Causal Inference* 2, 1 (2014), 13–74.

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[36] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2536–2544.

[37] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*.

[38] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*. 3356–3362.

[39] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in neural information processing systems* 32 (2019).

[40] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[41] Yuan Yuan, Kristen Altenburger, and Farshad Kooti. 2021. Causal network motifs: identifying heterogeneous spillover effects in A/B tests. In *Proceedings of the Web Conference 2021*. 3359–3370.

[42] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509* (2016).

[43] Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. 2022. Counterfactual fairness with partially known causal graph. *Advances in Neural Information Processing Systems* 35 (2022), 1238–1252.

A Details on Case Study

The Prosper dataset contains 1,048,575 loan records that occurred from November 2005 to September 2011. Each record contains nine features, including the Lender ID, Borrower ID, Timestamp, Amount, Status (a binary variable representing the risk of each borrower), Lender rate, Borrower rate, and the Rating. We refer readers to [20] for a more detailed introduction to each feature. The social network is naturally built for each individual by adding undirected edges among each borrower and lender [20]. However, we only concerned with the borrower network for individual judgment task, i.e., the network consists of individuals who has become a borrower more than once.

The first task, i.e. the individual social score judgment task, is performed as a prediction task on the borrower network, where the decision outcome is the social score for each borrower. In our case study, such variable is interpreted as the "Location" for each borrower. Of course, a loan decision should be discriminative on the locations of borrowers. Considering the social network built by loan relationships, this can be regarded as a node classification task. Especially, the original social score is set as follows:

$$s_i = \frac{1}{1+\left(\frac{\lambda}{1-\lambda}\right)^{g_i}\left(\frac{\lambda p+(1-\lambda)}{\lambda+(1-\lambda)p}\right)^{L_i}\left(\frac{\lambda+(1-\lambda)p}{\lambda p+(1-\lambda)}\right)^{H_i}}, \tag{7}$$

where H_i and L_i account for the number of high-risk and low-risk neighboring borrowers, and $g_i = \mathcal{I}(H_i > L_i)$. We set hyper-parameter $\lambda = 0.46$ and $p = 0.5$ following original implementations in [20]. Based on the value of $s_i \in [0, 1]$, we create discrete labels in $[0, 1, 2]$ by setting s_i falling in $[0.0, 0.33)$, $[0.33, 0.66)$, and $[0.66, 1.0]$, respectively. We then construct two models, i.e., one GNN and one MLP, for prediction[4]. The GNN model contains three Graph Convolutional Layers (GCL) with *relu* activation functions. The MLP model contains three fully connected layers with *relu* activation functions. All the latent representations are set to 64 neurons. We optimize both GNN and MLP using the Adam optimizer with an initial learning rate 0.01 and weight-decay as $5e-4$. Both the accuracy and unfair proportions metric are reported on the testing data, where we split the whole data in the ratio 8 : 2.

The second task, i.e. the loan risk judgment task, is performed as a prediction task on the overall loan records, where the decision outcome is the loan status, i.e., the risk of each loan. Following [20], we use the XGBoost classifier in Sklearn https://scikit-learn.org/stable/ to perform prediction. We here set the s_i into binary variables, where $s_i < 0.5$ corresponds to label 0 and $s_i >= 0.5$ corresponds to label 1. To test whether the accuracy and unfair proportion will change with and without modeling the network structure, we construct two XGBoost predictors with and without the social score as features, i.e., the XGB:S and XGB:non-S. Notably, we interpret the social feature here as a signal of the living/working location of each borrower, as the calculation of s_i relies on s_i's neighbors. Both the accuracy and unfair proportions metric are reported on the testing data, where we split the whole data in the ratio 8 : 2.

The unfair proportion we reported in the case study roughly characterizes interference-specific unfairness, which reports the proportion of individuals who receive unfair decisions caused by their neighboring relationships. For the first task, individuals reported by unfair proportion can be divided into two types: (1) individuals with low observed signals (0) and powerful connections (ground truth of social score $s == 2$) receive negative credit judgment (predicted $s == 0$); (2) individuals with high observed signals (1) and weak connections (ground truth of social score $s == 0$) receives negative credit judgment (predicted $s == 2$). Similarly, for the second task, the unfair proportion accounts for two types of samples: (1) loan record with low loan status (ground truth of low-risk borrower) and powerful connections (social score of the borrower $s > 0.9$) receive negative evaluation (high risk); (2) loan record with high loan status (ground truth of high-risk borrower) and weak connections (social score of the borrower $s < 0.1$) receive positive evaluation (low risk).

