ON DYADIC FAIRNESS: EXPLORING AND MITIGATING BIAS IN GRAPH CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Disparate impact has raised serious concerns in machine learning applications and its societal impacts. In response to the need of mitigating discrimination, fairness has been regarded as a crucial property in algorithmic design. In this work, we study the problem of disparate impact on graph-structured data. Specifically, we focus on dyadic fairness, which articulates a fairness concept that a predictive relationship between two instances should be independent of the sensitive attributes. Based on this, we theoretically relate the graph connections to dyadic fairness on link predictive scores in learning graph neural networks, and reveal that regulating weights on existing edges in a graph contributes to dyadic fairness conditionally. Subsequently, we propose our algorithm, **FairAdj**, to empirically learn a fair adjacency matrix with proper graph structural constraints for fair link prediction, and in the meanwhile preserve predictive accuracy as much as possible. Empirical validation demonstrates that our method delivers effective dyadic fairness in terms of various statistics, and at the same time enjoys a favorable fairness-utility tradeoff.

1 INTRODUCTION

The scale of graph-structured data has grown explosively across disciplines (e.g., social networks, telecommunication networks, and citation networks), calling for robust computational techniques to model, discover, and extract complex structural patterns hidden in big graph data. Research work has been proposed for inference learning on potential connections (Liben-Nowell & Kleinberg, 2007), and corresponding algorithms can be used for high-quality link prediction and recommendations (Adamic & Adar, 2003; Sarwar et al., 2001; Qi et al., 2006). In this work, we study the fairness issues in the prediction of dyadic relationships between two instances within a homogeneous graph.

Despite the practicality of link prediction algorithms, serious concerns raised by disparate impact (Angwin et al., 2016; Barocas & Selbst, 2016) should also be reckoned with by algorithm designers. The disparate impact describes the unfairness in influential decisions which essentially derives from the characteristics protected by anti-discrimination laws or social norms. Unfortunately, this negative impact derived from biased data and conventional algorithms occurs in many applications including link prediction. One example is that a user recommender system follows the proximity principle (individuals are more likely to interact with similar individuals) or existing connections with intrinsic bias. Such an operating mode would deliver a bias recommendation dominated by sensitive attributes, like users with the same religion or ethnicity are more likely to be recommended to a user, and consequently generate segregation in social relations by long-term accumulation (Hofstra et al., 2017). Another example can be noticed in news streaming. When a news app has collected the political profile from a user, in pursuit of the user preference in news streaming, the system might only deliver politicking that a user is predisposed to agree with, therefore skews a user's scope and narrow the view by selectively displaying reality (Pariser, 2011). To alleviate these concerns, an algorithm is supposed to be fair and call for conducting a link prediction without biased by the sensitive attribute of two instances, therefore stream diverse and preferred recommendations.

Motivated by the potential bias in real cases, we advance dyadic fairness for the link prediction problem in a homogeneous graph, where the dyadic fairness expects the predictive relationships to be executed independently of vertex sensitive attributes. In this work, we focus our scope on Graph Neural Networks (GNNs), which show remarkable capacity in graph representation learning by message passing following graph connections. Within the pipeline of GNNs, given an arbitrary

graph, we theoretically analyze the relationship between dyadic fairness and the graph connections. Our findings suggest adapting weights on existing edges in a graph can contribute to dyadic fairness conditionally. Continuing with our theoretical findings, we propose **FairAdj**, an algorithm to empirically learn a fair adjacency matrix by updating the normalized adjacency matrix while keeping the original graph structure unchanged. Integrating with a utility objective function, the proposed algorithm seeks supplied dyadic fairness and link predictive utility simultaneously.

For quantification of fairness, we extend statistical metrics in group fairness (Dwork et al., 2012; Kusner et al., 2017) to dyadic fairness. We form instances into sensitive groups, and the dyadic fairness requests standard statistics such as positive outcomes or false positive rate on link score to be approximately equalized across intra and inter sensitive groups. Thus, achieving the fairness can consequently close the gap on the number of desirable dyadic predictions with or without the same sensitive attribute, and mitigate social segregation if more interactions exist across sensitive groups.

We present empirical studies on six real-world social and citation networks. We conduct evaluation towards seven measurements of both utility and dyadic fairness, and improvements are observed from two aspects compared to other baseline methods (Kipf & Welling, 2016b; Grover & Leskovec, 2016; Rahman et al., 2019; Bose & Hamilton, 2019). Dyadic fairness metrics verify that our method can minimize the statistical gap between the predictions of intra and inter links. For utility, our results are consistent with the foregone consensus (Zhao & Gordon, 2019; Fish et al., 2016; Calders et al., 2009), that satisfying fairness is associated with the sacrifice in utility. However, concurrently, our algorithm enjoys a more favorable fairness-utility tradeoff (same in fairness but less sacrifice in utility, and vice versa) compare to previous work. Additionally, to approach the real application cases, we also showcase a direct product that comes from dyadic fairness: our method can effectively stream more diverse recommendations containing instances holding different kinds of sensitive attributes.

2 Related Work

We review the conceptually related work in fair machine learning and technically related work in graph representation learning, and introduce several studies addressing fair graph embedding.

