All of the Fairness for Edge Prediction with Optimal Transport

Charlotte Laclau, Ievgen Redko, Manvi Choudhary, Christine Largeron

All of the Fairness for Edge Prediction with Optimal Transport

Charlotte Laclau **Ievgen Redko** **Manvi Choudhary** **Christine Largeron**
Univ Lyon, UJM-Saint-Etienne, CNRS
Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516,
F-42023, Saint-Etienne, France

Abstract

Machine learning and data mining algorithms have been increasingly used recently to support decision-making systems in many areas of high societal importance such as healthcare, education, or security. While being very efficient in their predictive abilities, the deployed algorithms sometimes tend to learn an inductive model with a discriminative bias due to the presence of this latter in the learning sample. This problem gave rise to a new field of algorithmic fairness where the goal is to correct the discriminative bias introduced by a certain attribute in order to decorrelate it from the model's output. In this paper, we study the problem of fairness for the task of edge prediction in graphs, a largely underinvestigated scenario compared to a more popular setting of fair classification. To this end, we formulate the problem of fair edge prediction, analyze it theoretically, and propose an embedding-agnostic repairing procedure for the adjacency matrix of an arbitrary graph with a trade-off between the group and individual fairness. We experimentally show the versatility of our approach and its capacity to provide explicit control over different notions of fairness and prediction accuracy.

1 INTRODUCTION

We live in a world where an increasing number of decisions, with major societal consequences, are made or at least supported by algorithms that diligently learn the patterns from a training sample and gain their discriminating ability by identifying the key attributes correlated with the desired output. These attributes, however, can represent sensitive information that, in its turn, can lead to a significant bias in model's predictions when deployed on a previously unseen sample. For instance, when building a recommendation system supporting a recruitment company in finding a potential candidate suitable for their clients' needs, one wants to provide accurate recommendation of job offers (edge prediction task) with similar offers shown to people with similar profiles (individual fairness), while from a legal perspective, this process should not depend on criteria such as the gender or the ethnicity of a candidate (group fairness). In practice, such bias prevents minorities from gaining influence in the network as studied in [Stoica et al., 2018]. In practice, however, the training sample used to learn the model may have been collected in a biased manner with an unequal number of successive outcomes between the genders and/or ethnic groups. The recommendations of the learned model in this case will tend to follow the learned pattern thus reinforcing the already existent bias. Fairness in the context of online social graph can also be seen as a way to prevent online polarization by promoting links between users having opposite views, on controversial topics, such as political debates. In this context, link prediction algorithms tend to promote links between people having the same opinion (eg. belonging to the same political party) since they are more likely to be connected in the network, which leads to the increase in the formation of online bubbles and thus segregation of the network. Research works aiming at identifying and correcting such inductive bias form the core of the algorithmic fairness field, a scientific area that is constantly gaining more and more attention from the machine learning and data mining communities nowadays.

Algorithmic fairness methods are traditionally divided into one of the three following categories: (i) pre-processing methods that repair the original data to remove the bias, ii) methods that integrate fairness constraints or penalties in a given learning algorithm and iii) post-processing methods that debias directly the model's output. First family of methods can be further divided into two subfamilies where the first one corrects the input raw data to ensure that the inference of the sensitive attribute is impossible, regardless of the learning algorithm (e.g. classifier) used downstream [Feldman et al., 2015a, Calmon et al., 2017, Johndrow and Lum,

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

2019], while the other learns a new representation that is forced to be independent from the sensitive attribute [Zemel et al., 2013, Edwards and Storkey, 2016, Louizos et al., 2016, Madras et al., 2018]. These methods present the most generic solution to the considered problem as they allow to use any available algorithm on the repaired data and ensure that the generalization performance on this latter would be comparable to that obtained on the original data [Gordaliza et al., 2019]. The methods belonging to the second category [Zafar et al., 2017a, Zafar et al., 2017b, Corbett-Davies et al., 2017, Agarwal et al., 2018, Donini et al., 2018] are specific to the learning algorithm and thus the modification of this latter due to, for instance, a performance drop on another data set requires modifying the whole optimization procedure. Finally, the last category [Hardt et al., 2016, Kusner et al., 2017, Jiang et al., 2019, Chiappa, 2019, Zehlike et al., 2020] of methods has a virtue of debiasing directly the outputs of a learning algorithm, but similarly to the methods that impose fairness constraints while learning require the post-processing to be performed for each prediction.

Most of the works mentioned above address the problem of algorithmic fairness in the context of supervised classification and completely ignore learning from relational data given in form of structured objects or graphs, and the tasks associated to it. Such data, however, is ubiquitous in areas dealing with complex systems, especially in the social sciences where the relationships and interactions between people are studied. Several mining tasks can be defined for such data, such as edge prediction, node classification or community detection, to name a few.

Contributions In this paper, we propose a first theoretically sound embedding-agnostic method for group and individually fair edge prediction. This is done through the following contributions:

1. We analyze the group fair edge prediction task theoretically and show that one can efficiently repair the adjacency matrix of a graph by aligning the joint distributions of nodes appearing in different sensitive groups.

2. We derive an optimal transport (OT)-based algorithm from our analysis, add individual fairness constraints to it and implement it for binary and multi-class settings. The proposed algorithm outputs a repaired adjacency matrix by adding edges that obfuscate the dependence on the sensitive attribute. The repaired adjacency matrix can be used as input of any node embedding technique thus making it embedding agnostic.

3. We evaluate the efficiency of our approach through extensive experiments on several synthetic and real-world data sets and show that it provides an explicit control on the trade-off between the two notions of fairness and prediction accuracy.

Organisation The rest of this paper is organized as follows. We provide a theoretical analysis of group fair edge prediction in Section 2. In Section 3, we present a group and individually fair repair scheme for adjacency matrix of a graph. In Section 4, we evaluate our approach on synthetic and real-world networks and show the impact of the proposed repairing scheme both on the capacity of predicting the sensitive attribute from embeddings learned with repaired data and on the performance of edge prediction. Last section concludes the paper and gives a couple of hints for possible future research.

2 FAIR EDGE PREDICTION

In this section, we formulate the problem of fair edge prediction in graphs and give a definition of several key concepts related to it such as statistical parity, disparate impact and balanced error rate. We further analyze this problem and derive a theoretically sound approach allowing to solve it.

