Obtaining Dyadic Fairness by Optimal Transport

Moyi Yang[1*], Junjie Sheng[1*], Xiangfeng Wang[1†], Wenyan Liu[2], Bo Jin[1], Jun Wang[1], Hongyuan Zha[3]

[1] School of Computer Science and Technology, East China Normal University
[2] Software Engineering Institute, East China Normal University
[3] School of Data Science and AIRS, The Chinese University of Hong Kong(Shenzhen)

Abstract

Fairness has been taken as a critical metric on machine learning models. Many works studying how to obtain fairness for different tasks emerge. This paper considers obtaining fairness for link prediction tasks, measured by dyadic fairness. We propose a pre-processing methodology to obtain dyadic fairness through data repairing and optimal transport. To obtain dyadic fairness with satisfying flexibility and unambiguity requirements, we transform the dyadic repairing to the conditional distribution alignment problem based on optimal transport and obtain theoretical results on the connection between the proposed alignment and dyadic fairness. An optimal transport-based dyadic fairness algorithm is proposed for graph link prediction. The proposed algorithm shows superior results on obtaining fairness compared with the other pre-processing methods on two benchmark graph datasets.

Introduction

Machine learning has been widely adopted in the real world. Although remarkable results were achieved on predicting and decision-making scenarios, unexpected bias often happens (Stoica, Riederer, and Chaintreau 2018; Besse et al. 2018; Friedler et al. 2019). For instance, the famous new media company ProPublica found that black defendants were far more likely than white defendants to be incorrectly judged as having a higher risk of recidivism in the COMPAS system (Angwin et al. 2016). Amazon found that the AI hiring tool they developed to automate the hiring process is biased against women (Lauret 2019). Many works emerge to design algorithms to avoid such biases and obtain *fair* machine learning models further.

This work considers achieving fairness in link prediction tasks. The link prediction task is a common but essential problem in modern machine learning applications, not limited to recommendation systems and knowledge graph completion. The main goal is to predict whether the link between two nodes exists in a graph. Many existing algorithms, e.g., Node2Vec (Grover and Leskovec 2016) and GCN (Kipf and Welling 2017), have been proposed to solve the link prediction task with superior performance in many scenarios. However, the collected dataset for the model training procedure usually has various unexpected biases. This will lead to unfair results of the obtained link prediction model. For instance, when collecting data from social media platforms, early works highlighted that users were more interested in conversing with others of the same race and gender (Khanam, Srivastava, and Mago 2020). Link prediction models trained based on such unfair data will also tend to predict the existence of links between nodes with the same sensitive information, thus unfairly disadvantaging some users. To formally define such an unfair phenomenon, (Li et al. 2021; Masrour et al. 2020) introduced the dyadic fairness for link prediction of graphs. The dyadic fairness criterion expects the prediction results to be independent of the sensitive attributes from the given two nodes.

Recently, several existing works have been proposed to achieve dyadic fairness in link prediction tasks, which can be roughly divided into three categories: 1) in-processing scheme (Li et al. 2021) considers modifying the learning algorithm to eliminate bias; 2) post-processing scheme (Masrour et al. 2020) attempts to de-bias directly the model's output after training; 3) pre-processing scheme (Spinelli et al. 2021) aims to repair the graph data before the training procedure further the link prediction results can satisfy dyadic fairness. In this paper, our algorithm is established based on the pre-processing scheme. Compared with the other two schemes, the pre-processing scheme can be considered the most flexible fairness intervention (Nielsen 2020). Suppose the discrimination is removed from the data during the pre-processing stage. In that case, the processed data can train arbitrary downstream tasks with no need to be concerned about the fairness issue. Few works have studied obtaining dyadic fairness through pre-processing. FairDrop (Spinelli et al. 2021) proposed a heuristic repairing method that masks out edges based on the dyadic sensitive attributes. It is easy to implement but without a theoretical guarantee on achieving fairness. To design a theoretically sound pre-processing scheme, FairEdge (Laclau et al. 2021) firstly adopts the Optimal Transport (OT) theory to justify whether dyadic fairness can be obtained through a repairing scheme. FairEdge focuses on the plain graph (the node has no attribute) and proposes to repair adjacency information distributions (conditioned on sensitive attribute) to the corresponding Wasser-

*These authors contributed equally.
†Corresponding author.

stein barycenter. The dyadic fairness is obtained once the adjacency information distributions are all repaired as the Wasserstein barycenter. However, as for the practically common attributed graph, these existing methods can not guarantee fairness. This is because node i's feature is influenced by both its adjacency information and the multi-hop neighbors' features. Even if the discrimination from adjacency information is removed, the nodes' attributes can still introduce discrimination to the data. Thus, how to repair data to obtain dyadic fairness on the attributed graph is still under-exploited.