Fair Machine Learning. Various types of fairness notions have been advanced, including group fairness (Dwork et al., 2012; Kusner et al., 2017), individual fairness (Dwork et al., 2012; Kusner et al., 2017), subgroup fairness (Kearns et al., 2018; 2019), and preference-based notions (Zafar et al., 2017a; Ustun et al., 2019). Embracing these definitions, relevant algorithms involving fair constraints have been proposed. Zemel et al. (2013) propose a method to find a good representation to maximum utility while preserve both group and individual fairness, and following work on fair representation learning use autoencoder (Madras et al., 2018) or adversarial training (Zhao & Gordon, 2019; Zhao et al., 2019; Edwards & Storkey, 2015; Louizos et al., 2016) to simultaneously vanish the sensitive pattern while preserve enough information for prediction. Zafar et al. (2017b) optimize for decision boundary fairness through regularization in logistic regression and support vector machines, and some other work achieve fairness by optimal transport between sensitive groups (Gordaliza et al., 2019; Jiang et al., 2019) and fair kernel methods (Donini et al., 2018). However, most proposed learning algorithms for fairness are mainly built on independent and identically distributed data, but not suitable to be directly applied to graph-structured data with dyadic fairness.

Graph Representation Learning. Representation learning on graphs is formulated to convert a structural graph into a low-dimensional space while preserving the discriminative and structural representations. Efficient graph analytic methods (Von Luxburg, 2007; Tang et al., 2015; Perozzi et al., 2014; Grover & Leskovec, 2016; Xu et al., 2019) can benefit a series of downstream applications including node classification (Wang et al., 2017), node clustering (Nie et al., 2017), and link prediction (Zhang & Chen, 2018). Recently, Graph Neural Networks (GNNs) have shown remarkable capacity in graph representation learning, with emergent varieties (Kipf & Welling, 2016a; Veličković et al., 2017; Hamilton et al., 2017) consistently delivering promising results. Our work uses GNNs for graph representation learning but target at improving dyadic fairness in link prediction.

Fair Graph Embedding. As fairness in graph-structured data a relatively new topic for research, only a few studies have investigated the fair issues in graph embedding and graph representation learning. Rahman et al. (2019) pioneer Fairwalk, a random walk based graph embedding method that revises the transition probability according to the vertex's sensitive attributes. Following the

idea of adversarially removing sensitive patterns (Madras et al., 2018), Bose & Hamilton (2019) and Masrour et al. (2020) conduct adversarial training on vertex representations to minimize the marginal discrepancy. Other work but with less relevance includes fair collaborative filtering (Yao & Huang, 2017), item recommendation (Steck, 2018; Chakraborty et al., 2019) in bipartite graphs, and fair graph covering problem (Rahmattalabi et al., 2019).

3 PRELIMINARIES

Let $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ as a graph with a fix set of vertices $\mathcal{V}$ and edges $\mathcal{E}$, where vertex features with M dimensions are represented by $X \in \mathbb{R}^{N \times M}$. A nonnegative adjacency matrix $A \in \mathbb{R}^{N \times N}$ describes the relations between every pair of vertices. The element a_{vu} in A represents the weight on the linkage bridging v and u, and is set to zero if no link exists. Every vertex holds a sensitive attribute, and we use $S(v)$ to denotes the sensitive attribute as well as the sensitive group membership of v. Let $\Gamma(v)$ be the set of 1-hop neighbors of v including self-loop. Edge (v, u) is called intra if $S(v) = S(u)$, and inter implies $S(v) \neq S(u)$. $|S|$ denotes the cardinality of group S. For a binary sensitive attribute with two groups S_0 and S_1 separated from the graph, $\widetilde{S_0} := \{v \in S_0 | \Gamma(v) \cap S_1 \neq \varnothing\}$ represents the set of vertices in S_0 which locate on the boundary and has connections with S_1, and the same for $\widetilde{S_1}$. Set U to be the discrete uniform distribution over the set of vertices $\mathcal{V}$. Suppose a bivariate link prediction function $g(\cdot, \cdot) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, that given two vectors of the embedded vertices representations, a value is obtained showing the model belief that these two vertices are potentially linked.

Having these basic notations, we consider the disparity in link prediction bridging on intra and inter sensitive groups. The general purpose of dyadic fairness is to predict links independently of whether two vertices having the same sensitive attribute or not. We extend from demographic parity (Edwards & Storkey, 2015; Kipf & Welling, 2016b; Madras et al., 2018; Zemel et al., 2013) to formulate a specific criteria for dyadic fairness. In a binary classification problem, demographic parity expects a classifier gives positive outcome to two sensitive groups at the same rate. We turn the two groups in the content of demographic parity into the groups of intra and inter links. Ideally, achieving dyadic fairness will bring intra and inter link predictions at the same rate from a bag of potential links.

Having vertices representation v and u, dyadic fairness can be mathematically formulated as

Definition 3.1. A link prediction algorithm satisfies dyadic fairness if the predictive score satisfy

$$\Pr(g(u, v) | S(u) = S(v)) = \Pr(g(u, v) | S(u) \neq S(v)) \tag{1}$$

To quantify the fairness, we establish dyadic fairness on link prediction upon a fixed set of vertices, and models the expectation of absolute difference in score outcome across the groups of intra and inter links. Note that we also comprehensively evaluate our model on fairness by other four statistics gap extended from (Hardt et al., 2016) in Section 6.