2.1 Problem Setup

Let $\mathbb{V}$ denote an abstract vertex space, and let $\mathcal{V} = \{v_1, \dots, v_N\} \in \mathbb{V}^N$ be a set of N vertices drawn independently and identically (iid) from an arbitrary distribution over $\mathbb{V}$. Let $r : \mathbb{V} \times \mathbb{V} \to \{0, 1\}$ be a (symmetric) true edge prediction function that outputs 1 if there is an edge between two nodes and 0 otherwise. In the finite case, we further consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subseteq \mathbb{V} \times \mathbb{V}$ is labeled according to r such that a tuple $\{(v_i, v_j, r(v_i, v_j)\}$ defines an undirected edge $(i, j) \in \mathcal{E}$. Furthermore, we assume that all nodes have one categorical sensitive attribute $S : \mathbb{V} \to \mathbb{S}$ where, for simplicity, we assume that $\mathbb{S}$ is a set $\{0, 1\}$. In the context of fair edge prediction, this variable defines a potential source of *bias* in the graph where $S = 0$ stands for the minority (unfavored) class, while $S = 1$ stands for the default (favored) class. In what follows, we are interested in the edge prediction task where the goal is to find a function $h : \mathbb{V} \times \mathbb{V} \to \{0, 1\}$ such that h is as close as possible to r.

We define the notion of *statistical parity* of h for this scenario as the equality between the probability of h for predicting the same value, say 1, for both nodes belonging to the same and different classes. More formally, this definition is given below.

Definition 1. *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a function $h : \mathbb{V} \times \mathbb{V} \to \{0, 1\}$, we define the statistical parity for an edge predictor h on S with respect to (w.r.t) $\mathbb{V}$ as:*

$$\mathbb{P}(h(V, V') = 1 | S \neq S') = \mathbb{P}(h(V, V') = 1 | S = S')$$

or equivalently

$$\mathbb{P}(h(V, V') = 1 | S \oplus S' = 1) = \mathbb{P}(h(V, V') = 1 | S \oplus S' = 0),$$

where $\oplus$ stands for XOR operation and the probability is taken over random variables $((V, S), (V', S')) \sim \mathcal{D} \times \mathcal{D}$ with $\mathcal{D}$ denoting the joint distribution over $\mathbb{V} \times \mathbb{S}$.

This definition states that the probability for h to predict an edge between two nodes v and v' is the same whether v and v' belong to the same ($S=0,\ S'=0$ or $S=1,\ S'=1$) or to the different ($S=1,\ S'=0$ or $S=0,\ S'=1$) classes. One may note that it is different from the definition considered in (node) classification task with $\mathbb{V} \subset \mathbb{R}^d,\ h : \mathbb{V} \to \{0,1\}$ as this latter trivially reduces to a usual fair classification problem studied extensively in the literature. In our case, however, we have to deal with tuples of variables and implicitly attribute a sensitive variable defined by $S \oplus S'$ to each pair of nodes or, equivalently, to an edge. On a higher level, this transposes the initial problem into repairing adjacency matrices contrary to repairing the feature representation of nodes as done in the case of node classification. Bearing in mind the equivalent XOR representation, we further denote these two events by

$$\mathbb{P}_1(h) = \mathbb{P}(h(V,V')=1 | S \neq S'),$$
$$\mathbb{P}_0(h) = \mathbb{P}(h(V,V')=1 | S = S'). \quad (1)$$

Using these notations, we define two other important fairness measures, notably *disparate impact* (DI) and *balanced error rate* (BER). We give their definitions below.

Definition 2. *Given a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ and a function $h : \mathbb{V} \times \mathbb{V} \to \{0,1\}$, let $\mathbb{P}_0(h)$ and $\mathbb{P}_1(h)$ be defined as in* (1)*. We define disparate impact (DI) and balanced error rate (BER) for an edge predictor h on $S \oplus S'$ w.r.t. $\mathbb{V}$ as:*

$$\textit{DI}(h,\mathbb{V},S\oplus S') = \frac{\mathbb{P}_1(h)}{\mathbb{P}_0(h)},$$
$$\textit{BER}(h,\mathbb{V},S\oplus S') = \frac{\mathbb{P}_1(h)-\mathbb{P}_0(h)+1}{2}.$$

Each of these two measures has its own interpretation. DI is identical to statistical parity and it is equal to 1, when h is perfectly fair. In practice, we are interested in bounding this latter, i.e., $\text{DI}(h,\mathbb{V},S\oplus S') \leq \tau$, indicating that a classifier has a disparate impact at level $\tau \in (0,1]$. As for the BER, it stands for a misclassification error of the sensitive attribute S by h in a setting where $\mathbb{P}(S\oplus S'=1)=\mathbb{P}(S\oplus S'=0)$. This latter condition roughly tells us that the probability of drawing a pair of nodes belonging to the same class should be the same as the probability of them belonging to different classes. It is important to note that while higher values of disparate impact and τ indicate a more fair outcome, the best misclassification error in terms of fairness is equal to $\frac{1}{2}$ as in this case a classifier is not capable of predicting whether the nodes are from the same or different classes.

Remark 1. *In what follows, it will be also convenient to define both the disparate impact and the balanced error rate w.r.t. S by considering only the conditioning on one variable of the pair (S,S'). For this latter case, we write*

$$\textit{DI}(h,\mathbb{V},S) = \frac{\mathbb{P}(h(V,V')=1|S=0)}{\mathbb{P}(h(V,V')=1|S=1)}$$

and similarly for BER$(h,\mathbb{V},S)$.

We proceed to the analysis of our fairness setting below.

2.2 Analysis of Group Fair Edge Prediction

Several works [Feldman et al., 2015b, Gordaliza et al., 2019, Jiang et al., 2019] provided a theoretical analysis for the fair classification setting where one deals with one random variable $X : \Xi \to \mathbb{R}^d$ with Ξ being an arbitrary instance space and considers learning a hypothesis function $h : \mathbb{R}^d \to \{0,1\}$. Below, we provide the analysis for the edge prediction fairness and relate it to the statistical parity of h in predicting the sensitive attributed individually for one of the node's pair. Note that from the algorithmic point of view, working with edges given by pairs of nodes and their associated sensitive attributes $S \otimes S'$ is hard as they do not admit any representation allowing further repair. To this end, our goal would be to simplify the problem in a principled way by considering learning on the joint space of nodes with the sensitive attribute being related to only one node from a pair. In this case, we would be able to use the pair-wise information about the nodes as node representation.

To proceed, we first make the following assumptions.

A1. The probability of each node belonging to the favoured or unfavoured class is the same, i.e.,

$$\mathbb{P}(S=0)=\mathbb{P}(S'=0)=\mathbb{P}(S=1)=\mathbb{P}(S'=1)=\frac{1}{2}.$$

A2. The probability of predicting an edge given that both nodes are in the same class is higher than that of predicting an edge between the nodes of different classes, i.e., for all $s \in \{0,1\}$

$$\mathbb{P}(h(V,V')=1|S=s,S'=1-s) \leq \mathbb{P}(h(V,V')=1|S=s,S'=s).$$

We state our main result below[1].