This paper aims to exploit how to obtain dyadic fairness under the pre-processing scheme. Except for the dyadic fairness, two requirements are concluded to be satisfied when repairing: 1) flexibility: the repaired data needs to hold dyadic fairness for a wide range of embedding functions and predictors due to both of them being unknown during pre-processing; 2) unambiguity: each node's attribute and adjacency information should be determined without ambiguity after repairing. To properly obtain the dyadic fairness satisfying these two requirements, we propose to align the conditional distributions (node's attribute-adjacency distribution conditioned on the sensitive features) to the Wasserstein barycenter based on optimal transport. We can theoretically prove that dyadic fairness with flexibility and unambiguity is obtained when the conditional distributions are repaired to the same distribution. To make it practical, we derive an optimal transport-based repairing algorithm that aligns the conditional distributions to their Wasserstein barycenter. Experiments on the CORA and CiteSeer datasets, our proposed algorithm outperforms the state-of-the-art pre-processing methods on many fairness metrics.

Related Works

Fairness in link prediction

Although link prediction is a well-researched problem in applications related to graph data (Al Hasan et al. 2006; Masrour et al. 2015), since fairness in graph-structured data is a relatively new research topic, only a few works have investigated fairness issues in link prediction. In Spinelli et al. (2021), the authors proposed a biased dropout strategy that forces the graph topology to reduce the homophily of sensitive attributes. Meanwhile, to measure the improvements for the link prediction, they also defined a novel group-based fairness metric on dyadic level groups. In contrast, Masrour et al. (2020) considered generating more heterogeneous links to alleviate the filter bubble problem. In addition, they further presented a novel framework that combines adversarial network representation learning with supervised link prediction. Following the idea of adversarially removing unfair effects, Li et al. (2021) proposes the algorithm **FairAdj** to empirically learn a fair adjacency matrix with proper graph structural constraints for fair link prediction to ensure predictive accuracy as much as possible simultaneously. Most similar to our method, Laclau et al. (2021) formulated the problem of fair edge prediction and proposed an embedding-agnostic repairing procedure for the adjacency matrix with a trade-off between the group and individual fairness. However, they still ignore the node attributes, which impact both the predicting and fairness performance.

Fairness with Optimal Transport

In the context of ML fairness, several works have proposed using the capacity of optimal transport to align probability distributions, overcoming the limitation of most approaches that approximate fairness by imposing constraints on the lower order moments. Along with this motivation, most of the existing methods consider using optimal transport theory to match distributions corresponding to different sensitive attributes in the model input space or the model output space, which corresponds to pre-processing (Gordaliza et al. 2019; Feldman et al. 2015; Laclau et al. 2021) and post-processing (Jiang et al. 2019; Chzhen et al. 2020) methods, respectively. In addition, the in-processing (Jiang et al. 2019; Chiappa and Pacchiano 2021) methods based on optimal transport achieve fairness by imposing constraints in terms of the Wasserstein distance in the objective function.

Dyadic Fairness in Link Prediction

In this section, we formulate the dyadic fairness in the link prediction task and define two metrics (dyadic disparate impact and dyadic balanced error rate) to quantify the dyadic fairness. Then we conclude two desired properties for our repairing algorithm that try to obtain dyadic fairness, i.e., flexibility and unambiguity. We further theoretically discuss how these properties can be achieved and prove that aligning conditional attribute and adjacency distributions to the same distribution can obtain dyadic fairness with these properties.

Problem setup

Given the graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} := \{v_1, \ldots, v_N\}$ be the node set of the graph and $\mathcal{E} := \{e_1, \ldots, e_N\}$ be the edge set of the graph. Each node v_i be endowed with a vector $\mathbf{x}_i \in \mathbb{R}^M$ of attributes. Each edge e_i is the ith row of a non-negative adjacency matrix $A \in \{0,1\}^{N \times N}$ which summarizes the connectivity in the graph. $A_{ij} = 1$ if nodes v_i and v_j is connected; otherwise, $A_{ij} = 0$. The link prediction model usually identify whether the link between two nodes (i, j) exist based on their node representations, i.e., $g : \boldsymbol{z}_i \times \boldsymbol{z}_j \mapsto \{0, 1\}$ where the $\boldsymbol{z}_i$ denotes the node i's representation. The $\boldsymbol{z}_i$ is usually obtained by random walk or graph convolution on the whole graph: $\boldsymbol{z}_i = f(\mathcal{G})[i]$ where the $f : \mathcal{G} \mapsto \mathbb{R}^{N \times d}$ is called the embedding function. The d is the node representation's dimension, and the f can be Node2Vec, GCN, GAT, etc. The link predictor g takes two nodes' representations with the node representations and directly outputs whether a link between them exists. To study fairness on link prediction tasks, we assume that all nodes have one sensitive feature $S : \mathcal{V} \to \mathcal{S}$. We also take the binary sensitive feature $\mathcal{S} = \{0, 1\}$ first and let $S(i)$ denote the sensitive feature of node i. The binary sensitive feature will be relaxed later. Before proposing our algorithm, we make two assumptions:

