

Obtaining Dyadic Fairness by Optimal Transport

Moyi YANG*, Junjie SHENG*, Wenyan LIU*, Bo JIN*, Xiaoling WANG*, Xiangfeng WANG*

*School of Computer Science and Technology, East China Normal University, Shanghai, China 200062

Abstract—Fairness has been taken as a critical metric in machine learning models, which is considered as an important component of trustworthy machine learning. In this paper, we focus on obtaining fairness for popular link prediction tasks, which are measured by *dyadic fairness*. A novel pre-processing methodology is proposed to establish dyadic fairness through data repairing based on optimal transport theory. With the well-established theoretical connection between the dyadic fairness for graph link prediction and a conditional distribution alignment problem, the dyadic repairing scheme can be equivalently transformed into a conditional distribution alignment problem. Furthermore, an optimal transport-based dyadic fairness algorithm called DyadicOT is obtained by efficiently solving the alignment problem, satisfying flexibility and unambiguity requirements. The proposed DyadicOT algorithm shows superior results in obtaining fairness compared to other fairness methods on two benchmark graph datasets.

Index Terms—Optimal Transport, Dyadic Fairness, Link Prediction

I. INTRODUCTION

Machine learning has been widely adopted in real-world applications. Although remarkable results were achieved in the prediction and decision-making scenarios, unexpected bias occurs regularly [19]–[21]. For example, 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 [24]. The Amazon company found that the AI hiring tool they developed to automate the hiring process is biased against women [26]. Many works emerge to design algorithms to avoid such biases and aim to obtain *fair* machine learning models.

This work focuses on achieving fairness in link prediction tasks. The link prediction task is a fundamental 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 popular algorithms, e.g., Node2Vec [16] and GCN [15], have been proposed to solve the link prediction task with superior performance in many scenarios. However, the dataset collected for the model training procedure usually has various unexpected biases. This will lead to unfair results for the link prediction model obtained. For instance, after collecting data from social media platforms, early works highlighted that users were more interested in conversing with others of the same race and

gender [28]. 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. This will unfairly disadvantage some users. To formally define such an unfair phenomenon, [3], [4] introduced 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 works have been proposed to achieve dyadic fairness in link prediction tasks, which can be roughly divided into three categories: 1) in-processing scheme [4] considers modifying the learning algorithm to eliminate bias; 2) post-processing scheme [3] attempts to debias directly the model’s output after training; 3) pre-processing scheme [2] aims to repair the graph data before the training procedure, and ensures the link prediction results can satisfy dyadic fairness. In this paper, our proposed method is established under the pre-processing scheme. Compared to the in-processing and post-processing schemes, the pre-processing scheme should be the most flexible fairness intervention [27]. Suppose the discriminating information is removed from the data during the pre-processing stage, the processed data could be utilized to solve arbitrary downstream tasks without concern about the fairness issue. Few works have studied obtaining dyadic fairness through a pre-processing scheme. FairDrop [2] proposed a heuristic repairing method that can mask out edges based on the dyadic sensitive attributes. It is easy to implement but without a theoretical guarantee of achieving fairness. To design a theoretically sound pre-processing scheme, FairEdge [5] firstly adopts the Optimal Transport (OT) theory [13] 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 Wasserstein barycenter. Dyadic fairness is obtained once the adjacency information distributions are all repaired as the obtained Wasserstein barycenter. Unlike the previous approach, we expect to focus on attributed graphs (each node has attributes) that are more general in the real world. Because node attributes introduce bias even if the bias of adjacency information can be removed, those algorithms that simply consider plain graphs cannot solve this problem, and the achievement of dyadic fairness on attributed graphs is still underexploited.

This work was supported in part by STCSM (No. 20DZ1100300) and NSFC (No. 12071145). This paper is funded by Shanghai Trusted Industry Internet Software Collaborative Innovation Center. Corresponding author: Xiangfeng Wang.

II. RELATED WORKS

A. Fairness in Link Prediction

Link prediction is a well-researched problem in applications related to graph data [22], [23]. Since fairness in graph-structured data is a relatively new research topic, only a few works have investigated fairness issues in link prediction. In [2], 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, [3] 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, [4] 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, [5] formulated the problem of fair edge prediction and proposed an embedding-agnostic repairing procedure for the adjacency matrix with a trade-off between group and individual fairness. However, they still ignore the node attributes, which impact both the prediction and fairness performance.