Theorem 1. *Consider a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, sets $S,S' \in \{0,1\}$, an edge prediction function $h : \mathbb{V}\times\mathbb{V} \to \{0,1\}$ and assume that DI$(h,\mathbb{V},S) \leq \tau$ for some $\tau \in (0,1]$. Then, with the assumptions A1-A2 the following holds:*

$$\textit{DI}(h,\mathbb{V},S\oplus S') \leq \textit{DI}(h,\mathbb{V},S) \leq \tau.$$

Before discussing the implications of this theorem, we briefly note that the assumptions A1-A2 are not restrictive and capture one's intuition about the fair edge prediction setting. Indeed, A1 assumes that each node has an equal probability of belonging or not to the favoured class which is a reasonable assumption to make in practice when sampling vertices at random. This assumption is not related to

[1] All proofs are provided in the Supplementary material.

the sampling process and imbalanced datasets. For instance, $P(S = \text{man}) = P(S = \text{women})$ holds with almost exact equality but does not imply that the number man and women in the given graph is the same. This assumption is commonly used to draw the equivalence between $\text{BER}(h, X, S)$ and the missclassification error $\mathbb{P}(h(X) \neq S)$ [Gordaliza et al., 2019]. In its turn, A2 states that the probability of predicting an edge inside any of the two classes is higher than that of predicting an edge between different classes. This latter is related to assortativity effect and it is rather intuitive as sensitive attributes are often correlated with a certain latent structure of the graph and thus can be seen as a "community" indicator of each node.

As for the implications, several remarks are in order here. First, the theorem shows that the DI w.r.t. the sensitive attribute S of *individual* nodes provides an upper bound on the DI of h when considering the compositional sensitive attribute $S \oplus S'$ defined for pairs of nodes. An immediate consequence of this is that repairing a graph in this case can be done by considering only classes of nodes ($S = 0$ and $S = 1$) rather that classes of edges ($S = S'$ and $S \neq S'$) given by pairs of nodes. Second, the established inequality allows to further provide several results for $\text{BER}(h, \mathbb{V}, S \oplus S')$ that suggest an algorithmic solution for an OT-based repairing procedure. We give this result below.

Corollary 1. *With the assumption from Theorem 1, we have*

$$\textit{BER}(h, \mathbb{V}, S \oplus S') \leq \frac{1}{2} - \frac{\mathbb{P}_1(h)}{2}\left(\frac{1}{\tau} - 1\right),$$

$$\min_{h \in \mathcal{H}} \textit{BER}(h, \mathbb{V}, S \oplus S') = \frac{1}{2}(1 - \frac{1}{2} W_{\mathbf{1}_{\cdot \neq \cdot}}(\gamma_0, \gamma_1))$$

where $W_{\mathbf{1}_{\cdot \neq \cdot}}$ is the Wasserstein distance between true joint distributions γ_0, γ_1 over $\mathbb{V} \times \mathbb{V}$ given $S = 0$ and $S = 1$, respectively equipped with the Hamming cost function.

This corollary provides an upper bound on $\text{BER}(h, \mathbb{V}, S \oplus S')$ in terms of $\text{DI}(h, \mathbb{V}, S)$ and thus allows to control the former by maximizing the latter. Furthermore, the second part of the statement tells us that the balanced error rate of the best edge predictor depends on the divergence between the joint distributions over the nodes given that the sensitive attribute is equal to one of its possible values. This implication allows us to come up with an algorithmic implementation of the repair procedure based on aligning these joint conditional distributions with an OT coupling acting as a mapping. We use this idea as the backbone for our approach and further explore how one can constrain this mapping to ensure individual fairness too.

Remark 2. *Note that in the case of multi-class classification with C classes, $\mathbb{P}(S = i) = 1/C$, while $\mathbb{P}(S \neq S') = (C-1)/C$ implying*

$$(C-1)\mathbb{P}(S = i) = \mathbb{P}(S \neq S').$$

However, this does not change the final result as this factor will appear in both the denominator and numerator (see Supplementary materials, proof of Theorem 1). Finally, if $\mathbb{P}(S = 0) \neq \mathbb{P}(S = 1)$, then

$$\min_{h \in \mathcal{H}} \textit{BER}(h, \mathbb{V}, S \oplus S') = \frac{1}{2}(2(1 - \mathbb{P}_1(h)) - 2\alpha W_{\mathbf{1}_{\cdot \neq \cdot}}(\gamma_0, \gamma_1))$$

where for $\alpha > 0$, $\mathbb{P}(S = 0) = \alpha, \mathbb{P}(S = 1) = 1 - \alpha$ and similar adjustment will have to be made for the first statement as well.

3 ALGORITHMIC IMPLEMENTATION

We now present our algorithm, its multiclass extension and its positioning w.r.t. other related works.

3.1 Group Graph Fairness with OT

The algorithmic idea behind minimizing $W_{\mathbf{1}_{\cdot \neq \cdot}}(\gamma_0, \gamma_1)$ from Corollary 1 is to find an optimal transportation plan between γ_0 and γ_1 and to use it in order to map (push) one distribution on the other [2]. To do this in practice, we consider the adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ associated with the graph $\mathcal{G}$ defined previously and notice that γ_0 (resp. γ_1) corresponds to those rows (nodes) in $\mathcal{A}$ for which $S = 0$ (resp. $S = 1$). Let us denote such submatrices of $\mathcal{A}$ by $\mathcal{A}_0 \in \mathbb{R}^{N_0 \times N}$ and $\mathcal{A}_1 \in \mathbb{R}^{N_1 \times N}$ and assume that they contain N_0 and N_1 rows, respectively. For further convenience, we denote the ratio of each class by π_s, $s \in \{0, 1\}$ with $\pi_s = \frac{N_s}{N}$. We now aim to solve the following OT problem:

$$\min_{\gamma \in \Pi(\frac{\mathbf{1}}{N_0}, \frac{\mathbf{1}}{N_1})} \Omega_{\text{Group}}^{(\gamma, M)}, \quad \Omega_{\text{Group}}^{(\gamma, M)} := \langle \gamma, M \rangle \tag{2}$$

where $\frac{\mathbf{1}}{N_s}$ is a uniform vector with N_s elements, $s \in \{0, 1\}$ and M is the matrix of pairwise distances between the rows in $\mathcal{A}_0$ and $\mathcal{A}_1$, i.e., $M_{ij} = l(a_0^{(i)}, a_1^{(j)})$ for some distance l where $a_0^{(i)}, a_1^{(j)}$ denote the i^{th} and j^{th} rows of $\mathcal{A}_0, \mathcal{A}_1$, respectively. Note that while the Hamming distance $l = \mathbf{1}_{\cdot \neq \cdot}$ is advocated by the theoretical results given above, in practice we use the usual squared Euclidean distance as it tends to give better results and allows a simple closed-form repairing procedure as detailed below.