1. $\mathbb{P}(S \oplus S' = 1) = \mathbb{P}(S \oplus S' = 0) = \frac{1}{2}$, which is based on the fact that each node has an equal chance of being sampled regardless of its sensitive attribute value. For

instance, $\mathbb{P}(S = man) = \mathbb{P}(S = woman)$ is always an equivalence relationship independent of the sampling process and the obtained graph data itself;

2. $\mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 1|S(u) \oplus S(v) = 0) \geq \mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 1|S(u) \oplus S(v) = 1)$, which illustrates that the classifier we consider here will tend to predict the existence of links between nodes with the same sensitive attributes.

For link prediction problems, the main unfairness phenomenon is assigning high link probability to nodes with the same sensitive feature while assigning low probability for nodes with different sensitive features. For example, a user may be treated unfairly on social platforms because they are rarely recommended to users of a different gender or race. This unfairness can be defined mathematically as in (Li et al. 2021).

Definition 1 ***Dyadic Fairness:*** *A link predictor g obtains dyadic fairness if for node representation $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$*

$$\begin{aligned}\mathbb{P}(g(\boldsymbol{z}_i, \boldsymbol{z}_j)|S(i) \oplus S(j) = 1)\\ = \mathbb{P}(g(\boldsymbol{z}_i, \boldsymbol{z}_j)|S(i) \oplus S(j) = 0).\end{aligned} \tag{1}$$

When the link predictor predicts the link between two nodes in the same proportion regardless of whether they have the same sensitive attributes, the predictor is denoted as obtaining dyadic fairness. Actually, the dyadic fairness described in (1) is difficult to achieve in real data. Therefore, in order to better quantify the fairness, we adopt two other essential fairness metrics, namely dyadic disparate impact (DDI) and dyadic balanced error rate (DBER), which are defined as follows:

Definition 2 (Dyadic Disparate Impact) *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a function $g(\boldsymbol{z}_u, \boldsymbol{z}_v) : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$, we define the link prediction function g has Disparate Impact at level $\tau \in (0, 1]$ on $S(u) \oplus S(v)$ w.r.t.$\boldsymbol{Z}$ if:*

$$\mathrm{DDI}(g, \boldsymbol{Z}, \mathcal{S}) = \frac{\mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 1|S(u) \oplus S(v) = 1)}{\mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 1|S(u) \oplus S(v) = 0)} \leq \tau. \tag{2}$$

As defined, Dyadic Disparate Impact measures the fairness level of the predictor. The larger the value of τ, the fairer it is. Ideally, when the value of τ reaches 1, it means that the link predictor achieves dyadic fairness.

Definition 3 (Dyadic Balanced Error Rate) *For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a function $g(\boldsymbol{z}_u, \boldsymbol{z}_v) : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$, we define the dyadic balanced error rate of the predictor g as the average class-conditional error:*

$$\begin{aligned}\mathrm{DBER}(g, \boldsymbol{Z}, \mathcal{S}) = \frac{1}{2}(&\mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 0|S(u) \oplus S(v) = 1)\\ &+ \mathbb{P}(g(\boldsymbol{z}_u, \boldsymbol{z}_v) = 1|S(u) \oplus S(v) = 0)).\end{aligned} \tag{3}$$

As defined, Dyadic Balanced Error Rate measures the general misclassification error of sensitive attributes by g in the particular case of $\mathbb{P}(S \oplus S' = 1) = \mathbb{P}(S \oplus S' = 0) = \frac{1}{2}$. Then the DBER is guaranteed to be smaller than $\frac{1}{2}$. With the larger DBER, the data and predictor g will be fairer. If the DBER is $\frac{1}{2}$, the DDI will be 1, and dyadic fairness is achieved.

Obtaining Dyadic Fairness

This paper considers establishing dyadic fairness through pre-processing the graph data. Due to the pre-processing nature, our repairing procedure has no idea what the embedding function f and predictor g are. We thus need to ensure the repaired data can achieve the dyadic fairness for arbitrary embedding function and predictor. These are referred to as the **flexibility** requirements. Moreover, another straightforward requirement also needs to be considered: **unambiguity**. After repairing, each node's attribute and adjacency information should be determined without ambiguity.

