Breaking the Dyadic Barrier: Rethinking Fairness in Link Prediction Beyond Demographic Parity

João Mattos[1], **Debolina Halder Lina**[1], **Arlei Silva**[1,2]
[1]Computer Science Department, Rice University
[2]Ken Kennedy Institute, Rice University
Houston, TX, USA
{jrm28,dl73,arlei}@rice.edu

Abstract

Link prediction is a fundamental task in graph machine learning with applications ranging from social recommendation to knowledge graph completion. Fairness in this setting is critical, as biased predictions can exacerbate societal inequalities. Prior work adopts a dyadic definition of fairness, enforcing fairness through demographic parity between intra-group and inter-group link predictions. However, we show that this dyadic framing can obscure underlying disparities across subgroups, allowing systemic biases to go undetected. Moreover, we argue that demographic parity does not meet the desired properties for fairness assessment in ranking-based tasks such as link prediction. We formalize the limitations of existing fairness evaluations and propose a framework that enables a more expressive assessment. Additionally, we propose a lightweight post-processing method combined with decoupled link predictors that effectively mitigates bias and achieves state-of-the-art fairness–utility trade-offs.

Code — https://github.com/joaopedromattos/MORAL
Extended version — https://arxiv.org/abs/2511.06568

1 Introduction

Link prediction is the task of discovering potential missing or future links in a network. There is extensive literature on link prediction models (Liben-Nowell and Kleinberg 2007; Zhang and Chen 2018; Pan, Shi, and Dokmanić 2022; Zhu et al. 2021; Chamberlain et al. 2023), which demonstrates the relevance of this task and the variety of domains where it can be applied, ranging from social networks to knowledge graphs. Among these domains, many types of networks are prone to biases in their structure and features (Dai et al. 2024; Stoica, Riederer, and Chaintreau 2018; Karimi et al. 2018), requiring the adoption of fair machine learning methods to mitigate bias in model predictions (Li et al. 2021a, 2022a; Current et al. 2022a; Tsioutsiouliklis et al. 2022, 2021a). Our work is focused on the fair link prediction problem and, more specifically, on the evaluation and design of link prediction models under fairness considerations.

As a motivational example, consider the recommendation problem in social networks, where links represent social interactions. On a professional network, such as LinkedIn, biased connection recommendations can lead to persistent employment disparities between groups. Closed male networking circles provide more job leads and higher status connections than female/minority ones (Calvo-Armengol and Jackson 2004), resulting in white women receiving 33% fewer job leads on average, according to some models (McDonald, Lin, and Ao 2009). In addition, bias in friendship can lead to long-term accumulation (Gupta et al. 2021; Hofstra et al. 2017), with the large majority of links occurring only within the same communities, creating filter bubbles (Cinelli et al. 2021; Bakshy, Messing, and Adamic 2015).

The graph representation learning literature has focused on fair node representations (Zhu et al. 2024a,b; Ling et al. 2023; Agarwal, Lakkaraju, and Zitnik 2021; Laclau et al. 2021), and building upon these works, the current notion of fairness adopted by fair link prediction methods is *dyadic* (Li et al. 2021a; Current et al. 2022a; Li et al. 2022a; Luo et al. 2023). The main assumption behind dyadic fairness is to categorize groups according to a protected attribute of interest and obtain equalized positive outcomes across these groups. In the social network example, two communities can be identified (e.g., men and women), and we can define fairness as links between men and women (inter) having the same probability of occurring as links within these groups (intra). To mitigate the disparity between these probabilities, fair link prediction algorithms adopt different strategies, evaluated through group fairness statistical metrics (Dwork et al. 2012; Kusner et al. 2017; Masrour et al. 2020).

However, we show that fair node embeddings do not translate to fair link prediction. The limited expressive power of Graph Neural Networks (GNNs), outputs indistinguishable node representations in symmetric neighborhoods containing different sensitive groups. This prevents the adoption of fair node embeddings in training objectives that distinguish between edge groups to achieve parity.