4 HOW GRAPH CONNECTIONS AFFECT FAIRNESS

In this section, we propose a chain of theoretical analyses[1] established on a variant of demographic parity and graph neural networks to associate dyadic fairness with graph connections. We first demonstrate demographic parity in the outcomes of link prediction can be sufficiently reduced to the achievement of fair vertex representations when employing an inner product function for link prediction. Suggested by the sufficiency, we reveal how the pipeline of a one-layer graph neural network can affect the demographic parity, and draw the conclusion that for an arbitrary graph using GNNs for embedding, properly regulating the weights on existing graph connections can contribute to fairness conditionally. The theoretical findings motivate our algorithmic design as presented in the next section. Without loss of generality, in this section, we consider the sensitive attribute to be binary, where two sensitive groups S_0 and S_1 can be separated from the graph, but show the cases with sensitive attributes in multiple categorical values in our experimental section.

Proposition 4.1. For a link prediction function $g(\cdot, \cdot)$ modeled as inner product $g(v, u) = v^\top \Sigma u$, where $\Sigma \in \mathbb{S}^M_{++}$ is a positive-definite matrix, $\exists Q > 0, \forall v \sim \mathcal{V}$, $\|v\|_2 \leq Q$, for $\mathbb{E}_{v \sim U}[v] \in \mathbb{R}^M$, for

[1] Proofs for the proposition and theorem are in Appendix A.

dyadic fairness based on demographic parity, if $\|\mathbb{E}_{v\sim U}[v \mid v \in S_0] - \mathbb{E}_{v\sim U}[v \mid v \in S_1]\|_2 \leq \delta$,

$$\Delta_{\text{DP}} := |\mathbb{E}_{(v,u)\sim U\times U}[g(v,u) \mid S(v) = S(u)] - \mathbb{E}_{(v,u)\sim U\times U}[g(v,u) \mid S(v) \neq S(u)]| \leq Q\|\Sigma\|_2 \cdot \delta. \tag{2}$$

Remark 1. Proposition 4.1 can be applied for a general inner product function in Euclidean space, where Σ directionally and differently scales two input vectors. When setting Σ to an identity matrix, function $g(\cdot,\cdot)$ reduces to dot product and is widely used in a series of research work on link prediction (Kipf & Welling, 2016b; Trouillon et al., 2016; Yao & Huang, 2017).

The above proposition implies fair vertex representations is a sufficient condition to achieve demographic parity in link prediction. Suggested by the sufficiency, the approach to fairness could be reduced to achieving vertex representations with a small discrepancy between sensitive groups. With the proposition, we are ready to proceed to understand fairness within graph neural networks and reveal how the structure or connections of a graph could affect demographic parity.

A single layer GNN can be generically written as $\text{GNN}_\theta(X, \widetilde{A}) := \rho\left(\widetilde{A}XW_\theta\right)$, where ρ is a nonlinear activation function, $\widetilde{A}$ is the normalized adjacency matrix, and W_θ is the trainable weight matrix. One GNN layer can be decomposed into two disjoint phases: a vertex feature smoothing phase over the graph using $\widetilde{A}$, and a feature embedding phase using W_θ and ρ. Concretely, we consider left normalization $\widetilde{A} = D^{-1}A$ (D is the degree matrix) for feature smoothing. Equivalently, at an individual level, for each vertex it is one-hop mean-aggregation $\text{Agg}(v) := \text{deg}_w(v)^{-1}\sum_{u\in\Gamma(v)} a_{vu}u$, where $\text{deg}_w(v) := \sum_{u\in\Gamma(v)} a_{vu}$ stands for the weighted degree of vertex v.

We respectively abbreviate $\mathbb{E}_{v\sim U}[v|v{\in}S_0]$ and $\mathbb{E}_{v\sim U}[v|v{\in}S_1]$ as μ_0 and μ_1. Let σ denotes the maximal deviation of vertex representations, namely, $\forall v{\in}S_0$, $\|v{-}\mu_0\|_\infty \leq \sigma$, and $\forall v{\in}S_1$, $\|v{-}\mu_1\|_\infty \leq \sigma$. Let $D_{\text{max}} := \max_{v\in\mathcal{V}} \text{deg}_w(v)$ be the maximal weighted degree in $\mathcal{G}$, $m_w := \sum_{S(v)\neq S(u)} a_{vu}$ be the summation of weights on inter links. With these notations, we show how the discrepancy between μ_0 and μ_1 changes after conducting feature smoothing for one time over the graph.

Theorem 4.1. For an arbitrary graph with nonnegative link weights, after conducting one mean-aggregation over the graph, the consequent representation discrepancy between two sensitive groups $\Delta_{\text{DP}}^{\text{Aggr}} := \|\mathbb{E}_{v\sim U}[\text{Agg}(v) \mid v \in S_0] - \mathbb{E}_{v\sim U}[\text{Agg}(v) \mid v \in S_1]\|_2$ is bounded by

$$\max\{\alpha_{\min}\|\mu_0 - \mu_1\|_\infty - 2\sigma, 0\} \leq \Delta_{\text{DP}}^{\text{Aggr}} \leq \alpha_{\max}\|\mu_0 - \mu_1\|_2 + 2\sqrt{M}\sigma, \tag{3}$$

where $\alpha_{\min}{=}\min\{\alpha_1,\alpha_2\}$, $\alpha_{\max}{=}\max\{\alpha_1,\alpha_2\}$, $\alpha_1{=}|1{-}\frac{m_w}{D_{\text{max}}}(\frac{1}{|S_0|}{+}\frac{1}{|S_1|})|$, $\alpha_2{=}|1{-}\frac{|\widetilde{S_0}|}{|S_0|} - \frac{|\widetilde{S_1}|}{|S_1|}|$.