Firstly, to obtain the wide applicability on predictors (flexibility), we consider the DBER of the most unfair predictor with the repaired data, i.e.,

$$\boldsymbol{Z}^* = \arg\max_{\boldsymbol{Z}} \min_{g} \ \mathrm{DBER}\left(g, \boldsymbol{Z}, \mathcal{S}\right). \tag{4}$$

Suppose the repaired data $\boldsymbol{Z}^*$ ensures high DBER under the most unfair predictor. In that case, it obtains dyadic fairness with wide applicability on predictors. Although this makes the problem a bi-level optimization one, the closed form of g can be obtained with the Bayes formula as in (Gordaliza et al. (2019)).

Theorem 1 *The smallest DBER for data $\boldsymbol{Z}$ is equal to:*

$$\min_{g} \ \mathrm{DBER}\left(g, \boldsymbol{Z}, \mathcal{S}\right) = \frac{1}{2}\left(1 - \frac{1}{2} W_{1.\neq}\left(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1\right)\right), \tag{5}$$

where $W_{1.\neq}$ is the Wasserstein distance between the conditional joint distributions of node representation with Hamming cost function. The $\hat{\boldsymbol{\gamma}}_0$ and $\hat{\boldsymbol{\gamma}}_1$ are over $\boldsymbol{Z} \times \boldsymbol{Z}$ given $S(u) \oplus S(v) = 0$ and $S(u) \oplus S(v) = 1$.

As showing the theorem, the dyadic balanced error rate of the most unfair predictor depends on the Wasserstein distance between the two conditional dyadic node representation distribution $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1)$. When $W_{1.\neq}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1) = 0$, which means that the two conditional distributions are identical, i.e.,

$$\mathbb{P}(\boldsymbol{z}_u, \boldsymbol{z}_v|S(u) \oplus S(v) = 1) = \mathbb{P}(\boldsymbol{z}_u, \boldsymbol{z}_v|S(u) \oplus S(v) = 0). \tag{6}$$

The DBER can achieve the optimal $\frac{1}{2}$ and the $\boldsymbol{Z}$ are taken as dyadic fairness on the sensitive feature S. Thus ensuring the (6) makes the repaired data achieve dyadic fairness with wide applicability on arbitrary predictor g.

One straightforward repairing is directly moving the two conditional distributions to the same distribution. However node i's representation $\boldsymbol{z}_i$ often occurs multi times in the $\hat{\boldsymbol{\gamma}}_0$ and $\hat{\boldsymbol{\gamma}}_1$. When repairing the $\hat{\boldsymbol{\gamma}}_0$ and $\hat{\boldsymbol{\gamma}}_1$, the $\boldsymbol{z}_i$ will be likely to assign multi values. For example, as shown in Figure 1, the direct repairing leads to ambiguity on the A's attribute. To achieve the unambiguity repairing, we propose the following proposition.

Proposition 1 *The dyadic fairness* (6) *is satisfied if and only if the following equation satisfied:*

$$\mathbb{P}(\boldsymbol{z}_u|S(u) = 0) = \mathbb{P}(\boldsymbol{z}_v|S(v) = 1). \tag{7}$$

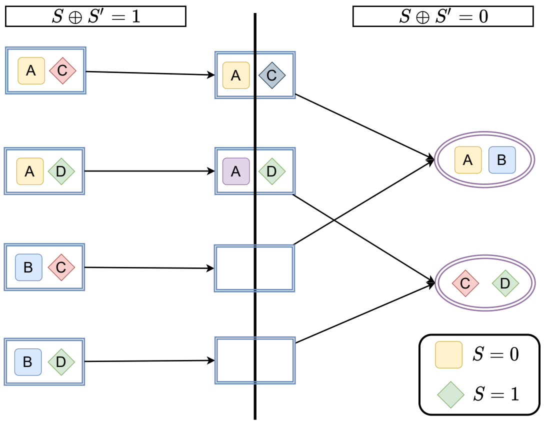


Figure 1: The ambiguity illustration of dyadic repairing. The pair (A, C) and (A, D) are repaired to the two pairs in the black line. The A's original attribute is yellow, while in the repaired data, it has multiple values ('yellow' and 'purple'), which leads to ambiguity.

Proof: For the sufficient part, if the (7) is satisfied, then for arbitrary representation a and b, the

$$\begin{aligned} &\mathbb{P}(\boldsymbol{z}_u = a, \boldsymbol{z}_v = b | S(u) \oplus S(v) = 0) \\ =&\textstyle\sum_{i=0}^{1} \mathbb{P}(\boldsymbol{z}_u = a | S(u) = i) \times \mathbb{P}(\boldsymbol{z}_v = b \mid S(v) = i) \\ =&\textstyle\sum_{i=0}^{1} \mathbb{P}(\boldsymbol{z}_u = a | S(u) = i) \times \mathbb{P}(\boldsymbol{z}_v = b \mid S(v) = 1 - i) \\ =&\mathbb{P}(\boldsymbol{z}_u = a, \boldsymbol{z}_v = b | S(u) \oplus S(v) = 1). \end{aligned}$$

As follows, the (6) is satisfied accordingly. For the necessary part, it can be easily proved by contradiction. The above proposition implies that a fair node representation is sufficient for achieving dyadic fairness in the optimal case. Repairing based on (7) allows us to obtain dyadic fairness and unambiguity requirement due to the node's representation being only repaired once.