3.2 Individual Fairness with Laplacian Regularization

Intuitively, one expects from an individually fair mapping to respect the initial relationships between the studied objects when learning their fair representation. This intuition of individually fair mapping was formally captured in the seminal work of [Dwork et al., 2012] and we present its adaptation to graphs below.

[2] For more details about optimal transport problem, we refer the reader to the Supplementary material.

Definition 3. *A mapping $\phi : \mathbb{V} \to \mathbb{V}$ satisfies the (D, d)-Lipschitz property if for every $v, v' \in \mathbb{V}$, and two metrics $D, d : \mathbb{V} \times \mathbb{V} \to \mathbb{R}_+$, we have*

$$D(\phi(v), \phi(v')) \leq d(v, v').$$

In our case, it means that for any two rows $a^{(i)}, a^{(j)} \in \mathcal{A}$, the adjacency matrix repaired by ϕ should preserve the initial similarities between them. We can further rewrite this definition in the form of constraints as follows:

$$D(\phi(a^{(i)}), \phi(a^{(j)}))k(a^{(i)}, a^{(j)}) \leq 1,$$

where $k(a^{(i)}, a^{(j)})$ is a similarity function inversely proportional to $d(a^{(i)}, a^{(j)})$ (eg, as a popular RBF kernel). We can further assume that the mapping $\phi(\cdot)$ is given by a linear transformation with an unknown linear operator matrix Φ. If D is taken to be the squared ℓ_2 norm, we get the following pair-wise constraints for all $a^{(i)}, a^{(j)}$:

$$||\Phi^T a^{(i)} - \Phi^T a^{(j)}||_2^2 k(a^{(i)}, a^{(j)}) \leq 1.$$

One can further incorporate this term into the objective function by introducing a regularization term that aggregates pairwise constraints over all pairs:

$$\min_{\Phi} \sum_{i,j} ||\Phi^T a^{(i)} - \Phi^T a^{(j)}||_2^2 k(a^{(i)}, a^{(j)}).$$

We call this term "individual fairness" term, denote it by $\Omega_{\text{Indiv.}}$ and rewrite it as a graph Laplacian with similarity matrix $(K)_{ij} = k(a^{(i)}, a^{(j)})$:

$$\Omega_{\text{Indiv.}}(\Phi, \mathcal{A}, k) = \text{trace}(\mathcal{A}^T \Phi^T L_K \Phi \mathcal{A}),$$

where $L_K = \text{diag}(K) - K$.

Below, we put all the ingredients together to propose a unified repair procedure that allows us to control the extent of both group and individual fairness when solving the edge prediction task. The group fairness term ensures that edge prediction does not depend on a given protected attribute, and it is the only type of fairness which have been addressed by the related work in the context of fair graph embeddings. Different from this, the individual fairness term ensures that the edge prediction remains consistent with respect to the original graph structure. For instance, let us consider two nodes v_i, v_j having the same sensitive attribute ($S = 0$) and that are similar (in terms of some similarity measure, eg, KNN graph) in the original graph: the first term create edges between these two nodes and nodes with ($S = 1$) in the corrected graph, while the second term ensure that v_i and v_j preserve their original neighbors in the corrected graph. As a result, the proposed objective function allows us to control explicitly the trade-off between individual and group fairness as encouraging one is degrading the other.

3.3 Repairing the Adjacency Matrix

The proposed optimization problem for both group and individually fair adjacency matrix repair takes the following form:

$$\min_{\gamma \in \Pi(\frac{1}{N_0}, \frac{1}{N_1})} \Omega_{\text{Group}}^{(\gamma, M)} + \lambda \sum_{i=0}^{1} \Omega_{\text{Indiv.}}(\Phi_i(\gamma), \mathcal{A}_i, \text{KNN}_3) \quad (3)$$

where $\Phi_0(\gamma) = N_1 \gamma^T$ and $\Phi_1(\gamma) = N_0 \gamma$ are barycentric projections used to push the points of one distribution to those of the other [Ferradans et al., 2013], and KNN_3 is the adjacency matrix of a k-nearest neighbor graph with $k = 3$ calculated from the raw adjacency matrix. Note that we choose to calculate the Laplacian using a KNN graph instead of the raw adjacency matrix as it provides a richer structural information about the graph. Once a solution γ_λ^* to Problem (3) is found, we use it to align the two joint conditional distributions by mapping both $\mathcal{A}_0$ and $\mathcal{A}_1$ on the mid-point of the geodesic path between them [Villani, 2009] as follows:

$$\tilde{\mathcal{A}}_0 = \pi_0 \mathcal{A}_0 + \pi_1 \gamma_\lambda^* \mathcal{A}_1,$$
$$\tilde{\mathcal{A}}_1 = \pi_1 \mathcal{A}_1 + \pi_0 \gamma_\lambda^{*T} \mathcal{A}_0.$$

Note that the closed-form expression given above is valid only when one uses the squared Euclidean distance between the nodes' representations. However, for any arbitrary distance we may obtain the equivalent solution by solving the pre-image problem for each row $\tilde{a}_0^{(i)}$ of $\tilde{\mathcal{A}}_0$, $i = \{1, \ldots, N_0\}$ as follows

$$\tilde{a}_0^{(i)} = \pi_0 a_0^{(i)} + \pi_1 \underset{a \in \mathbb{R}^N}{\text{argmin}} \sum_{j=1}^{N_0} \gamma_\lambda^*(i, j) l\left(a, a_1^{(j)}\right)$$

and similarly for $\tilde{\mathcal{A}}_1$. Such an optimization procedure can be easily parallelized for all $\tilde{a}_0^{(i)}$ with each individual problem solved efficiently by any quasi-Newton method.