B. 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 [5]–[7] and post-processing [8], [10] methods, respectively. In addition, the in-processing [8], [11] methods based on optimal transport achieve fairness by imposing constraints in terms of the Wasserstein distance in the objective function.

III. DYADIC FAIRNESS IN LINK PREDICTION

In this section, we formulate dyadic fairness in the link prediction task and define two metrics (dyadic disparate impact and dyadic balanced error rate) to quantify 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.

A. Problem Formulation

Given the graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} := \{v_1, \dots, v_N\}$ be the node set of the graph and $\mathcal{E} := \{e_1, \dots, e_N\}$ 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 i th row of a non-negative adjacency matrix $A \in \{0, 1\}^{N \times N}$ which summarizes the connectivity in the graph. If nodes v_i and v_j are connected, then $A_{ij} = 1$; otherwise, $A_{ij} = 0$. The link prediction model usually identifies whether the link between two nodes (i, j) exists based on their node representations, i.e., $g : \mathbf{z}_i \times \mathbf{z}_j \mapsto \{0, 1\}$ where the \mathbf{z}_i denotes the node i 's representation. The \mathbf{z}_i is usually obtained by random walk or graph convolution on the whole graph: $\mathbf{z}_i = f(\mathcal{G})[i]$ where the $f : \mathcal{G} \mapsto \mathbb{R}^{N \times d}$ is called the embedding function. The d is the dimension of the node representation, 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 exists between them. To study the fairness of link prediction tasks, we assume that all nodes have one sensitive feature $S : \mathcal{V} \rightarrow \mathcal{S}$. 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 the following two assumptions:

1). Equivalence assumption

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

which is based on the fact that each node has the same chance of being sampled regardless of its sensitive attribute value. For instance, $\mathbb{P}(S = \text{man}) = \mathbb{P}(S = \text{woman})$ is always an equivalence relationship independent of the sampling process and the obtained graph data itself;

2). Propensity assumption

$$\begin{aligned} \mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 1 \mid S(u) \oplus S(v) = 0) \\ \geq \mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 1 \mid S(u) \oplus S(v) = 1), \end{aligned}$$

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 to 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 [4].

Definition 1 (Dyadic Fairness): A link predictor g obtains dyadic fairness if for node representation \mathbf{z}_i and \mathbf{z}_j

$$\mathbb{P}(g(\mathbf{z}_i, \mathbf{z}_j) \mid S(i) \oplus S(j) = 1) = \mathbb{P}(g(\mathbf{z}_i, \mathbf{z}_j) \mid S(i) \oplus S(j) = 0). \quad (1)$$

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

Definition 2 (DDI: Dyadic Disparate Impact): Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a function $g(\mathbf{z}_u, \mathbf{z}_v) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \{0, 1\}$, 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. \mathbf{Z} if:

$$\text{DDI}(g, \mathbf{Z}, S) = \frac{\mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 1 \mid S(u) \oplus S(v) = 1)}{\mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 1 \mid S(u) \oplus S(v) = 0)} \leq \tau. \quad (2)$$

DDI measures the fairness level of the predictor. The higher 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 (DBER: Dyadic Balanced Error Rate): For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a function $g(\mathbf{z}_u, \mathbf{z}_v) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \{0, 1\}$, we define the dyadic balanced error rate of the predictor g as the average class-conditional error:

$$\text{DBER}(g, \mathbf{Z}, S) = \frac{1}{2} [\mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 0 \mid S(u) \oplus S(v) = 1) + \mathbb{P}(g(\mathbf{z}_u, \mathbf{z}_v) = 1 \mid S(u) \oplus S(v) = 0)]. \quad (3)$$

DBER measures the general misclassification error of sensitive attributes by g in the particular case of $\mathbb{P}(S \oplus S' = 1) = \mathbb{P}(S \oplus S' = 0) = \frac{1}{2}$. DBER can be guaranteed to be smaller than $\frac{1}{2}$. With a larger DBER, the data and predictor g will be more fair. If DBER equals $\frac{1}{2}$, then DDI will be 1, and dyadic fairness will be achieved.