After achieving wide applicability on predictors and unambiguity, we consider obtaining the wide applicability on embedding function f. The embedding function takes the whole graph $\mathcal{G}$ as input and outputs the node representation $\boldsymbol{z}_i$ based on the graph.

Proposition 2 *For any node u, v in the graph $\mathcal{G}$, if they have the same node attributes and adjacency status, i.e,*

$$\mathbf{x}_u = \mathbf{x}_v \quad \text{and} \quad \mathbf{e}_u = \mathbf{e}_v, \tag{8}$$

then for any embedding function f, $f(\mathcal{G})[u] = f(\mathcal{G})[v]$. Here the $\mathbf{x}_u$, $\mathbf{x}_v$ denote the attribute of node u and node v, respectively. And the $\mathbf{e}_u$, $\mathbf{e}_v$ denote the 1-hop adjacency information, which means the local topology structure of node u and node v.

This proposition enables us to transform (7) to the following one:

$$\mathbb{P}(\mathbf{x}_u, \mathbf{e}_u | S(u) = 0) = \mathbb{P}(\mathbf{x}_v, \mathbf{e}_v | S(v) = 1). \tag{9}$$

Suppose the above (9) can be obtained. In that case, the dyadic fairness (1) can further be satisfied for arbitrary predictors. In the following, we propose an algorithm to make (9) happen.

Algorithmic Framework

This section introduces a practical algorithm to achieve dyadic fairness on link prediction tasks based on optimal transport theory. It can be easily extended to multi-valued sensitive attributes problems, which relaxes the binary sensitive value constraint.

Dyadic fairness with OT

In order to achieve the dyadic fairness through (9), we first represent the graph $\mathcal{G}$ as a matrix $\mathbb{R}^{N\times(d+N)}$ where each row represent one node's attribute ($\mathbf{x}_u$) and adjacency information ($\mathbf{e}_u$). According to the sensitive feature of each node, we further split the $\mathcal{G}$ to $\mathcal{G}_0 \in \mathbb{R}^{N_0\times(d+N)}$ and $\mathcal{G}_1 \in \mathbb{R}^{N_1\times(d+N)}$ where the N_0 and N_1 are the number of nodes with $S = 0$ and $S = 1$. To bridge it with the optimal transport theory, we assume graph $\mathcal{G}_0$ and $\mathcal{G}_1$ form uniform distributions $\hat{\boldsymbol{\gamma}}_0$ and $\hat{\boldsymbol{\gamma}}_1$. Then our goal can be explicitly described as $\min_{\mathcal{G}} W_{1,\neq}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1)$. To achieve that goal, we transform the two distributions to their Wasserstein barycenter:

$$\boldsymbol{\Gamma}^* = \min_{\boldsymbol{\Gamma}\in\Pi(\frac{1}{N_0},\frac{1}{N_1})} \langle \boldsymbol{\Gamma}, \mathbf{C} \rangle, \tag{10}$$

where N_s is the number of nodes in the graph and $\frac{1}{N_s}$ is the uniform vector with N_s elements, i.e., $s \in \{0, 1\}$.

Cost matrix Considering the distribution $\hat{\boldsymbol{\gamma}}_0$ and $\hat{\boldsymbol{\gamma}}_1$ encodes two important parts of information about the node, i.e., feature $\mathbf{x}_u$ and local topology structure $\mathbf{e}_u$, our cost matrix $\mathbf{C}$ will consist of two components with hyperparameter η as a trade-off between the feature term and structure term.

$$\mathbf{C}_{ij} = \eta||\mathbf{x}_i, \mathbf{x}_j||_2^2 + (1-\eta)||\mathbf{e}_i, \mathbf{e}_j||_2^2. \tag{11}$$

Note that although the Hamming distance is used in the above theoretical results, in practice, to get better results, we use the squared Euclidean distance.