In addition, the dyadic fairness applied by previous fair link prediction methods (e.g., intra vs. inter) is unable to capture biases that occur *within* sensitive groups of node pairs, which is known as fairness gerrymandering (Kearns et al. 2018). In the social network example, male-male connections might be systematically more likely than female-female connections, and the dyadic fairness evaluation metrics considered (in particular, demographic parity) are insensitive to this phenomenon, thereby masking underlying bi-

ases. The consequence of this limitation is a "glass ceiling" effect (Stoica, Riederer, and Chaintreau 2018), in which an under-representation of a subgroup of pairs goes undetected.

A third limitation of existing work on fair graph machine learning (including link prediction) is evaluating fairness as a subset selection problem, instead of based on ranking (Han et al. 2023; Kleinberg, Ryu, and Tardos 2024; Stoica, Litvak, and Chaintreau 2024; Li et al. 2021b). We claim that by not considering the ranking of candidate links, a link prediction method can be prone to *exposure bias*. In this fashion, one key property of a fair link prediction algorithm should be to ensure equality of probabilities and exposure, preventing one group from dominating the top positions of the ranking.

Our work is the first attempt at formalizing and demonstrating empirically the above limitations in the context of fair link prediction. To address these limitations, we propose using an exposure-based fairness evaluation metric previously applied in information retrieval (Draws, Tintarev, and Gadiraju 2021). Our experiments show that several existing approaches (Rahman et al. 2019; Dong et al. 2022; Ling et al. 2023; dai 2021; Wang et al. 2022b) fail at achieving a good tradeoff between accuracy and fairness under this evaluation metric. This motivates the design of a new post-processing algorithm that can be combined with any existing link prediction model to generate fair outcomes based on the new metric, outperforming existing alternatives.

We summarize the contributions of this work as follows: (1) we expose the limitations of demographic parity as a fairness metric in the context of link prediction; (2) we propose using an exposure-based fairness metric that overcomes the limitations of demographic parity and show that existing approaches for fair link prediction are ineffective under the new metric; and (3) we introduce **MORAL**, a post-processing algorithm that can de-bias the outputs of any link prediction model and achieves good accuracy vs. fairness tradeoffs under the new evaluation metric.

2 Preliminaries

Let a graph $G = (V, E, S)$, $v \in V$ is the set of nodes, $(u, v) \in E$ the set of existing edges in the graph, and $S \in \{0, 1\}^n$ is the vector of sensitive attributes, where s_v indicates the sensitive attribute of node v. In this work, we consider binary sensitive attributes for simplicity and the availability of datasets, but our analysis is also valid for categorical and/or multiple sensitive attributes. The binary sensitive attribute S produces three subgroups of node pairs, which for the remainder of the paper we denote as $E_{s\text{-}s'} = \{(u, v) \in V \times V \mid s_u = 1, s_v = 0\}$, $E_{s\text{-}s} = \{(u, v) \in V \times V \mid s_u = 1, s_v = 1\}$, and $E_{s'\text{-}s'} = \{(u, v) \in V \times V \mid s_u = 0, s_v = 0\}$, for simplicity. We consider a link prediction classifier a score function $f(\cdot, \cdot)$ that maps a given input pair (u, v) to a score $\hat{Y} \in [0, 1]$. We denote the scores of all candidate pairs as $R \in [0, 1]^{|C|}$, where C is the set of candidate pairs.

Definition 1 (Demographic Parity - Δ_{DP})**.** Let $\hat{Y}$ be the prediction of a binary classifier. Demographic parity is defined as: $|P(\hat{Y} \mid (u, v) \in E_{\text{intra}}) - P(\hat{Y} \mid (u, v) \in E_{\text{inter}})|$.

In this scenario, we can define the objective of fair link prediction with an output R as $\min_R \lambda \mathscr{L}_A(R)$, where $\mathscr{L}_A$ is an accuracy loss (usually Binary Cross Entropy), subject to a constraint (or regularization term) based on a bias loss $\mathscr{L}_B(R) \leq \beta$. Previous works (Li et al. 2021a, 2022a; Current et al. 2022a) follow the fair graph representation learning community practice of defining the task of fair link prediction in a *dyadic* framework. In particular, these works propose to divide pairs into intra-pairs ($E_{\text{intra}} = \{(u, v) \in E \mid E_{s\text{-}s} \bigcup E_{s'\text{-}s'}\}$) and inter-pairs ($E_{\text{inter}} = \{(u, v) \in E \mid E_{s'\text{-}s}\}$), and are trained and evaluated considering demographic parity (Δ_{DP}) or one of its surrogates as the main fairness metric ($\mathscr{L}_B \approx \Delta_{DP}$).