[4]We note that all GNN-related models are constructed by the Pytorch-geometric package in https://pytorch-geometric.readthedocs.io/en/latest/

B Details on Comparison between our IAF and Conventional Fairness Notions

We first detail the constructed counterexample to distinguish our IAF from DP. To be more intuitive, we present the unfair decision model, i.e., Y, as follows:

A_1	A_2	$P(\hat{Y}=1 \mid A_1, A_2)$
a	a	0.1
a'	a'	0.9
a	a'	0.5
a'	a	0.5

(8)

, together with $P(A_1 = a \mid A_2 = a') = 0.99$, $P(A_1 = a \mid A_2 = a) = 0.01$ and $P(A_2 = a') = 0.5 = P(A_2 = a)$, then we first derive the DP for individual 2 as follows:

$$\begin{aligned}
&P\left(\hat{Y}_2 = 1 \mid A_2 = a'\right) \\
&= \sum_{a_1 \in \{a,a\}} P\left(\hat{Y}_2 = 1 \mid A_1 = a_1, A_2 = a'\right) P\left(A_1 = a_1 \mid A_2 = a'\right) = 0.5 \\
&P\left(\hat{Y}_2 = 1 \mid A_2 = a\right) \\
&= \sum_{a_1 \in \{a,a\}} P\left(\hat{Y}_2 = 1 \mid A_1 = a_1, A_2 = a\right) P\left(A_1 = a_1 \mid A_2 = a\right) = 0.496.
\end{aligned}$$

We then derive the posterior probabilities as follows:

$$\begin{aligned}
&P\left(A_1 = a'\right) \\
&= \sum_{a_2 \in \{a,a\}} P\left(A_1 = a' \mid A_2 = a_2\right) P\left(A_2 = a_2\right) = 0.5 \\
&P\left(A_1 = a\right) \\
&= \sum_{a_2 \in \{a,a\}} P\left(A_1 = a \mid A_2 = a_2\right) P\left(A_2 = a_2\right) = 0.5,
\end{aligned}$$

and we have:

$$\begin{aligned}
P\left(A_2 = a \mid A_1 = a'\right) &= \frac{P\left(A_1 = a' \mid A_2 = a\right) P\left(A_2 = a\right)}{P\left(A_1 = a'\right)} = 0.99 \\
P\left(A_2 = a \mid A_1 = a\right) &= \frac{P\left(A_1 = a \mid A_2 = a\right) P\left(A_2 = a\right)}{P\left(A_1 = a\right)} = 0.01.
\end{aligned}$$

Hence, we derive similar results for individual 2 as follows:

$$\begin{aligned}
P\left(\hat{Y}_1 = 1 \mid A_1 = a'\right) &= 0.504 \\
P\left(\hat{Y}_1 = 1 \mid A_1 = a\right) &= 0.505
\end{aligned}$$

Algorithm 1 Training procedure of Our DR Framework

Input The dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n$ with the adjacent matrix G, the GNN-based feature encoder V, the neighborhood exposure mapping Φ, the hyper-parameter ϵ.
for itr = 1 to $\mathcal{I}$ **do**
  Embedding covariates X by $\text{Emb}_X = V(X, G)$;
  Computing the neighbor exposure mapping for each node as $\tilde{A} = \Phi(A, G)$;
  Compute (observed) hashed value $A^{sum} = \Phi(A, \tilde{A})$;
  Compute the prediction head $\hat{Y}^{nn}(Emb_X, A^{sum})$;
  Compute the score head $\hat{A}^{nn}(A^{sum} \mid Emb_X)$;
  $\mathcal{L}_{reg} = 0$;
  for All possible values $(a^*, \tilde{a}^*)$ in $A \times \tilde{A}$: **do**
    Compute hashed value $A^{*,sum} = \Phi(a^*, \tilde{a}^*)$;
    Compute $\hat{Y}^{reg}(A^{*,sum}, Emb_X)$;
    Compute $\hat{A}^{nn}(A^{*,sum} \mid Emb_X)$;
    Compute $\hat{Y}^{reg}(A^{*,sum}, Emb_X)$ based on above terms;
    Accumulate $\mathcal{L}_{reg}$;
  end for
  for All possible values $(a^*, \tilde{a}^*)$ in $A \times \tilde{A}$: **do**
    Compute $\phi^{reg}(a^*, \tilde{a}^*)$;
  end for
  Compute $\mathcal{L}_f$ based on $\{\phi^{reg}(a^*, \tilde{a}^*)\}_{a^* \in A, \tilde{a}^* \in \tilde{A}}$;
  Compute $\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(Y_i, \hat{Y}_i^{nn}(A^{sum}, g_X(x_i)))^2$;
  Compute $\mathcal{L}_A = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(A_i, \hat{A}^{nn}(A^{sum} \mid g_X(x_i)))$;
  Update overall objective $\mathcal{L}_{sum}$;
end for
Output Using $\hat{Y}^{nn}$ as the decision model.

C Algorithm Procedure

We detail the procedure of our proposed IAF+DR in Alg. 1.