Remark 2. Theorem 4.1 shows the lower and upper bound given by the graph structure and the maximum deviation σ of vertex representations in each sensitive group on demographic parity after conducting one aggregation function on vertices. The contraction coefficient $\alpha_{\max}$ is a maximum of two absolute terms α_1 and α_2, where α_2 is a constant predetermined by the graph connections. It is worth pointing out that although in the worst case $\alpha_{\max}$ could be 1, e.g., in a complete bipartite graph, in most practical graphs it is strictly less than 1, hence the above upper bound corresponds to a contraction lemma but under some additional error introduced by deviation σ. We also provide several illustrative graph diagrams for the contraction of this theorem in Appendix B.

To approximate demographic parity after feature smoothing, Theorem 4.1 inspires a strategy to regulate the weights on graph connections to change α_1 in $\alpha_{\max}$ so as to minimize the upper bound for one-layer mean-aggregation. For term α_1, obviously if the summation of inter weights are too small ($m_w \to 0$) or too large ($m_w \to D_w \cdot \min\{|S_0|, |S_1|\}$), α_1 will approximate to 1. This indicates that increasing the weights on inter links cannot always guarantee to achieve a better demographic parity although inter group connections are always the minority in links, but should regulate this part to a proper range depending on the size of sensitive groups and the connected situation of a graph.

We combine the upper bound with the second feature embedding phase. Here we denote $\mu_i' := \mathbb{E}_{v\sim U}[\text{GNN}_\theta(v, \widetilde{A}) \mid v \in S_i]$, $i = 0, 1$, $Q' := \sup\{\|\text{GNN}_\theta(v, \widetilde{A})\|_2 \mid v \in \mathcal{V}\}$.

Corollary 4.1. For Δ_{DP} on vertices after passing one layer $\text{GNN}_\theta(X, \widetilde{A}) = \rho\left(\widetilde{A}XW_\theta\right)$, we have:

$$\Delta_{\text{DP}} \leq\ Q'\|\Sigma\|_2 \cdot \|\mu_0' - \mu_1'\|_2 \leq QL^2\|\Sigma\|_2\|W_\theta\|_2^2 \cdot (\alpha\|\mu_0 - \mu_1\|_2 + 2\sqrt{M}\sigma), \tag{4}$$

Algorithm 1: Algorithmic routine for **FairAdj**

Input: vertex features X, adjacency matrix A, GNNs parameters θ, learning rates η_θ and $\eta_{\widetilde{A}}$

Normalize adjacency matrix $\widetilde{A} \leftarrow D^{-1}A$

Fix the elements with zero in $\widetilde{A}$ and select the non-zero elements for optimization

while θ *or* $\widetilde{A}$ *has not converged* **do**

  for $t = 1$ **to** T_1 ▷ `optimize for utility`

  do

    Compute $\mathcal{L}_{\text{util}}$ by Eq. (5), $\quad g_\theta \leftarrow \nabla_\theta \mathcal{L}_{\text{util}}, \quad \theta \leftarrow \theta + \eta_\theta \cdot \text{Adam}(\theta, g_\theta)$

  for $t = 1$ **to** T_2 ▷ `optimize for fairness`

  do

    $Z \leftarrow \text{GNN}_\theta(X, \widetilde{A}), \quad \hat{A} \leftarrow ZZ^\top$ ▷ *`reconstruct graph connections`*

    Compute $\mathcal{L}_{\text{fair}}$ by Eq. (6), $\quad g_{\widetilde{A}} \leftarrow \nabla_{\widetilde{A}} \mathcal{L}_{\text{fair}}$

    for $v = 1$ **to** N ▷ `projected gradient descent`

    do

      Sort n non-zero elements in $[\widetilde{A} - \eta_{\widetilde{A}} g_{\widetilde{A}}]_{v,*}$ in descending order: $e_1 \geq e_2 \geq \cdots e_n$

      $\gamma \leftarrow \sum_{j=1}^{n} \mathbf{1}(e_j + \frac{1}{j}(1 - \sum_{i=1}^{j} e_i) \geq 0)$ ▷ $\mathbf{1}(\cdot)$:*`the indicator function`*

      $\beta \leftarrow \frac{1}{\rho}(1 - \sum_{i=1}^{\gamma} e_i)$

      for $u = 1$ **to** n ▷ *`update`* $\widetilde{A}$

      do

        $[\widetilde{A}]_{v,u} \leftarrow \max\{[\widetilde{A} - \eta_{\widetilde{A}} g_{\widetilde{A}}]_{v,u} + \beta,\ 0\}$

Output: Link predictive score between vertex v and $u \leftarrow \text{sigmoid}(\text{GNN}_\theta(v, \widetilde{A})^\top \text{GNN}_\theta(u, \widetilde{A}))$

where L is the Lipschitz constant for ρ. The first inequality holds by Proposition 4.1, the second one is by Theorem 4.1 and the definition of spectral norm and Lipschitz constant, also realizing that $Q' \leq L\|W_\theta\|_2 Q$. Multiple layers of GNNs can be reasoned out similarly.