Multi-class extension In order to extend our method to the case of $|S| > 2$, i.e., non-binary attributes, we propose to use a recently proposed method for computing the free-support Wasserstein barycenters introduced in [Cuturi and Doucet, 2014, Algorithm 2] and add a Laplacian regularization to it. This leads to the following optimization problem:

$$\tilde{\mathcal{A}}_{\text{bary}} = \underset{\mathcal{A} \in \mathbb{R}^{N \times N}}{\text{argmin}} \frac{1}{|S|} \sum_{i=1}^{|S|} \min_{\gamma_i \in \Pi(\frac{1}{N}, \frac{1}{N_i})} \Omega_{\text{Group}}^{(\gamma_i, M_i)} + \lambda \Omega_{\text{Indiv.}}(N_i \gamma_i^T, \mathcal{A}, \text{KNN}_3),$$

where M_i is the cost matrix between $\mathcal{A}$ and $\mathcal{A}_i$ for $i \in \{1, \ldots, |S|\}$. Contrary to the binary setting of Problem (3), we have only one fairness term here applied to a projection of each sensitive group on the barycenter. Once the optimal

solution for this problem is obtained, we use barycentric mapping to repair each individual submatrix $\mathcal{A}_i$ as follows:

$$\tilde{\mathcal{A}}_i = N_i \gamma_i^{*T} \tilde{\mathcal{A}}_{\text{bary}},$$

where $N_i = |S = i|$ and in general we do not require $N_i = N_j,\ i \neq j$. Note that contrary to the binary case, this mapping projects each matrix $\mathcal{A}_i$ on the barycenter and not on the mid-point of the geodesic path as before.

Complexity Solving OT with Laplacian regularization relies on Frank-Wolfe algorithm where at each iteration a linearization of the loss function under the linear constraints (LP) is solved [Courty et al., 2017]. This results in $O(n^3)$ complexity of each iteration plus the complexity of the node embedding technique used once the repair is done. This cost, however, should not be directly compared with that of other baselines (for instance $O(n^2 d)$ for CNE used by [Buyl and Bie, 2020], on par with node2vec used by [Rahman et al., 2019]) as they provide a solution for one particular embedding technique without individual fairness constraints, while our method is versatile and allows to choose any embedding technique without needing to repeat the repair procedure and deals with individually fair constraints.

Illustration To illustrate the different steps needed to repair an adjacency matrix, we provide in Figure 1 a visual explanation of our proposed approach for a graph having 9 nodes ($\mathcal{A}$). In this graph, the nodes numbered from 1 to 4 belong to the class "Female" ($S = 0$, $\mathcal{A}_0$), while the nodes from 5 to 9 belong to the class "Male" ($S = 1$, $\mathcal{A}_1$). The matrix M calculated in the **first step** and given in Figure 1 contains higher values (darker squares) for the node pairs that are far away from each other in terms of the used distance (e.g., 1 and 5) and lower values for those close to each other (e.g., 3 and 6). The solution obtained in the **second step** highlights the difference between the group fair (EMD, $\lambda = 0$) and both group and individually fair (Laplace, $\lambda > 0$) repair as illustrated by the adjacency matrix $[\tilde{\mathcal{A}}_0 \quad \tilde{\mathcal{A}}_1]$ obtained in the **third step**. Here, we see that group fair repair adds edges that obfuscate the original graph structure both within and across the sensitive groups, while adding individually fair regularization keeps the original structure withing groups almost intact.

3.4 Related Works

Fairness for graphs To the best of our knowledge, very few articles proposed group fair repair schemes for relational data. In [Rahman et al., 2019], the authors proposed FAIRWALK algorithm that produces fairness-aware node embeddings using a modification of the random walk stage of the popular NODE2VEC algorithm [Grover and Leskovec, 2016]. Similarly, DEBAYES [Buyl and Bie, 2020] is an adaptation of Conditional Network Embedding (CNE) [Kang et al., 2019], a Bayesian approach based on integrating prior knowledge through prior distribution for the network. DEBAYES is an adaptation of CNE where the sensitive information is modeled in the prior distribution. Contrary to these two approaches, our proposal is embedding-agnostic, enjoys a theoretical justification and takes individual fairness into account as well. Unlike the previous methods, [Bose and Hamilton, 2019] introduces an adversarial framework that enforces fairness by filtering out the information related to the sensitive attribute from node embeddings obtained with any embedding technique. To obtain a trade-off between fairness and accuracy, the optimization process minimizes alternatively the loss w.r.t. the filtering and its opposite w.r.t. the discriminator. In that case, fairness is defined in terms of invariance, in other words independence, according to the mutual information, between the node embedding and the sensitive attribute. A main drawback of such procedure is that it does not debias relational data given by pairs of nodes but only node embeddings themselves. Consequently, this algorithm seems more designed for fair node classification but not tailored to specifically tackle the fair edge prediction task that takes node tuples as input.

Fairness with OT Several works used the capacity of OT to align probability distributions for fair classification [Gordaliza et al., 2019, Jiang et al., 2019, Zehlike et al., 2020]. The origin of such idea is close to the use of OT in domain adaptation [Courty et al., 2017] where two distributions are aligned using the barycentric mapping. Our work is close to this line of research and extends it in two ways. First, we show that Laplacian regularization previously used in OT for color transfer [Ferradans et al., 2013] and domain adaptation [Courty et al., 2017] leads to an individually fair repair. Second, we use the free-support barycenter algorithm to provide a multi-class version of repair that can deal with sensitive attributes taking non-binary values.

4 EXPERIMENTAL EVALUATION

We investigate the efficiency of our contribution at different levels for both synthetic graphs (see supplementary materials) and three real-world networks [3]. Overall, we aim to answer the following questions: **(Q1) Impact on the structure of the graph**: we investigate the structural changes of the considered graph resulting from the repairing mechanism and the impact of the Laplacian regularization parameter used to promote individual fairness. **(Q2) Impact on node embeddings**: we analyze the impact of our approach on the embeddings obtained from the repaired graph with traditional fairness-unaware node embedding methods. Specifically, we aim to verify whether one can infer the sensitive attribute from the node embedding vectors. **(Q3) Impact on edge prediction**: we study the influence of the

[3] The code reproducing the experimental results is publicly available at `https://github.com/laclauc/FairGraph`.