D Experimental Details

D.1 Dataset Details

Synthetic Dataset Description Following [5], we construct the synthetic data containing the gender $A \in \{0, 1\}$, qualification Q, number of children D, physical strength M_θ, and hiring decision outcome $Y \in \{0, 1\}$ following a pre-defined SCM. We first randomly sample the social network by generating the adjacent matrix G: $G_{ij} \sim \text{Bernoulli}(0.5)$. We then generate each feature attribute as follows:

$$\begin{aligned}
&Q = \lfloor U_Q \rfloor, \quad Q \in \mathcal{R}^n \quad U_Q \sim \mathcal{N}\left(2, 5^2\right), \\
&D = A + \lfloor 0.5 Q U_D \rfloor, \quad D \in \mathcal{R}^n \quad , U_D \sim \text{Tr}\,\mathcal{N}\left(2, 1^2, 0.1, 3.0\right), \\
&M = 3A + 0.4 Q U_M, \quad M \in \mathcal{R}^n, \quad U_M \sim \text{Tr}\,\mathcal{N}\left(3, 2^2, 0.1, 3.0\right), \\
&A \sim \text{Bernoulli}\left(\text{expit}(G^T Q - 40)\right), \quad A \in \mathcal{R}^n,
\end{aligned} \tag{9}$$

where expit refers to the sigmoid function. We then perform exposure mapping for $A_{\mathcal{N}_i}$ for each $in[n]$ and thus the mapping of A:

$$\tilde{A}_i = \mathcal{I}(G^T A > 0.5 * \text{mean}(G^T A)),$$

where mean refers to the mean-value of the vector $G^T A$, and $\mathcal{I}$ refers to the indicator function. Intuitively, our exposure mapping here implies that $\tilde{A}_i = 1$ if the number of positive neighbors of i is larger than the mean level of the overall population. Furthermore, we generated hashed A from $\tilde{A}_i$ and A_i:

$$A^{hash} = \begin{cases} 0, \text{if } \tilde{A}_i = 0 \& A = 1 \\ 1, \text{if } \tilde{A}_i = 1 \& A = 0 \text{ or } \tilde{A}_i = 0 \& A = 1 \\ 2, \text{if } \tilde{A}_i = 1 \& A = 1 \end{cases} \tag{10}$$

Hence, Y are generated as follows:

$$Y \sim \text{Bernoulli}(\text{expit}(-20 + 10A^{hash} + 0.1 * (G^T Q + G^T D + G^T M)))$$

NBA dataset This is extended from a Kaggle dataset 1 containing around 400 NBA basketball players. The performance statistics of players in the 2016-2017 season and other various information e.g., nationality, age, and salary are provided. The social network is constructed by collecting the relationships of the NBA basketball players on Twitter with its official crawling API 2. The sensitive, i.e., the nationality, is binarized into two categories, i.e., U.S. players and overseas players. The classification, in task is to predict whether the salary of the player is over the median [7].

Credit Default Dataset Credit Default Dataset is comprised of 30,000 nodes, where each node represents individuals who are utilizing some form of credit. Each node contains 13 attributes. Individuals are connected by their spending and payment behavior. The classification task is to determine whether an individual will default on the credit card payment. Age is used as the sensitive attribute [1].

D.2 Implementation Details

To unify backbone embedding model across all baselines, we set the GNN model with three GCN layers throughout our experiments (except for the GAT baseline). The activation function is the *relu* function, and the optimizer is set as the Adam optimizer with initial learning rate 0.001 and weight_decay as $5e-4$. The training epoch for all baselines is set to 300. We run experiments on all baselines using the open-source project, i.e., the pygdebias package in https://github.com/yushundong/PyGDebias. For the NIFTY [1], we set the penalty parameter as 0.6 following original implementations. For fairGNN method [7], we set $\alpha = 100$ and $\beta = 1$ following their optimal setting. For CrossWalk in [17], the random walk length is set to $d = 5$, and the number of walks is set to $r = 500$, and their $\alpha = 0.5$ and $p = 2$. For InfoRM_GNN in [16], we set $\alpha = 10$ and η across all datasets by combining their original suggestions and our cross-fold validation results. For GEAR [21], we follow their original parameter setting, which states that $\lambda = 0.6, C = 2, \lambda_s = 0.4, \beta = 10, \mu = 1e-5, k = 20, B = 4$. For our IAF+DR method, we map $\tilde{A}$ and A to a scalar value by simply operations as $A^{hash} = \tilde{A} + 2 * A$, as we hold the belief that individuals with low A and low $\tilde{A}$ will receives the worst decisions, individuals with high A and high $\tilde{A}$ will receive best decisions, and individuals with low A and high $\tilde{A}$ or high A but low $\tilde{A}$ will lie between. Besides, we set $\alpha = 1$ with $\epsilon = 0.1$ throughout our experiments.