B. Obtaining Dyadic Fairness

In this paper, we consider establishing dyadic fairness through pre-processing the graph data. Due to the nature of pre-processing, our repairing procedure has no relationship with the embedding function f and predictor g . As a result, it becomes important to ensure that the repaired data can achieve dyadic fairness for arbitrary embedding function and predictor. These can be considered as the requirements **flexibility**. Furthermore, another straightforward requirement needs to be emphasised, i.e., **unambiguity**. After repairing, the attribute and adjacency information of each node should be determined without ambiguity.

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

$$\mathbf{Z}^* = \arg \max_{\mathbf{Z}} \min_g \text{DBER}(g, \mathbf{Z}, S). \quad (4)$$

Suppose that the repaired data \mathbf{Z}^* ensures high DBER under the most unfair predictor. In that case, it obtains dyadic fairness with wide applicability to 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 [6].

Theorem 1: The smallest DBER for the data \mathbf{Z} is equal to:

$$\min_g \text{DBER}(g, \mathbf{Z}, S) = \frac{1}{2} \left(1 - \frac{1}{2} W_{1,\neq}(\hat{\gamma}_0, \hat{\gamma}_1) \right), \quad (5)$$

where $W_{1,\neq}$ denotes the Wasserstein distance between the conditional joint distributions of the node representation with the

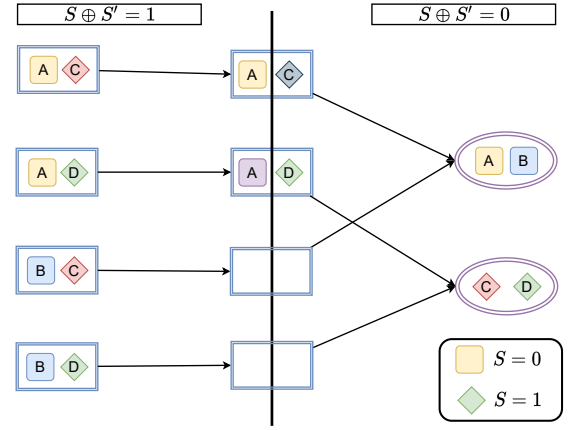


Fig. 1: The ambiguity illustration of dyadic repairing. These pairs (A, C) and (A, D) are repaired respectively to the pairs in the black line. A 's original attribute is yellow, while in the repaired data, it has multiple values ("yello" and "purple"), which might lead to ambiguity.

Hamming cost function. $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are conditional distributions over $\mathbf{Z} \times \mathbf{Z}$ given $S(u) \oplus S(v) = 0$ and $S(u) \oplus S(v) = 1$. The detailed proof of this theorem has been elaborated in work [6]. As shown in the theorem, the dyadic balanced error rate of the most unfair predictor depends on the Wasserstein distance between the two conditional dyadic node representation distributions $(\hat{\gamma}_0, \hat{\gamma}_1)$. When $W_{1,\neq}(\hat{\gamma}_0, \hat{\gamma}_1) = 0$, which means that the two conditional distributions are identical, i.e.,

$$\mathbb{P}(\mathbf{z}_u, \mathbf{z}_v \mid S(u) \oplus S(v) = 1) = \mathbb{P}(\mathbf{z}_u, \mathbf{z}_v \mid S(u) \oplus S(v) = 0). \quad (6)$$

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

One straightforward repairing scheme is directly moving the two conditional distributions to the same distribution. However, the representation of node i ' \mathbf{z}_i often occurs more times in $\hat{\gamma}_0$ and $\hat{\gamma}_1$. When repairing $\hat{\gamma}_0$ and $\hat{\gamma}_1$, \mathbf{z}_i will probably assign multiple values. For example, as shown in Figure 1, the direct repairing leads to ambiguity in 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 is satisfied:

$$\mathbb{P}(\mathbf{z}_u \mid S(u) = 0) = \mathbb{P}(\mathbf{z}_v \mid S(v) = 1). \quad (7)$$

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

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

which indicates the satisfactory of (6) accordingly. For the necessary part, it can be easily proved by contradiction. The above proposition implies that a fair representation of nodes is sufficient to achieve dyadic fairness in the optimal case. Repairing based on (7) allows us to obtain the 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 \mathbf{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.,

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

then for any embedding function f , $f(\mathcal{G})[u] = f(\mathcal{G})[v]$. $\mathbf{x}_u, \mathbf{x}_v$ denote the attribute of node u and node v , respectively. $\mathbf{e}_u, \mathbf{e}_v$ 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) into the following one:

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

Based on (9), the dyadic fairness (1) can be further satisfied for arbitrary predictors. In the following, we aim to propose an efficient algorithm to guarantee (9).