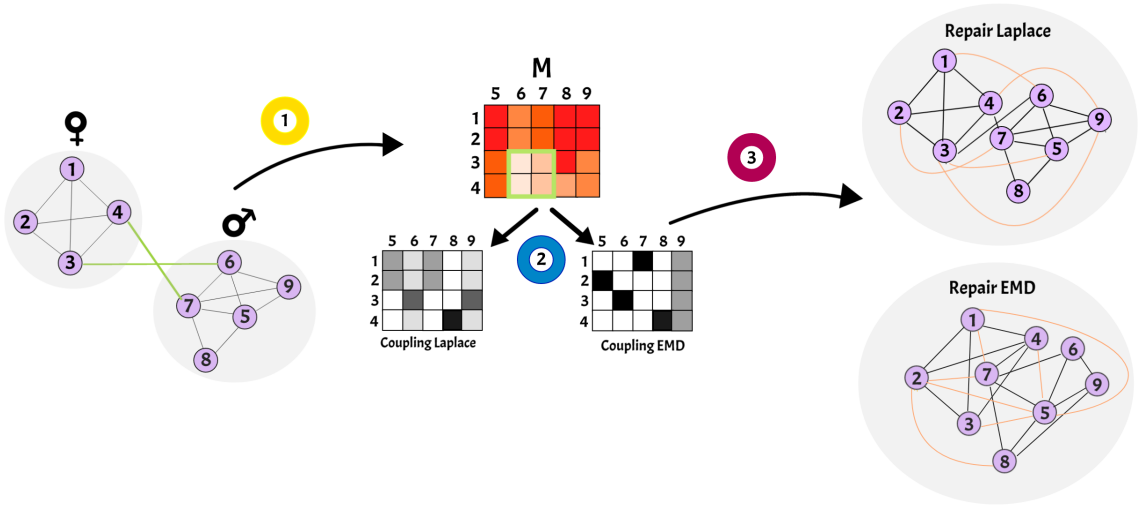

Figure 1: Illustration of the three steps performed to repair the adjacency matrix of a graph.

alterations of the graph on the edge prediction accuracy.

4.1 Baselines

For **(Q2)** and **(Q3)**, we consider two embedding methods, namely NODE2VEC (N2VEC) and CNE, and compare our approach with their fair versions described below.

NODE2VEC and FAIRWALK To evaluate the impact of our approach on node embeddings, we first consider the very popular N2VEC as a baseline. This approach builds a representation of a node in based on its neighborhood following a two-step procedure:

1. Generate a corpus of traces by performing random walks. Formally, denoting by c_i the i-th node in a given walk, the next node is selected among all neighbors of c_i, i.e.,

$$\mathbb{P}(c_{i+1} = v | c_i = u) = \begin{cases} \frac{\pi_{vu}}{C} & \text{if } \{u, v\} \in \mathcal{E} \\ 0 & \text{otherwise,} \end{cases}$$

where π_{uv} denotes the unnormalized transition probability between nodes u and v and C corresponds to a normalization constant. The transition probability is set so as to reflect the neighborhood of u.

2. Use the generated corpus to learn the embedding vectors through a SkipGram architecture that maximizes the log-probability of observing a network neighborhood for a node conditioned on its feature representation:

$$\underset{Z}{\text{argmax}} \prod_{u \in \mathcal{V}} \prod_{v \in N_u} \mathbb{P}(v | Z(u)).$$

We compare our approach with FAIRWALK, a version of N2VEC, designed for fair node embeddings. This latter modifies the transition probability of N2VEC for the generation of unbiased traces. Step (1) of N2VEC becomes

$$\mathbb{P}(c_{i+1} = v | c_i = u) = \begin{cases} \frac{1/k}{|S^k_{N_u}|} & \text{if } S^k_v = 1 \text{ and } \{u, v\} \in \mathcal{E} \\ 0 & \text{otherwise,} \end{cases}$$

where $k = \{1, \cdots, K\}$ denotes the modality of the sensitive attribute S, $S^k_{N_u}$ is the number of nodes in the neighborhood of u belonging to the group k and $S^k_v = 1$ indicates that node v belongs to the k-th group of the sensitive attribute. As a result, each generated random walk has a higher probability to contain nodes of different groups.

CNE and DEBAYES We also study the impact of our algorithm on CNE and compare our algorithm with its recently proposed fair extension DEBAYES. Given a graph $G = (V, E, S)$, CNE finds an embedding $\mathbf{Z}$ by maximizing $P(G|\mathbf{Z}) = \frac{P(\mathbf{Z}|G)P(G)}{P(\mathbf{Z})}$. In our experiments, we use the prior knowledge about the node degree modeled by the prior distribution $P(G)$ expressed by the following constraint:

$$\sum_{v \in V} P((v, v') \in E) = \sum_{v \in V} \mathbb{1}\Big((v, v') \in \mathcal{E}\Big). \quad (4)$$

DEBAYES extends CNE, with a prior to model the sensitive attribute by replacing the constraint (4) with:

$$\sum_{v \in V_s} P((v, v') \in E \mid S(v) = s) = \sum_{v \in V_s} \mathbb{1}\Big((v, v') \in \mathcal{E}\Big),$$

where $V_s = \{v | S(v) = s\}$. With this prior, debiased embeddings containing less information about sensitive information are obtained during the training step. Then, the debiased link predictions are computed using these embeddings and $P(G)$ instead of the biased prior distribution $P(G|S)$.

Random To illustrate the fact that the OT repairing schema is efficient in choosing where to add edges to reduce

Table 1: Statistics for all networks: number of nodes ($|\mathcal{V}|$), number of edges ($|\mathcal{E}|$), type of the protected attribute.

Network	$\mathcal{V}$	$\mathcal{E}$	Type of S	$\lvert S\rvert$
POLBLOGS	$1,490$	$19,090$	binary	2
FACEBOOK	$4,039$	$88,234$	binary	2
DBLP	3,790	6,602	multiclass	5

Table 2: Comparison of assortativity coefficient w.r.t the protected attribute between the original and the repaired graphs. We report the values obtained for $\lambda \in \{0.005, 1, 5\}$

Dataset	Original	EMD	$\text{Lap}_{.005}$	Lap_1	Lap_5
POLBLOGS	.81	.14	.59	.68	.77
FACEBOOK	.09	.04	.04	.05	.06
DBLP	.83	$-.003$	$-.002$	.04	.03

the bias, while maintaining a reasonable accuracy for link prediction, we also compare it with an approach that adds random edges between nodes from different groups for the sensitive attribute. This method is referred to as RANDOM.

For the sake of reproducibility, hyperparameters used in the experiments are provided in the Supplementary materials.

4.2 Datasets

We present the experimental results obtained on three real-world publicly available networks described below. Their key characteristics are summarized in Table 1.

Political Blogs [Adamic and Glance, 2005][4] is a network representing the state of the political blogosphere in the US in 2005. Nodes represent blogs and vertices represent hyperlinks between two blogs. For each node, the sensitive variable indicates the political leaning of the blog.

Snap Facebook [Leskovec and Mcauley, 2012][5] data set consists of ego networks collected through the Facebook app. We use the combined version which contains the aggregated networks of ten individual's Facebook friends list. The gender is the sensitive attribute of each node.

DBLP is a co-authorship network originally built from DBLP, a computer science bibliography database. We use the version proposed by [Buyl and Bie, 2020] where the sensitive attribute corresponds to the continent extracted from authors' affiliation.