Repairing procedure Once we obtained the optimal transport plan $\boldsymbol{\Gamma}^*$, we apply it to repair the node feature, and adjacency information by mapping both $\mathcal{G}_0 \in \mathbb{R}^{N_0\times(N+d)}$ and $\mathcal{G}_1 \in \mathbb{R}^{N_1\times(N+d)}$ to the mid-point of the geodesic path between them (Villani 2009) as follows:

$$\begin{cases} \tilde{\mathcal{G}}_0 = \pi_0\mathcal{G}_0 + \pi_1\boldsymbol{\Gamma}^*\mathcal{G}_1, \\ \tilde{\mathcal{G}}_1 = \pi_1\mathcal{G}_1 + \pi_0{\boldsymbol{\Gamma}^*}^\top\mathcal{G}_0. \end{cases} \tag{12}$$

Multi-class extension In order to extend our approach to the case of the non-binary sensitive attribute, it would be necessary to compute the Wasserstein barycenter of the conditional distributions.

$$\bar{\mathcal{G}} = \underset{\bar{\mathcal{G}}\in\mathbb{R}^{N\times(N+d)}}{\operatorname{argmin}} \frac{1}{|S|}\sum_{k=1}^{|S|} \min_{\boldsymbol{\Gamma}_k\in\Pi\left(\frac{1}{N},\frac{1}{N_k}\right)} \langle \boldsymbol{\Gamma}_k, \mathbf{C}_k \rangle, \tag{13}$$

where $\mathbf{C}_k$ is the cost matrix between $\mathcal{G}_k$ and $\bar{\mathcal{G}}$. Once we got the Wasserstein barycenter $\bar{\mathcal{G}}$ and the optimal transport plan between the Wasserstein barycenter and each sensitive attribute group, i.e.,$\boldsymbol{\Gamma}_k$, we will repair $\mathcal{G}_k$ as follows

$$\tilde{\mathcal{G}}_k = N_k{\boldsymbol{\Gamma}_k^*}^\top\bar{\mathcal{G}}. \tag{14}$$

Algorithm 1: Dyadic fairness with OT

1: Initialize η and $\mathbf{\Gamma}^0 \in \Pi(\frac{1}{N_0}, \frac{1}{N_1})$;
2: Split the graph $\mathcal{G} \in \mathbb{R}^{N\times(d+N)}$ into $\mathcal{G}_0 \in \mathbb{R}^{N_0\times(d+N)}$ and $\mathcal{G}_1 \in \mathbb{R}^{N_1\times(d+N)}$;
3: Compute the cost matrix $\mathbf{C}$ with (11);
4: **repeat**
5: $\quad \mathbf{\Gamma}^* = \min_{\mathbf{\Gamma}\in\Pi(\frac{1}{N_0},\frac{1}{N_1})}\langle \mathbf{\Gamma}, \mathbf{C}\rangle$;
6: **until** Convergence;
7: Repair the $\mathcal{G}_0$ and $\mathcal{G}_1$ with (12).

Experiments

This section specifies the experimental procedure of our approach on the link prediction tasks and summarizes the analysis of the experimental results.

Experiment Setup

We first describe the experimental setup, including real-world datasets, baselines, evaluation metrics, and experiment details.

Datasets Our proposed algorithm is evaluated on two real-world network datasets. Statistical for datasets are summarized in Table 1.

Table 1: Statistic for datasets in experiments

Dataset	#Nodes	#Edges	#Node attributes	$\lvert S\rvert$
Cora	2708	5278	2879	7
CiteSeer	2110	3668	3703	6

- Cora[1] is a citation network consisting of 2708 scientific publications classified into seven classes. Each node in the network is a publication, characterized by a bag-of-words representation of the abstract. The link between nodes represents undirected citations, and the sensitive attributes are the category of the publication;
- CiteSeer[2] dataset consists of 2110 scientific publications classified into one of six classes. Similar to the Cora dataset, the node in the CiteSeer network is also a publication. Its sensitive attribute is the publication's category.

Baselines In our experiments, we choose the current two pre-processing dyadic fairness baselines listed below:

- FairDrop(Spinelli et al. 2021) is a biased dropout strategy that forces the graph topology to reduce the homophily of sensitive attributes. Specifically, they generate a fairer random copy of the original adjacency matrix to reduce the number of connections between nodes sharing the same sensitive attributes;
- FairEdge(Laclau et al. 2021) is a theoretically sound embedding-agnostic method for group and individually fair edge prediction. They repair the adjacency matrix of plain graphs based on the optimal transport theory and directly ignore the influence of node attributes.

Evaluation metrics Firstly, we expect insight into the structural changes between the repaired and the original graph for the pre-processing mechanism. To measure this change, we use Assortativity Coefficient (AC)(Laclau et al. 2021) to evaluate the correlation between the sensitive attributes of every pair of nodes that are connected. The AC values between -1 and 1, and the value close to 0 means no strong association of the sensitive attributes between connected nodes.