IV. ALGORITHMIC FRAMEWORK

In this section, we introduce a practical and efficient algorithm called **DyadicOT** to achieve dyadic fairness in link prediction tasks based on optimal transport theory. It can be easily extended to multi-valued sensitive attributes problems, which can relax the binary sensitive value constraint.

A. Dyadic fairness with optimal transport

In order to achieve dyadic fairness through (9), we first represent the graph \mathcal{G} as a matrix $\mathbb{R}^{N \times (d+N)}$ where each row represents the attribute of one node (\mathbf{x}_u) and adjacency information (\mathbf{e}_u). According to the sensitive feature of each node, we further split \mathcal{G} into $\mathcal{G}_0 \in \mathbb{R}^{N_0 \times (d+N)}$ and $\mathcal{G}_1 \in \mathbb{R}^{N_1 \times (d+N)}$ where 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 $\hat{\gamma}_0$ and $\hat{\gamma}_1$. Our goal can be explicitly described as $\min_{\mathcal{G}} W_{1,\neq}(\hat{\gamma}_0, \hat{\gamma}_1)$. To achieve that goal, we solve the following optimal transport problem:

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

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., $s \in \{0, 1\}$.

1) *Define the cost matrix \mathbf{C} :* Considering the distribution $\hat{\gamma}_0$ and $\hat{\gamma}_1$ encodes two important parts of information about the node, i.e., feature \mathbf{x}_u and the 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 the structure term.

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

To emphasis, although the Hamming distance is used in the above theoretical results, we practically employ the squared Euclidean distance.

2) *The DyadicOT algorithm:* The optimal transport plan $\mathbf{\Gamma}^*$ can be obtained, and further $\mathbf{\Gamma}^*$ can be utilized to repair the node feature and the 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 [13], i.e.,

$$\begin{cases} \tilde{\mathcal{G}}_0 = \pi_0 \mathcal{G}_0 + \pi_1 \mathbf{\Gamma}^* \mathcal{G}_1, \\ \tilde{\mathcal{G}}_1 = \pi_1 \mathcal{G}_1 + \pi_0 \mathbf{\Gamma}^{*\top} \mathcal{G}_0. \end{cases} \quad (12)$$

Following the above schemes (10)-(12), the proposed DyadicOT algorithm can be concluded as follows.

Algorithm 1 DyadicOT: Dyadic fairness with OT

- 1: Initialize η and $\mathbf{\Gamma}^0 \in \Pi(1/N_0, 1/N_1)$;
 - 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: Transform the distributions to their Wasserstein barycenter by solving (10);
 - 5: Repair the \mathcal{G}_0 and \mathcal{G}_1 with (12).
-

3) *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 [29] of the conditional distributions. Specifically, since each node has $|S|$ possible values of sensitive attribute, we first divide the graph \mathcal{G} into $|S|$ sensitive attribute groups $\mathcal{G}_k \in \mathbb{R}^{N_k \times (d+N)}$ where N_k is the number of nodes with $S = k$. Then, we compute the Wasserstein barycenter $\bar{\mathcal{G}}^*$ of these groups as follows:

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

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

$$\tilde{\mathcal{G}}_k = N_k \mathbf{\Gamma}_k^{*\top} \bar{\mathcal{G}}^*. \quad (14)$$

V. EXPERIMENT RESULTS

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

A. Experiment Setup

At the beginning, 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. The statistical results for these two datasets are summarized in the following Table I.

TABLE I: Statistic for datasets in experiments

Dataset	#Nodes	#Edges	#Node attributes	S
CORA	2708	5278	2879	7
CiteSeer	2110	3668	3703	6

- CORA¹ 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 sensitive attributes are set to be the categories of the publication;

- CiteSeer² 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 set to be the publication’s categories.