From the above theorem, we see Δ_{DP} can be processed with a tighter upper bound by regulating the weights on edges, and it is also dependent on the property of W_θ, ρ, and the error term $O(\sigma)$. When setting W_θ fixed, our theoretical findings provide a feasible solution that regulating weights on graph connections can achieve better demographic parity on link prediction, and as supplementary, also indicate where or when it cannot perform well with finite layers of GNNs: (1) The solution cannot guarantee arbitrary fairness if we want to preserve the graph structure due to the resistance of α_1 in lower bound in Theorem 4.1. (2) When it is already fair enough in the original graph data, which means $\|\mu_0 - \mu_1\|_2$ is small and additional error $O(\sigma)$ is comparable to it, the upper bound cannot be reduced significantly and the solution may not further mitigate the bias. We include a dataset to investigate the potential limitations empirically in response to above analysis in Appendix D. In the following sections, we implement the inspired algorithm and demonstrate that multiple real-world networks accept this solution with favorable results and a better fairness-utility tradeoff.

5 Learning Fair Graph Connections

The above discussion indicates that when employing GNNs for graph embedding, adjusting the adjacency matrix can assist the model with achieving fairness conditionally. However, searching for the optimal adjacency matrix within hierarchical graph neural networks is a non-trivial problem. In this section, continuing with the preceding analysis, we develop **FairAdj** algorithm to adjust the graph connections and learn a fair adjacency matrix by updating $\widetilde{A}$ while preserving the original graph structure unchanged. In overview, we implement the algorithm by separately optimizing the parameter W_θ of GNNs towards utility, and adjusting $\widetilde{A}$ towards dyadic fairness by gradient descent and empirical risk minimization with structural and right stochastic constraints. Therefore, **FairAdj** is able to pursue the supplied dyadic fairness and link predictive utility simultaneously.

We employ variational graph autoencoder (Kipf & Welling, 2016b) for feature embedding. A two-layer graph neural network is used as inference model $\text{GNN}_\theta(\cdot,\cdot)$. Z denotes the embedded representations, and dot product between embedded representations is the generative model $p(\cdot)$. The KL-divergence term $KL[\cdot\|\cdot]$ punishes the discrepancy between latent distribution and a Gaussian prior. The objective function to reconstruct graph connections from latent variable can be written as:

$$\max_{\theta} \quad \mathcal{L}_{\text{util}} := \mathbb{E}_{\text{GNN}_\theta(Z|X,\widetilde{A})}[\log p(A \mid Z)] - KL[\text{GNN}_\theta(Z \mid X, \widetilde{A}) \| \mathcal{N}(0,1)]. \tag{5}$$

For fairness, we impel $\mathcal{L}_{\text{fair}}$ to empirically seek for better graph connections, then update $\widetilde{A}$ with constraints. Specifically, we optimize the normalized adjacency matrix $\widetilde{A}$ as follows:

$$\begin{aligned} \min_{\widetilde{A}} \quad & \mathcal{L}_{\text{fair}} := \|\mathbb{E}_{v,u\sim U\times U}[\hat{a}_{vu} \mid S(v) = S(u)] - \mathbb{E}_{v,u\sim U\times U}[\hat{a}_{vu} \mid S(v) \neq S(u)]\|^2, \\ \text{s.t.} \quad & (1).\ [\widetilde{A}]_{vu} = 0, \text{ if } [A]_{vu} = 0, \quad (2).\ \widetilde{A}\mathbb{1} = \mathbb{1},\ \widetilde{A} \geq 0, \end{aligned} \tag{6}$$

where $\hat{a}_{vu}$ takes value in $\hat{A} = ZZ^\top$ and $\mathbb{1}$ is the all-one vector with size N. The two constraints are necessary for optimizing $\widetilde{A}$: (1) Elements with zero value should be maintained, meaning no new links can be established during optimization. This restriction is proposed to preserve utility, due to adding fictitious links might mislead the directions of message passing, and further corrupt the representation learning. Therefore, we only adapt weights on existing edges and preserve the original graph structure. (2) In consistent with the initial left normalization $\widetilde{A} = D^{-1}A$, the optimized matrix should still remain a right stochastic matrix. This can restrict the largest eigenvalue of adjacency matrix $\|\widetilde{A}\|_2 = 1$, hence avoid numerical instabilities and exploding gradients when training W_θ towards $\mathcal{L}_{\text{util}}$. In practice, we observe explosions during training with no constraints applied on $\widetilde{A}$.