4.3 Experimental Results

Q1: impact on the graph structure In order to gain insights on the structural changes resulting from the repairing with our OT-based method, we propose to look at the coefficient of assortativity given the sensitive attribute of the original graph and its repaired versions. We recall that assortativity coefficient takes values in the range between -1 and 1, and that its high values indicate a preference for nodes within the group to be connected with each other w.r.t. a given attribute. Therefore, in our context, one can see the assortativity w.r.t. the protected attribute as a measure of *how much biased the graph structure is*, where values close to 1 indicate a strong bias. From Table 2, we observe that POLBLOGS and DBLP are strongly biased with assortativity coefficients close to 1, while FACEBOOK presents no particular bias w.r.t. its protected attribute as indicated by the assortativity coefficient close to 0. Consequently, we expect 1) to significantly reduce this coefficient for POLBLOGS and DBLP after the repair, and only slightly for FACEBOOK, 2) to preserve the original bias more and more with the increasing strength of the Laplacian regularization. Both these expectations are confirmed by the results provided in Table 2 where the desired behavior is clearly observed.

Q2: impact on node embeddings We proceed by studying the impact of the fair graph repair on the information carried by the node embeddings. In particular, we follow a standard protocol and use 10-fold cross-validated logistic regression to predict the sensitive attribute S from the learned embeddings in order to understand whether applying these latter on a repaired graph maintains the desirable level of fairness. We use the resulting AUC score as a measure of bias, also termed Representation Bias (RB) in [Buyl and Bie, 2020], and recall that in this context the ideal RB corresponds to the optimal value of BER and should be around 0.5. These results are presented in Table 3. From it, we can see that all repairing procedures manage to decrease the RB score successfully and that this decrease is more pronounced for DEBAYES method and, in general, when using CNE embedding. We believe that this embedding is inherently more sensitive to the considered score and we leave the question on why this is the case as an open research avenue.

Q3: Impact on edge prediction For N2VEC- and CNE-based approaches, we follow the protocol of [Rahman et al., 2019] and [Buyl and Bie, 2020] [6], respectively. Our goal here is to identify which approach provides the best trade-off in terms of fairness and prediction accuracy.

To this end, Table 3 reports the AUC for link prediction, the disparate impact (DI) and the consistency (Cons) scores, where the two latter are measures of group and individual fairness (see [Zemel et al., 2013]), respectively. From these results, we make the following observations. First, we recall that POLBLOGS and DBLP present a true challenge for fair edge prediction as the original results obtained with classical embeddings approaches are characterized by a low DI and a high RB score. This is contrary to FACEBOOK graph em-

[4] www-personal.umich.edu/~mejn/netdata/

[5] snap.stanford.edu/data/ego-Facebook.html

[6] github.com/aida-ugent/DeBayes

Table 3: AUC score for link prediction, Representation Bias (RB), Disparate Impact (DI) and Consistency (Cons.). For the Laplacian, results corresponds to the regularization parameter set to 1. RANDOM results are N2Vec-based.

	Metric	N2VEC	FAIRWALK	N2VEC$^{\text{EMD}}$	N2VEC$^{\text{LAP}}$	CNE	DEBAYES	CNE$^{\text{EMD}}$	CNE$^{\text{LAP}}$	RANDOM
POLBLOGS	AUC	$\mathbf{.75 \pm .01}$	$\mathbf{.75 \pm .01}$	$.66 \pm .01$	$.73 \pm .01$	$\mathbf{.93 \pm .01}$	$.88 \pm .01$	$.86 \pm .01$	$.91 \pm .02$	$.53 \pm .01$
	RB	$.97 \pm .01$	$.96 \pm .01$	$\mathbf{.78 \pm .01}$	$.94 \pm .01$	$.97 \pm .01$	$\mathbf{.64 \pm .04}$	$.73 \pm .03$	$.94 \pm .04$	$.63 \pm .01$
	DI	$.10 \pm .02$	$.20 \pm .01$	$.54 \pm .07$	$.25 \pm .02$	$.03 \pm .02$	$.53 \pm .05$	$\mathbf{.83 \pm .05}$	$.19 \pm .03$	$.43 \pm .02$
	Cons.	$.75 \pm .02$	$.73 \pm .01$	$.77 \pm .10$	$\mathbf{.91 \pm .01}$	$.89 \pm .01$	$.89 \pm .01$	$.90 \pm .01$	$\mathbf{.93 \pm .01}$	$.90 \pm .04$
FACEBOOK	AUC	$\mathbf{.98 \pm .01}$	$.85 \pm .00$	$.96 \pm .00$	$.96 \pm .00$	$\mathbf{.99 \pm .01}$	$\mathbf{.99 \pm .03}$	$\mathbf{.99 \pm .01}$	$.98 \pm .01$	$.49 \pm .04$
	RB	$.64 \pm .01$	$\mathbf{.61 \pm .01}$	$\mathbf{.61 \pm .00}$	$.63 \pm .00$	$.58 \pm .02$	$.57 \pm .02$	$\mathbf{.54 \pm .03}$	$.58 \pm .02$	$.56 \pm .02$
	DI	$.80 \pm .01$	$\mathbf{.83 \pm .00}$	$.80 \pm .01$	$.80 \pm .00$	$.93 \pm .03$	$.91 \pm .03$	$.98 \pm .01$	$\mathbf{.99 \pm .05}$	$.84 \pm .02$
	Cons.	$.96 \pm .00$	$.94 \pm .00$	$.96 \pm .01$	$.96 \pm .00$	$\mathbf{.97 \pm .01}$	$.96 \pm .00$	$\mathbf{.97 \pm .01}$	$\mathbf{.97 \pm .00}$	$.89 \pm .01$
DBLP	AUC	$\mathbf{.98 \pm .01}$	$\mathbf{.98 \pm .01}$	$.78 \pm .03$	$.81 \pm .04$	$\mathbf{.98 \pm .01}$	$\mathbf{.98 \pm .01}$	$.77 \pm .03$	$.82 \pm .05$	$.54 \pm .01$
	RB	$.77 \pm .00$	$.77 \pm .01$	$\mathbf{.58 \pm .04}$	$\mathbf{.58 \pm .02}$	$.55 \pm .02$	$\mathbf{.51 \pm .02}$	$.52 \pm .01$	$\mathbf{.51 \pm .02}$	$.59 \pm .01$
	DI	$.14 \pm .01$	$.14 \pm .01$	$\mathbf{1.26 \pm .04}$	$1.02 \pm .05$	$.03 \pm .01$	$.04 \pm .01$	$\mathbf{1.29 \pm .04}$	$.98 \pm .05$	$.43 \pm .03$
	Cons.	$.91 \pm .01$	$.91 \pm .01$	$.93 \pm .02$	$\mathbf{.95 \pm .01}$	$.91 \pm .01$	$.90 \pm .02$	$.94 \pm .01$	$\mathbf{.97 \pm .01}$	$.86 \pm .01$