Baselines. The following two pre-processing dyadic fairness baseline methods are chosen to be compared as follows:

- FairDrop [2] is a biased dropout strategy that forces the graph topology to reduce the homophily of sensitive attributes. Specifically, it generates a fairer random copy of the original adjacency matrix to reduce the number of connections between nodes sharing the same sensitive attributes;
- FairEdge [5] is a theoretically sound embedding-agnostic method for group and individually fair edge prediction. It aims to repair the adjacency matrix of plain graphs based on the optimal transport theory and directly ignore the influence of node attributes.

Evaluation metrics. In order to measure the structural changes between the repaired and the original graph for the pre-processing mechanism, we use Assortativity Coefficient (AC) [5] to evaluate the correlation between the sensitive attributes of every pair of nodes that are connected. The values of AC always belongs to $[-1, 1]$, and the value close to 0 denotes that there is no strong association of the sensitive attributes between the connected nodes.

To evaluate the fairness, which is the main concern of our work, Representation Bias (RB) [18] is employed to measure whether the embedding is well-obfuscated, i.e., contains no sensitive information. Further more, we introduce a new dyadic fairness evaluation metric called DyadicRB through extending classical RB metric. Similar with RB, the DyadicRB

TABLE II: Assortativity Coefficient

Dataset	Original	FairEdge	FairDrop	DyadicOT
CORA	.771	.668	-.089	.397
CiteSeer	.673	.645	-.065	.567

is calculated based on the accuracy of dyadic sensitive feature classification problem, which can be calculated as

$$\text{DyadicRB} = \sum_{s=0}^1 \frac{|\mathcal{E}_s|}{|\mathcal{E}|} \text{Accuracy}(S(u) \oplus S(v) | \mathbf{Z}_{u,v}).$$

where $\mathbf{Z}_{u,v}$ is the edge embedding as the concatenation of the embeddings of the two nodes u and v connected by the link. And $\text{Accuracy}(\cdot)$ is the accuracy of predicting the dissimilarity of sensitive information $S(u) \oplus S(v)$ based on edge embedding $\mathbf{Z}_{u,v}$. Without limiting ourselves to unbiased embeddings, we utilize DDI (2) to measure the fairness properties of the predictions themselves. The effectiveness of our method on link prediction tasks from both the utility and fairness perspectives will be further evaluated. As for the utility index, the Accuracy (ACC) is considered to measure the predictor’s performance.

Experiment Details. Node2Vec [16] and support vector classifier are employed for all experiments as our embedding function and link predictor, respectively. The dimension of the node’s embedding is 128, and all values are collected with 5 different random seeds. For easy reproduction of the results, our codes are open-sourced in Github³, and more details can be found there.

B. Experiment Results

In this section, we will evaluate and compare the effectiveness of our proposed DyadicOT method with other SOTA algorithms on real-world datasets at different stages along the pipeline of the link prediction task.

Impact on the graph structure. Table II shows that the AC values of the two original graphs are relatively high, indicating that the links often appear between nodes with the same sensitive attributes. This leads to discrimination against nodes with different sensitive attributes. The three repairing methods can reduce the assortativity coefficient from the original graph. Specifically, DyadicOT achieves smaller AC than FairEdge, which indicates the effectiveness of DyadicOT. FairDrop could achieve a much smaller AC, and the resulting negative AC indicates that the different sensitive attribute nodes are more likely to connect. However, the prediction accuracy of FairDrop may be highly influenced, and this phenomenon has been shown in Table III and Table IV.

Impact on node embeddings. Comparison on the impact on node embeddings among different repairing methods is another important concern. Two aforementioned metrics are used, i.e., RB and DyadicRB, to quantify the fairness of the node

¹<https://networkrepository.com/cora.php>

²<https://networkrepository.com/citeseer.php>

³<https://github.com/mail-ecnu/OTDyadicFair>

TABLE III: Results on CORA. \uparrow (\downarrow) denotes the higher (lower) the better respectively.

	ACC \uparrow	DDI \uparrow	RB \downarrow	DyadicRB \downarrow
Original	.829 \pm .007	.266 \pm .012	.834 \pm .004	.726 \pm .009
FairEdge	.663 \pm .008	.393 \pm .073	.655 \pm .004	.596 \pm .031
FairDrop	.533 \pm .019	.657 \pm .087	.467 \pm .015	.522 \pm .018
DyadicOT	.614 \pm .006	.836 \pm .106	.172 \pm .018	.522 \pm .013

TABLE IV: Results on CiteSeer.