For the first constraint in Eq. (6), we only compute gradients and update the elements in $\widetilde{A}$ which have non-zero initialization. For the second one, after selecting the variable to optimize, we employ projected gradient descent by (Wang & Carreira-Perpinán, 2013) to satisfy the constraint while minimizing $\mathcal{L}_{\text{fair}}$. Once given computed gradients on $\widetilde{A}$ denoted as $\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}}$, with the corresponding learning rate $\eta_{\widetilde{A}}$, we have the following optimization problem that update $\widetilde{A}$ by projecting $\widetilde{A} - \eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}}$ into the feasible region with the minimum Euclidean distance:

$$\begin{aligned} \min_{\widetilde{A}-(\eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}})'} \quad & \sum_v \|[\widetilde{A} - \eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}}]_{v,*} - [\widetilde{A} - (\eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}})']_{v,*}\|^2 \\ \text{s.t.} \quad & (\widetilde{A} - (\eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}})')\mathbb{1} = \mathbb{1} \text{ and } \widetilde{A} - (\eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}})' \geq 0, \end{aligned} \tag{7}$$

where $[\widetilde{A}]_{v,*}$ are the row-wise elements for $\widetilde{A}$, namely, all the connections on v. $\widetilde{A} - (\eta_{\widetilde{A}}\nabla_{\widetilde{A}}\mathcal{L}_{\text{fair}})'$ is the update for $\widetilde{A}$ after projection. Since $\widetilde{A}$ is row-wise independent and the objective function is strictly convex for this quadratic program, there exists a unique solution for each row. Solution details for projected gradient descent are restated as a part of our algorithmic pipeline.

The algorithmic routine is elaborated in Algorithm 1. θ and $\widetilde{A}$ are optimized iteratively with T_1 and T_2 epochs for co-adaptation. Compared to adversarial training method on graph embedding (Bose & Hamilton, 2019), which requires a hyperparameter to control the fairness-utility tradeoff, we also find regulating the convergence of $\widetilde{A}$ has a similar effect as well. That is because the more $\widetilde{A}$ changes, the further $\widetilde{A}$ is away from the original graph connections, and consequently, the more damage in utility but more enhancement in fairness. Thanks to this favorable property, in experiments, comparing to adversarially remove sensitive attributes, we present multiple options for T_2 to control the convergence and observe a more favorable fairness-utility tradeoff shown by our method.

6 EXPERIMENTS

We present empirical analysis on six real-world datasets, compared with baseline methods in terms of seven evaluative metrics on both fairness and utility. Approaching to applications, we testify that our method can enhance the diversity in recommendations. Due to the space limitation, we only showcase partial results but defer the rest in Appendix D. Moreover, as an intermediate result, we demonstrate vertex representations are embedded more fairly assessed by fair clustering in Appendix E.

Table 1: Statistic for datasets in experiments.

Dataset	# Vertex	# Edge	# Class	# Intra	# Inter	Intra Ratio	Inter Ratio	Dis. Ratio
Oklahoma97	3,111	73,230	2	46,368	26,862	1.92e-2	1.11e-2	1.73
UNC28	4,018	65,287	2	36,212	29,075	8.76e-3	7.38e-3	1.19
Facebook#1684	786	14,024	2	7,989	6,035	4.76e-2	4.30e-2	1.11
Cora	2,708	5,278	7	4,275	1,003	6.51e-3	3.30e-4	19.73
Citeseer	3,312	4,660	6	2,089	2,571	2.13e-3	5.70e-4	3.74
Pubmed	19,717	44,327	3	33,443	10,884	4.80e-4	9.00e-5	5.33

Table 2: Experimental results on *UNC28*.

Method	AUC ↑	AP ↑	Δ_{DP} ↓	Δ_{true} ↓	Δ_{false} ↓	Δ_{FNR} ↓	Δ_{TNR} ↓
VGAE	**87.63** ± 0.56	**88.69** ± 0.65	2.24 ± 0.42	1.50 ± 0.41	0.44 ± 0.36	7.62 ± 0.84	2.18 ± 0.72
node2vec	87.22 ± 0.30	87.10 ± 0.37	2.75 ± 0.78	1.30 ± 0.53	1.05 ± 0.93	12.56 ± 1.12	2.24 ± 0.92
Fairwalk	87.18 ± 0.30	87.07 ± 0.37	2.79 ± 0.70	1.17 ± 0.49	0.90 ± 0.92	12.71 ± 1.11	2.20 ± 0.96
$\text{FairAdj}_{\text{T2=5}}$	86.98 ± 0.54	87.75 ± 0.65	**1.53** ± 0.35	**0.32** ± 0.29	**0.41** ± 0.35	2.84 ± 0.74	2.22 ± 0.68
$\text{FairAdj}_{\text{T2=20}}$	87.04 ± 0.55	87.80 ± 0.65	1.57 ± 0.36	0.34 ± 0.31	0.42 ± 0.35	**2.76** ± 0.75	**2.16** ± 0.73

6.1 SETTINGS

Datasets. We conduct experiments on real-world social networks and citation networks including *Oklahoma97*, *UNC28* (Traud et al., 2011), *Facebook#1684*, *Cora*, *Citeseer*, and *Pubmed*. *Oklahoma97* and *UNC28* are two school social networks. A link represents a friendship relation in social media, and every user has a profile for vertex features, including student/faculty status, gender (sensitive attribute), major, etc. *Facebook#1684* is a social ego network from Facebook app. As the rest three citation networks, each vertex represents an article with bag-of-words descriptions as features. A link stands for a citation regardless the direction. We set the category of an article as the sensitive attribute. Statistic for datasets are summarized in Table 1, where #Class is the number of sensitive groups, #Intra/Inter Ratio represents the ratio that the number of actual intra/inter links v.s. the number of links if the graph is fully connected. These two terms show the density of intra/inter links. Dis. Ratio (abbreviated from disparity ratio) is calculated by divide intra ratio into inter ratio. Dis. Ratio equaling to one implies intra/inter connections are perfectly balanced in existing graph, and the degree of deviation from 1 indicates how skew the link connections are.