For fairness of concern in this paper, we will first use Representation Bias (RB)(Buyl and De Bie 2020) to measure whether the embedding is well-obfuscated, i.e., contains no sensitive information. Secondly, we extend RB further to the dyadic fairness considered in this paper: DyadicRB. The (Dyadic) RB is calculated based on the accuracy of (Dyadic) sensitive feature classification problem. Take the DyadicRB as an example; it can be calculated based on the following:

$$DyadicRB = \min_{g} \frac{1}{|\mathcal{E}|} \sum_{(u,v)\in\mathcal{E}} (S(u) \oplus S(v) | \boldsymbol{z}_u, \boldsymbol{z}_v).$$

Finally, without limiting to unbiased embeddings, we use the DDI defined in (2) to measure the fairness properties of the predictions themselves. We evaluate the effectiveness of our method on the link prediction tasks from both utility and fairness perspectives. For utility, we use Accuracy (ACC) to measure the predictor's performance trained on the data.

Experiment Details For all the experiments, we use Node2Vec(Grover and Leskovec 2016) as our embedding function and the support vector classifier as our link predictor. The node embedding's dimension is 128, and all the values are collected with 5 different random seeds. To enable results reproducing, our codes are open-sourced in Github[3], and more details can be found in it.

Experimental Results

We will evaluate the effectiveness of our proposed method on real-world datasets at different stages along the pipeline of the link prediction task.

Impact on the graph structure From Table 2, we can find that the AC values of the two original graphs are relatively high, which indicates that the links often appear between nodes that have the same sensitive attributes. This introduces the discrimination on the nodes with different sensitive attributes. The three repairing methods can reduce the assortativity coefficient from the original graph. Specifically, our method achieves smaller AC than the FairEdge, which indicates the effectiveness of our method. Moreover, the FairDrop achieves the smallest AC. It reduces the AC to negative numbers, indicating that the different sensitive attribute nodes are more likely to connect. Although the lower AC it is, the prediction accuracy may be highly influenced, and this phenomenon is shown later.

[1] https://networkrepository.com/cora.php
[2] https://networkrepository.com/citeseer.php
[3] https://github.com/mail-ecnu/OTDyadicFair

Impact on node embeddings Secondly, we consider the impact on the node embeddings with different repairing methods. We adopt the two aforementioned metrics i.e., RB and DyadicRB, to quantify the fairness of the node embedding. As shown in Tables 3 and 4, our method achieves the best score of RB and DyadicRB. This indicates that after repairing with our method, both the sensitive attribute predicting and dyadic sensitive attribute relation predicting is hard. The embedding thus is concluded as obtaining the fairness and the dyadic fairness. To better understand the impact of our repairing on node embedding, we use PCA to reduce the learned embedding to the 2-dimension space. As shown in Figure 2, the embedding learned in the original graph is distributed with highly correlated to the node's sensitive feature, which corresponds to the high RB. Embedding learned from the repaired graph of our method is less correlated with the sensitive features, corresponding to the low RB. Moreover, we consider dyadic embedding. As shown in Figure 3, the learned dyadic embedding of our method is less correlated than the original graph, which indicates the less predictable of the dyadic sensitive features' relationship (the low DyadicRB).

Impact on link prediction We then consider the performance of the link prediction task. Two metrics are adopted, i.e., ACC and DDI. The ACC indicates the utility of the predictor, while the DDI is the quantity of dyadic fairness the predictor achieves. For the Cora dataset, all the three repairing methods lose the ACC to gain dyadic fairness. Both the FairDrop and our method achieve a high quantity of dyadic fairness (DDI), but our method achieves a higher ACC. Compared with FairEdge, our method achieves a similar ACC performance while dominating the dyadic fairness performance. For the CiteSeer dataset, things are different. FairEdge is hard to gain better fairness after repairing. Compared with the FairDrop, our method achieves high DDI with a smaller loss of accuracy.

Table 2: Assortativity Coefficient

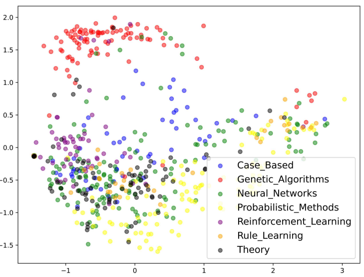
Dataset	Original	FairEdge	FairDrop	Ours
Cora	.771	.668	$-.089$	.397
CiteSeer	.673	.645	$-.065$	.567

Table 3: Experimental Results on Cora. The ↑ (↓) denotes the higher (lower) the better it is.

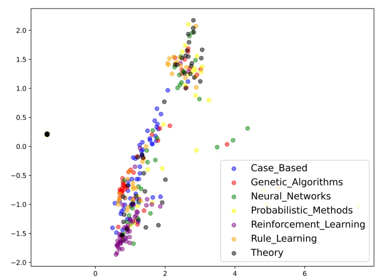
	ACC ↑	DDI↑	RB ↓	DyadicRB ↓
Original	**.829** ± .007	.266 ± .012	.834 ± .004	.726 ± .009
FairEdge	.663 ± .008	.393 ± .073	.655 ± .004	.596 ± .031
FairDrop	.533 ± .019	.657 ± .087	.467 ± .015	**.522** ± .018
Ours	.614 ± .006	**.836** ± .106	**.172** ± .018	**.522** ± .013

Table 4: Experimental Results on CiteSeer. ↑ (↓) denotes the higher (lower) the better respectively.