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

embedding. As shown in Tables III and Table IV, DyadicOT achieves the best score of both RB and DyadicRB. These results indicate that both the sensitive attribute prediction and the dyadic sensitive attribute relation prediction are hard after repairing through DyadicOT.

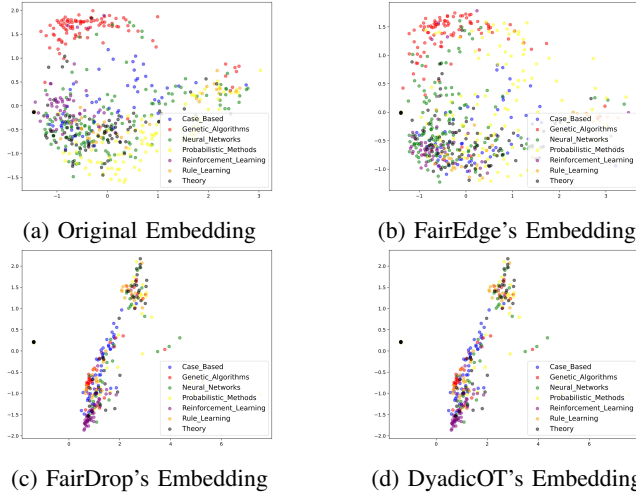


Fig. 2: Visualization of node embedding learned by Node2Vec on CORA. Different colors indicate different sensitive attributes. (a) and (b) denote the node embeddings learned from the original graph or the graph repaired by DyadicOT respectively.

To better understand the impact of our repairing on node embedding, we employ the PCA method to reduce the learned embedding into 2-dimension space. As shown in Figure 2, the learned embedding from the original graph is distributed with highly correlated to the node's sensitive feature, which corresponds to higher RB. The embedding learned from the repaired graph by DyadicOT is less correlated with the sensitive features compared with the baselines, corresponding to lower RB. The comparison of dyadic embedding is shown in Figure 3. The learned dyadic embedding by DyadicOT is less correlated than the original graph, indicating less

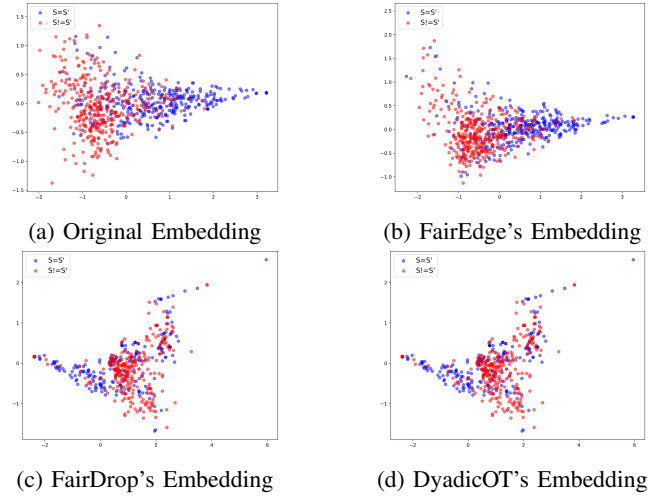


Fig. 3: Visualization of dyadic node embedding learned by Node2Vec on CORA. Here, the red colour represents node embeddings with different sensitive attributes, while the blue colour indicates node embeddings with the same sensitive attributes.

predictability of the dyadic sensitive features' relationship (lower DyadicRB).

Impact on link prediction. Finally, we consider the performance comparison on the link prediction task through two basic metrics, i.e., ACC and DDI. ACC indicates the utility of the predictor, while DDI denotes the quantity of dyadic fairness the predictor achieves. For the CORA dataset, all three repairing methods lose ACC while obtaining dyadic fairness. Compared with FairEdge and FairDrop, DyadicOT achieves the best quantity of dyadic fairness (DDI), and the ACC decreases within the tolerance range.

As for the other CiteSeer dataset, FairEdge nearly cuts no ice on fairness. However, compared to FairDrop, DyadicOT achieves higher DDI with less accuracy decrease, which indicates the better performance of DyadicOT.