3 Limitations of Demographic Parity for Fair Link Prediction

Despite recent advances in fair graph learning, existing bias mitigation techniques for link prediction remain limited in three fundamental ways. First, many approaches rely on node representations from GNNs with constrained expressive power, restricting their ability to model the nuanced structural and demographic patterns necessary for fairness. Second, common fairness metrics such as demographic parity (Δ_{DP}) operate under a dyadic assumption that aggregates across sensitive subgroups, masking imbalances that occur within groups. Finally, these metrics are often permutation-invariant and insensitive to ranking order, making them unsuitable for applications where exposure and position matter. We systematically analyze these limitations to motivate a new fairness criterion for link prediction.

3.1 Expressive Power Impact on Fairness

Fair link prediction methods typically fall into two categories: node-level and link-level approaches. Node-level methods assume that fair node representations—often learned via Graph Neural Networks (GNNs)—will naturally result in fair link predictions (Li et al. 2021a). In contrast, link-level methods attempt to directly enforce fairness on edge predictions, either by learning fairness-aware edge embeddings or by modifying the graph structure (e.g., through unbiased adjacency matrices). These methods often rely on adversarial training or fairness-specific loss functions to enforce demographic parity across intra- and inter-group links (Li et al. 2021a; Current et al. 2022a). However, both approaches are fundamentally constrained by the limited expressive power of standard GNN architectures.

Most GNNs operate within the expressivity limits of the Weisfeiler-Lehman (1-WL) test, which restricts their ability to distinguish certain graph structures and node interactions (Xu et al. 2019). This poses a problem for fairness in link prediction, which often requires capturing subtle topological and demographic asymmetries across node pairs. For instance, a model must distinguish between link types (e.g., $E_{s'\text{-}s'}$ vs. $E_{s'\text{-}s}$) and detect systematic under-representation. Yet, 1-WL GNNs tend to produce similar embeddings for nodes in symmetric neighborhoods, failing to differentiate pairwise combinations that matter for fairness. See the extended version for an example of this limitation.

3.2 Bias Within Edge Groups ($E_{s\text{-}s}$ vs. $E_{s'\text{-}s'}$)

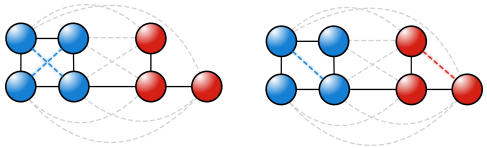

Figure 1: Toy example showing how Δ_{DP} fails to distinguish subgroup bias within aggregated edge groups. Both panels depict top-10 link predictions (dashed edges) over the same graph, achieving the optimal Δ_{DP} value despite the left scenario overrepresenting $E_{s\text{-}s}$ (blue) relative to $E_{s'\text{-}s'}$ (red), while $E_{s'\text{-}s}$ (gray) is the protected group.

As discussed, the limited expressivity of node-level representations motivates the use of fair edge representations for link prediction. However, most existing methods aim to directly minimize Δ_{DP} , which can introduce unintended biases. Specifically, considering Δ_{DP} at the edge level without accounting for the combinatorial nature of sensitive attributes can mask disparities between subgroup pairings.