	ACC ↑	DDI↑	RB ↓	DyadicRB ↓
Original	.820 ± .011	.372 ± .019	.661 ± .005	.658 ± .009
FairEdge	**.821** ± .013	.389 ± .018	.655 ± .004	.623 ± .023
FairDrop	.532 ± .024	**.717** ± .081	.493 ± .021	.510 ± .037
Ours	.585 ± .014	.653 ± .181	**.211** ± .027	**.506** ± .036

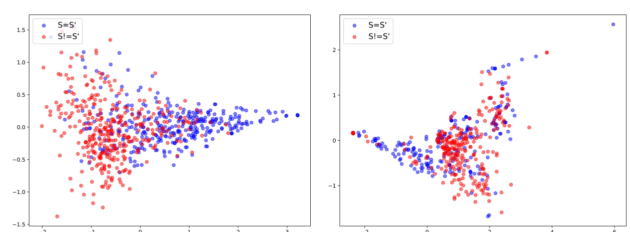


(a) Original Embedding



(b) Our Embedding

Figure 2: Visualization with PCA of node embedding learned by the Node2Vec on the Cora dataset. The different colors indicate different sensitive attributes. The left is the node embedding learned from the original graph, while the right is learned from the graph repaired by our method.



(a) Original Dyadic Embedding (b) Our Dyadic Embedding

Figure 3: Visualization with PCA of dyadic node embedding learned by the Node2Vec on Cora dataset. Different colors indicate whether two nodes' sensitive features are the same.

Conclusion

This paper proposes a pre-processing method to achieve dyadic fairness in link prediction tasks. By transforming the dyadic fairness obtaining problem to a conditional distribution alignment problem, dyadic fairness can be obtained with flexibility and unambiguity. Further, we propose a practical repairing implementation based on optimal transport theory. Experiments on Cora and CiteSeer show that our method has significant results in obtaining the dyadic fairness of link prediction.

Acknowledgement: This work was supported in part by NSFC (No. 12071145), STCSM (No. 20DZ1100300 and 20DZ1100304), the Open Research Projects of Zhejiang Lab (NO.2021KE0AB03) and a grant from Shenzhen Institute of Artificial Intelligence and Robotics for Society.

References

Al Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*, volume 30, 798–805.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica (2016). *Google Scholar*, 23.

Besse, P.; del Barrio, E.; Gordaliza, P.; and Loubes, J.-M. 2018. Confidence intervals for testing disparate impact in fair learning. *arXiv*, 1807.06362.

Buyl, M.; and De Bie, T. 2020. Debayes: a bayesian method for debiasing network embeddings. In *ICML*, 1220–1229.

Chiappa, S.; and Pacchiano, A. 2021. Fairness with Continuous Optimal Transport. *arXiv*, 2101.02084.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair Regression with Wasserstein Barycenters. In *NeurIPS*.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *FAccT*, 329–338.

Gordaliza, P.; Barrio, E. D.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining Fairness using Optimal Transport Theory. In *ICML*, 2357–2365.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *KDD*, 855–864.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2019. Wasserstein Fair Classification. In *UAI*, 862–872.

Khanam, K. Z.; Srivastava, G.; and Mago, V. 2020. The homophily principle in social network analysis. *arXiv*, 2008.10383.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Laclau, C.; Redko, I.; Choudhary, M.; and Largeron, C. 2021. All of the Fairness for Edge Prediction with Optimal Transport. In *AISTATS*, 1774–1782.

Lauret, J. 2019. Amazon's Sexist AI Recruiting Tool: How Did It Go so Wrong?'. *Becoming Human*, 16.

Li, P.; Wang, Y.; Zhao, H.; Hong, P.; and Liu, H. 2021. On Dyadic Fairness: Exploring and Mitigating Bias in Graph Connections. In *ICLR*.

Masrour, F.; Barjesteh, I.; Forsati, R.; Esfahanian, A.-H.; and Radha, H. 2015. Network completion with node similarity: A matrix completion approach with provable guarantees. In *ASONAM*, 302–307.

Masrour, F.; Wilson, T.; Yan, H.; Tan, P.; and Esfahanian, A. 2020. Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. In *AAAI*, 841–848.

Nielsen, A. 2020. *Practical Fairness*. O'Reilly Media.

Spinelli, I.; Scardapane, S.; Hussain, A.; and Uncini, A. 2021. Biased Edge Dropout for Enhancing Fairness in Graph Representation Learning. *arXiv*, 2104.14210.

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*, 923–932.

Villani, C. 2009. *Optimal transport: old and new*, volume 338. Springer.