VI. CONCLUSION

This paper proposes a pre-processing method to achieve dyadic fairness in link prediction tasks. By transforming the dyadic fairness obtaining problem into a conditional distribution alignment problem, dyadic fairness can be obtained with flexibility and unambiguity. Furthermore, a practical repairing method is introduced based on optimal transport theory. Experiments on CORA and CiteSeer show that the proposed DyadicOT method has significant results in obtaining the dyadic fairness of link prediction.

REFERENCES

- [1] Dai, Enyan, and Suhang Wang. "Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information." Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021.
- [2] Spinelli I, Scardapane S, Hussain A, et al. FairDrop: Biased edge dropout for enhancing fairness in graph representation learning[J]. IEEE Transactions on Artificial Intelligence, 2021, 3(3): 344-354.

- [3] Masrour F, Wilson T, Yan H, et al. Bursting the filter bubble: Fairness-aware network link prediction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 841-848.
- [4] Li P, Wang Y, Zhao H, et al. On dyadic fairness: Exploring and mitigating bias in graph connections[C]//International Conference on Learning Representations. 2021.
- [5] Laclau C, Redko I, Choudhary M, et al. All of the fairness for edge prediction with optimal transport[C]//International Conference on Artificial Intelligence and Statistics. 2021: 1774-1782.
- [6] Gordaliza P, Del Barrio E, Fabrice G, et al. Obtaining fairness using optimal transport theory[C]//International Conference on Machine Learning. 2019: 2357-2365.
- [7] Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 259-268.
- [8] Jiang R, Pacchiano A, Stepleton T, et al. Wasserstein fair classification[C]//Uncertainty in Artificial Intelligence. 2020: 862-872.
- [9] Dai E, Wang S. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 680-688.
- [10] Chzhen E, Denis C, Hebiri M, et al. Fair regression with wasserstein barycenters[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 7321-7331.
- [11] Silvia Chiappa and Aldo Pacchiano. Fairness with continuous optimal transport, 2021.
- [12] Oneto L, Chiappa S. *Recent Trends in Learning From Data*[M]. Springer International Publishing, 2020.
- [13] Villani C. *Optimal transport: old and new*[M]. Berlin: Springer, 2009.
- [14] Zhang S, Yao L, Sun A, et al. Deep learning based recommender system: A survey and new perspectives[J]. *ACM Computing Surveys*, 2019, 52(1): 1-38.
- [15] Kipf, Thomas and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks." *ArXiv abs/1609.02907* (2017): n. pag.
- [16] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.
- [17] Rahman, Tahleen A. et al. "Fairwalk: Towards Fair Graph Embedding." *IJCAI* (2019).
- [18] Buyl M, De Bie T. Debayes: a bayesian method for debiasing network embeddings[C]//International Conference on Machine Learning. 2020: 1220-1229.
- [19] Stoica A A, Riederer C, Chaintreau A. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 923-932.
- [20] Besse, Philippe C. et al. "Confidence Intervals for Testing Disparate Impact in Fair Learning." *ArXiv abs/1807.06362* (2018): n. pag.
- [21] Friedler S A, Scheidegger C, Venkatasubramanian S, et al. A comparative study of fairness-enhancing interventions in machine learning[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019: 329-338.
- [22] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM06: Workshop on Link Analysis, Counter-terrorism and Security. 2006, 30: 798-805.
- [23] Masrour F, Barjesteh I, Forsati R, et al. Network completion with node similarity: A matrix completion approach with provable guarantees[C]//2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015: 302-307.
- [24] Angwin J, Larson J, Mattu S, et al. *Machine bias*[M]//Ethics of Data and Analytics. Auerbach Publications, 2016: 254-264.
- [25] Nguyen T T, Hui P M, Harper F M, et al. Exploring the filter bubble: the effect of using recommender systems on content diversity[C]//Proceedings of the 2014 World Wide Web Conference. 2014: 677-686.
- [26] Lauret J. Amazon's Sexist AI Recruiting Tool: How Did It Go so Wrong? [J]. Medium.[Online] Available: <https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e>, 2019.
- [27] Nielsen A. *Practical Fairness*[M]. O'Reilly Media, 2020.
- [28] Khanam K Z, Srivastava G, Mago V. The homophily principle in social network analysis[J]. *arXiv preprint arXiv:2008.10383*, 2020.
- [29] Cuturi M, Doucet A. Fast computation of Wasserstein barycenters[C]//International Conference on Machine Learning. 2014: 685-693.