bedding for which we obtain a high DI value indicating that it requires no particular repair. The goals of the repairing methods for each of these data sets are thus quite different: for POLBLOGS and DBLP we would like to increase the DI value and reduce the predictability of the sensitive attribute by trading off as little of the edge prediction AUC as possible, while for FACEBOOK the algorithms should mainly maintain the existing graph structure and not hinder the edge prediction with unnecessary repairing. From the obtained results, we first note that all fairness-aware methods increase DI score compared to the original one on POLBLOGS and FACEBOOK, while maintaining a decent prediction accuracy well-above the random guessing threshold observed in the case of the RANDOM repair. On the other hand, on DBLP dataset only our approach improves fairness scores, but this comes at a price of a drop in terms of the performance. Most likely, this is due to the imbalance between different sensitives groups that hinders the performance of OT. As for the consistency, only Laplacian regularization significantly improves this criterion, while it remains almost unchanged for other baselines after the repair. While different repair methods have their distinct strong sides making it difficult to choose the "best" one, we note that our proposed approach is versatile and allows to explicitly control the trade-off between the fairness and prediction accuracy and to be used with different embeddings. Finally, Figure 2 shows the results for different regularization parameters for the Laplacian OT on POLBLOGS. Once can see that as the value of the regularization parameter increase, the group fairness metric (DI) decreases while the individual fairness metric (Cons.) increases.

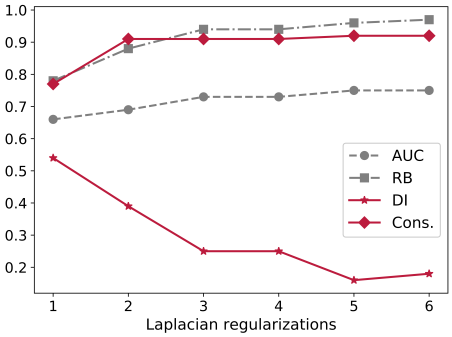

Figure 2: Impact of the Laplacian regularization on the different metrics for POLBLOGS.

5 CONCLUSION

In this paper we addressed an important problem of fair edge prediction in graphs. Contrary to fair classification, fair edge prediction in graphs has received a very limited amount of attention from the research community and has mainly been solved using heuristic embedding-dependent procedures and only in group fairness context. To bridge this gap, we provide a first embedding-agnostic repair procedure for the adjacency matrix of a graph with both group and individual fairness constraints. We show through extensive experimental evaluations that our approach provides a flexibility of choosing explicitly to which extent one wants to ensure group and individually fair constraints.

Further research directions of this work are many. First, we would like to study the impact of different embedding techniques on the bias in the adjacency matrix of a graph as empirical evidence suggests that some embedding techniques reinforce the bias in the data making it even more apparent. We also plan to use a recent theoretical analysis of popular node embedding methods [Qiu et al., 2018] to provably show their effect on the correlation between the estimated output and the sensitive attribute.

Acknowledgement

This work has been supported by IDEXLYON ACADEMICS Project ANR-16-IDEX-0005 of the French National Research Agency.

References

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *3rd International Workshop on Link Discovery*, page 36–43.

[Agarwal et al., 2018] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *ICML*, pages 60–69.

[Bose and Hamilton, 2019] Bose, A. and Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. In *ICML*, pages 715–724.

[Buyl and Bie, 2020] Buyl, M. and Bie, T. D. (2020). Debayes: a bayesian method for debiasing network embeddings. In *Proceedings ICML*, pages 2537–2546.

[Calmon et al., 2017] Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *NeurIPS*, pages 3992–4001.

[Chiappa, 2019] Chiappa, S. (2019). Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808.

[Corbett-Davies et al., 2017] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *KDD*, page 797–806.

[Courty et al., 2017] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865.

[Cuturi and Doucet, 2014] Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In *ICML*, pages 685–693.

[Donini et al., 2018] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In *NeurIPS*, pages 2791–2801.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In *ITCS*, pages 214–226.

[Edwards and Storkey, 2016] Edwards, H. and Storkey, A. J. (2016). Censoring representations with an adversary. In *ICLR*.

[Feldman et al., 2015a] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015a). Certifying and removing disparate impact. In *SIGKDD*, page 259–268.

[Feldman et al., 2015b] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015b). Certifying and removing disparate impact. In *SIGKDD*, page 259–268.

[Ferradans et al., 2013] Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., and Aujol, J.-F. (2013). Regularized discrete optimal transport. In Kuijper, A., Bredies, K., Pock, T., and Bischof, H., editors, *Scale Space and Variational Methods in Computer Vision*, pages 428–439.

[Gordaliza et al., 2019] Gordaliza, P., del Barrio, E., Gamboa, F., and Loubes, J. (2019). Obtaining fairness using optimal transport theory. In *ICML*, pages 2357–2365.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *SIGKDD*, pages 855–864.

[Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323.

[Jiang et al., 2019] Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). Wasserstein fair classification. In *UAI*, page 315.

[Johndrow and Lum, 2019] Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

[Kang et al., 2019] Kang, B., Lijffijt, J., and Bie, T. D. (2019). Conditional network embeddings. In *7th International Conference on Learning Representations, ICLR*.

[Kusner et al., 2017] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *NeurIPS*, pages 4066–4076.

[Leskovec and Mcauley, 2012] Leskovec, J. and Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *NIPS*, pages 539–547. Curran Associates, Inc.

[Louizos et al., 2016] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. (2016). The variational fair autoencoder. In *ICLR*.

[Madras et al., 2018] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *ICML*, pages 3384–3393.

[Qiu et al., 2018] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. (2018). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *WSDM*, pages 459–467.

[Rahman et al., 2019] Rahman, T., Surma, B., Backes, M., and Zhang, Y. (2019). Fairwalk: Towards Fair Graph Embedding. In *IJCAI*, pages 3289–3295.

[Stoica et al., 2018] Stoica, A.-A., Riederer, C., and Chaintreau, A. (2018). Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *WWW*, page 923–932.

[Villani, 2009] Villani, C. (2009). *Optimal transport : old and new*. Springer, Berlin.

[Zafar et al., 2017a] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, page 1171–1180.

[Zafar et al., 2017b] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017b). Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, pages 962–970.

[Zehlike et al., 2020] Zehlike, M., Hacker, P., and Wiedemann, E. (2020). Matching code and law: achieving algorithmic fairness with optimal transport. *Data Min. Knowl. Discov.*, 34(1):163–200.

[Zemel et al., 2013] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *ICML*, pages 325–333.