Baselines and Protocols. We involve four baseline methods. Variational graph autoencoder (VGAE) (Kipf & Welling, 2016b) inherits from variational autoencoder, which uses two GNN-layer as the inference model and leverages latent variables to reconstruct the graph connections. Node2vec (Grover & Leskovec, 2016) is a widely used graph embedding approach based on random walk. Fairwalk (Rahman et al., 2019) is built upon node2vec and designed specifically for fairness issues. It modifies the transition probability for one vertex according to the sensitive attribute of its neighbors. The last one is adversarial training on vertex representations (Bose & Hamilton, 2019), which aims to minimize the discrepancy between different sensitive groups by optimizing parameters in GNN. Besides the standard pipeline for utility, it additionally trains the networks to confuse a discriminator, meanwhile training the discriminator to distinguish the embedded features with different sensitive attributes. A hyperparameter λ is used in the overall objective function to balance the tradeoff between utility and fairness. We vary λ in experiments and make comparisons to various results given by adversarial training. For all experiments, we randomly remove 10% links from the graph and reserve them for evaluation, and equivalently, the same number of false links are sampled in the evaluation phase. For one dataset, we repeat experiments with different train/test splits for 20 times. Full experimental configurations are available in Appendix C.

Metrics. We evaluate the utility of link prediction using Area Under the Curve (AUC) and Average Precision (AP). Fairness is evaluated towards Δ_{DP}, as well as the disparity on the expected score on all the true samples Δ_{true} and false samples Δ_{false}. Besides these, following the suggested fairness notions (Hardt et al., 2016), we compute the maximum gap of true negative rate (TNR) and false negative rate (FNR). Illustratively, let the conditional cumulative distribution function of score R be evaluated by a threshold τ with a given label written as $F_y^s(\tau) := \Pr(R{\leq}\tau|Y{=}y, S{=}s)$, $y{\in}\{0,1\}$, $s{\in}\{\text{intra}, \text{inter}\}$. With these notations, the maximum gap in true negative rate can be expressed as $\Delta_{\text{TNR}} := \max_\tau |F_0^{\text{intra}}(\tau) - F_0^{\text{inter}}(\tau)|$, and similar for false positive rate $\Delta_{\text{FNR}} :=$

Table 3: Experimental results on *Citeseer*.

Method	AUC ↑	AP ↑	Δ_{DP} ↓	Δ_{true} ↓	Δ_{false} ↓	Δ_{FNR} ↓	Δ_{TNR} ↓
VGAE	**81.77** ± 1.23	**85.57** ± 1.39	11.24 ± 1.83	3.37 ± 2.33	2.14 ± 1.26	10.81 ± 3.61	11.31 ± 2.99
node2vec	81.21 ± 1.35	84.69 ± 1.26	14.49 ± 3.38	4.02 ± 2.74	6.82 ± 4.62	7.37 ± 3.07	12.50 ± 4.21
Fairwalk	81.69 ± 1.50	84.97 ± 1.23	13.50 ± 2.97	3.30 ± 2.49	5.33 ± 3.83	**7.26** ± 3.30	11.34 ± 3.19
FairAdj$_{\text{T2=2}}$	80.45 ± 1.34	84.47 ± 1.43	9.57 ± 1.84	2.55 ± 2.02	1.70 ± 1.47	9.87 ± 3.17	10.27 ± 3.23
FairAdj$_{\text{T2=20}}$	78.84 ± 1.38	82.74 ± 1.46	**7.81** ± 1.80	**1.85** ± 1.66	**1.30** ± 1.41	9.22 ± 2.95	**10.01** ± 3.13

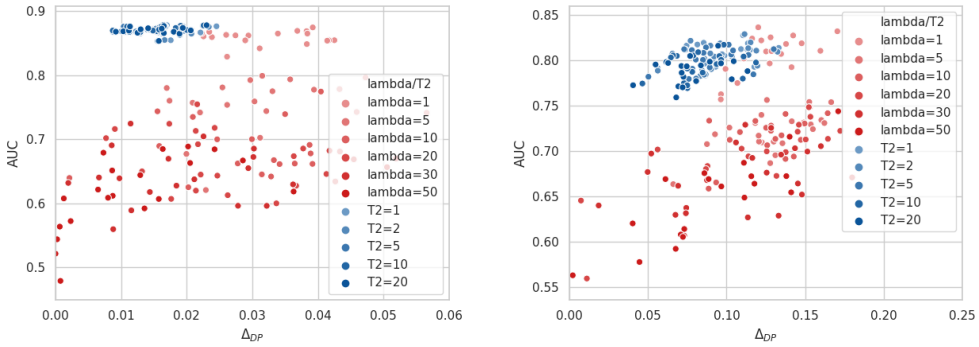

Figure 1: Comparison with adversarial training method (Bose & Hamilton, 2019) in terms of the tradeoff between utility and fairness. Left: *UNC28*; Right: *Citeseer*. Blue points denote **FairAdj** with different T_2 values, red points represent (Bose & Hamilton, 2019) with different λ values.