In the case of a binary sensitive attribute (e.g., gender), the graph naturally decomposes into three edge types: $E_{s\text{-}s}$, $E_{s'\text{-}s'}$, and $E_{s'\text{-}s}$, corresponding to all possible node pairings. Existing methods often aggregate these into broader categories—such as intra-group ($E_{s\text{-}s} \cup E_{s'\text{-}s'}$) and inter-group ($E_{s'\text{-}s}$) edges—to enforce fairness constraints. However, we show that such aggregation may overlook systematic biases within the aggregated subgroups. For example, a model may consistently overpredict links of type $E_{s\text{-}s}$ relative to $E_{s'\text{-}s'}$, yet still satisfy Δ_{DP} under the aggregated grouping. More generally, the metric $\Delta_{\text{max}} = \max_{g_1,g_2\in\mathcal{G},g_1\neq g_2} |P(\hat{Y} \mid g_1) - P(\hat{Y} \mid g_2)|$ fails to penalize subgroup-level disparities when groupings $\mathcal{G}$ ignore pairwise composition. As a result, traditional dyadic fairness assumptions break down in the link prediction setting.

Example Figure 1 demonstrates how demographic parity can obscure subgroup imbalances. Here, $E_{s'\text{-}s}$ (gray) represents inter-group links (designated as the protected class), while $E_{s\text{-}s}$ (blue) and $E_{s'\text{-}s'}$ (red) represent intra-group links. Although the model disproportionately favors $E_{s\text{-}s}$ over $E_{s'\text{-}s'}$ in the left panel, both panels yield identical Δ_{DP} scores due to aggregation into E_{intra}. Even alternative groupings (e.g., $E_{s\text{-}s} \cup E_{s'\text{-}s}$ vs. $E_{s'\text{-}s'}$) fail to resolve the issue: underexposure of certain subgroups remains undetected.

Toward Subgroup-Sensitive Fairness. Over multiple prediction rounds, this skew can amplify structural disparities, akin to exposure bias (Singh and Joachims 2018), reinforcing phenomena such as the glass ceiling effect (Stoica, Riederer, and Chaintreau 2018). To address this issue, we propose a fairness criterion that preserves the distributional structure of edge types as observed in the original graph.

Property 1 (Non-Dyadic Distribution-Preserving Fairness)**.** A fairness metric should treat all sensitive attribute pairings as distinct subgroups, and aim to align the predicted edge distribution with that of the original graph. Formally, let $\boldsymbol{\pi} = (\pi_{s\text{-}s}, \pi_{s'\text{-}s}, \pi_{s'\text{-}s'})$ denote the empirical distribution of edge types defined by a binary sensitive attribute S. Let $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{s\text{-}s}, \hat{\pi}_{s'\text{-}s}, \hat{\pi}_{s'\text{-}s'})$ be the predicted distribution. Then, a fair link predictor should minimize $dist(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi})$, where $dist$ is a suitable divergence metric (e.g., KL divergence).

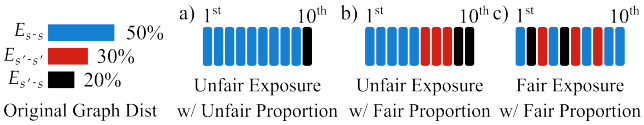

Figure 2: Δ_{DP} fails to capture exposure and subgroup proportion bias. Three models (a)–(c) output top-10 link predictions over a graph with original subgroup edge proportions of 50% $E_{s\text{-}s}$ (blue), 30% $E_{s'\text{-}s'}$ (red), and 20% $E_{s'\text{-}s}$ (black).

3.3 Bias Across Edge Groups (E_{intra} vs. E_{inter})

Fair link prediction methods commonly frame the task as binary classification, using utility metrics such as AUC-ROC to evaluate model performance (Li et al. 2022a, 2021a; Current et al. 2022a; Masrour et al. 2020). However, this evaluation protocol often misaligns with real-world applications, where link prediction decisions are based on ranked candidate lists (Tsioutsiouliklis et al. 2021b, 2022). Recent work addresses this limitation using ranking-based evaluation metrics (Mattos et al. 2025).

In terms of fairness, many approaches adopt the Δ_{DP} reduction between E_{inter} and E_{intra} as the primary fairness metric, which does not consider the *proportion* of pairs from each group in the final ranking. As a result, a biased model could produce a ranking heavily dominated by one pair type (e.g., E_{intra}) and still be evaluated as fair if the predicted scores are numerically similar. We argue that Δ_{DP} is inherently ranking-insensitive, making it an inadequate fairness measure for ranking-based tasks such as link prediction.