$\max_\tau |F_1^{\text{intra}}(\tau) - F_1^{\text{inter}}(\tau)|$. These two terms also reflect the disparity in true positive rate (TPR) and false positive rate (FPR) due to TPR $= 1 -$ FNR and FPR $= 1 -$ TNR.

6.2 Results and Analysis

Table 2 and 3 list quantitative results on *UNC28* and *Citeseer* comparing to VGAE, node2vec, and Fairwalk. Two choices of T_2 are presented here, where T_2 in a smaller value pursues a comparable performance in utility (a little bit lower in AUC but higher in AP) to random walk based methods, and at the same time performs much better in fairness. We also present T_2 = 20 since we observe a convergence on the adjacency matrix. The results indicate that **FairAdj** achieves the best on various statistics for dyadic fairness and with only a small sacrifice in predictive utility.

A Better Tradeoff. Figure 1 plots every experimental result for our method and adversarial training with various fairness-utility regulated hyperparameters. Two observations explain why our method surpasses the adversarial technique. (1) Blue dots are closer to the left top corner than the red on the whole, meaning the same level of fairness is achieved with less sacrifice in utility. An explanation for this favorable property is that, the adversarial training method neglects the graph connections and only diminish the group discrepancy, where two irrelevant instances with no connection but from different groups may be closely mapped, thus greatly damage the utility. The optimization on $\widetilde{A}$ does facilitate the feature smoothing across groups which is not indicated in the original adjacency matrix, but still considers the graph connections. (2) Additionally, blue dots are more aggregated, suggesting our methods escape from the instability of min-max optimization and acting more robust to different train/test splits. However, as shown, **FairAdj** cannot achieve arbitrary small in Δ_{DP} as red dots do. This is indicated in Section 4 as the first potential limitation.

Diversity in Recommendations. We examine the top-scored links in evaluation at a certain proportion in terms of its diversity and utility, shown in Figure 2. This exploration can be useful when conducting recommendations according to scores in descending order. For a fixed proportion, we report the diversity as the number of inter links divided by the number of intra links, and the utility as the recall rate among these recommendations. Figures show that as a direct product by dyadic fairness, **FairAdj** enhances the diversity in recommendations but is achieved at a sacrifice of utility.

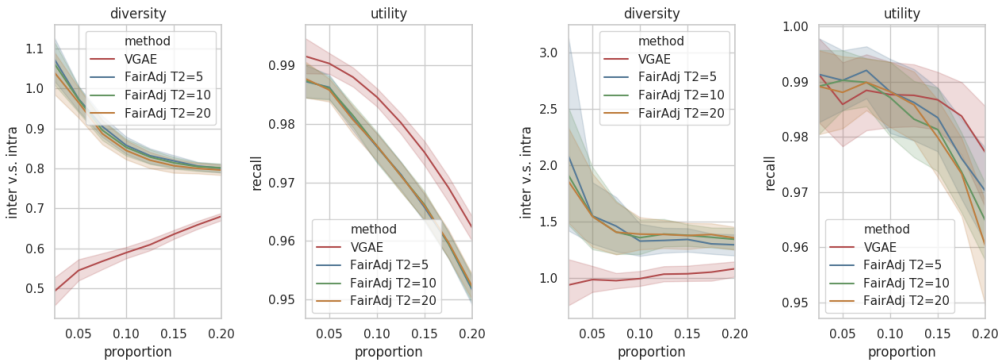

Figure 2: Diversity and utility in recommendations. Left: *UNC28*; Right: *Citeseer*.

7 Conclusion

We studied the dyadic fairness in graph-structured data. We theoretically analyzed how the connections in graph links affect dyadic fairness of demographic parity when employing graph neural networks for representation learning. On the basis of the foregoing analysis, we proposed **FairAdj** to learn a fair adjacency matrix, and pursued the dyadic fairness and prediction utility simultaneously. Empirical validations demonstrated the achievement of fairness and a better fairness-utility tradeoff.

REFERENCES

Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23: 2016, 2016.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 715–724, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/bose19a.html.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 129–138, 2019.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

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

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365, 2019.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Bas Hofstra, Rense Corten, Frank Van Tubergen, and Nicole B Ellison. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review*, 82(3):625–656, 2017.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 100–109, 2019.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016a.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016b.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *ICLR*, 2016. URL `http://arxiv.org/abs/1511.00830`.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. Bursting the filter bubble: Fairness-aware network link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 841–848, 2020.

Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised large graph embedding. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.

Yanjun Qi, Ziv Bar-Joseph, and Judith Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, 63(3):490–500, 2006.

Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: towards fair graph embedding. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3289–3295. AAAI Press, 2019.

Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, and Milind Tambe. Exploring algorithmic fairness in robust graph covering problems. In *Advances in Neural Information Processing Systems*, pp. 15750–15761, 2019.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 154–162, 2018.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pp. 1067–1077, 2015.

Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. International Conference on Machine Learning (ICML), 2016.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382, 2019.

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

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning*, pp. 6932–6941, 2019.

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pp. 2921–2930, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pp. 229–239, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 5165–5175, 2018.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *Advances in neural information processing systems*, pp. 15649–15659, 2019.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019.