Even if the final ranking distribution of pairs of each type in a ranking approximates the distribution from the original graph, Δ_{DP} is also limited by being permutation invariant. This characteristic also enables another instance of exposure bias (Singh and Joachims 2018), but this time in a dyadic fashion. For instance, a (biased) link prediction algorithm can output a ranking of pairs that promotes an unfair number of E_{intra} ($E_{s\text{-}s} \bigcup E_{s'\text{-}s'}$) pairs to the top positions while still maintaining low values of demographic parity, characterizing a case of exposure disparity against E_{inter}. This is problematic in many applications, in which user attention is concentrated on the few top-ranked items.

Example Figure 2 presents three hypothetical top-10 rankings produced by different link prediction models, each evaluated over the same input graph with subgroup edge distribution $\boldsymbol{\pi} = (0.5, 0.2, 0.3)$. Model (a) exhibits severe bias by over-representing $E_{s\text{-}s}$ pairs, distorting both the group proportions and their relative exposure in top ranks. Model (b) preserves the global edge distribution but places E_{intra} edges disproportionately high, resulting in exposure bias despite matching overall proportions. Model (c) maintains both proportionality and fair exposure across groups. Despite these clear disparities, all three rankings could yield

the same Δ_{DP} score, which is blind to ranking order and group-specific position effects—highlighting the need for rank-aware fairness evaluation.

Property 2 (Rank awareness)**.** A fairness metric for link prediction should be sensitive to the proportion and rank of every type of pair. Specifically, it should ensure that any group, when ranked, does not systematically have higher or lower ranks compared to other groups. Let $\hat{\boldsymbol{\pi}}_k$ denote the distribution of each pair type in the top-k ranking of pairs outputted by a link prediction classifier, and $\mathcal{C}$ the set of candidate pairs. A rank-aware fairness metric should minimize $\min_{\hat{\boldsymbol{\pi}}_k} \sum_{k=1}^{|\mathcal{C}|} dist(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi})\delta_k$ where δ_k denotes the proportional exposure decay attributed to the k-th position on the ranking, which is usually monotonically decreasing.

Previous definitions of group exposure (Singh and Joachims 2018; Zehlike and Castillo 2020) consider the sum of the scores of items from the group weighted by the top-1 position bias. Such a definition assumes that items in top positions receive more attention (Joachims et al. 2005) and, thus, should have larger weights associated with their scores. Following the same assumption, Property 2 ensures that significant deviations from the original graph distribution in the top positions are more heavily penalized than deviations occurring at the bottom positions.

4 Distribution-Preserving and Ranking-Aware Fair Link Prediction

We first demonstrate how fair ranking evaluation metrics can address previous limitations associated with Δ_{DP} , and then we propose a simple post-processing algorithm for bias mitigation in link prediction methods.

4.1 Fair Ranking Metrics

The limitations of Δ_{DP} suggest the need for accounting not only for group proportions but also for exposure in ranked outputs. Inspired by previous works on fair ranking (Zehlike et al. 2017), we pose fair link prediction as a group-aware ranking problem over candidate links.

In standard fair ranking, items are ranked by relevance, and fairness is enforced by controlling the representation of protected groups in top positions. Although link prediction differs from classical ranking—most notably, edges do not carry scalar relevance scores—we observe that a model still induces a ranking over predicted edges. Moreover, this induced ranking affects which types of node pairs (e.g., $E_{s\text{-}s}$, $E_{s'\text{-}s'}$) are prioritized in downstream model decisions.

Our setting breaks away from dyadic fairness assumptions by considering group-level edge categories as ranking units. Given this framing, we seek metrics and methods that jointly account for (i) consistency with group-level edge proportions and (ii) equitable exposure across ranking positions.

In this fashion, we adopt the *Normalized Cumulative KL-Divergence* (NDKL) as our fairness metric. NDKL penalizes deviation between the cumulative exposure of group categories in the top-k ranked edges and their expected distribution, capturing exposure fairness and subgroup proportions.

Definition 2 (Normalized Cumulative KL-Divergence - NDKL)**.** Let $\mathcal{C}$ be the set of candidate pairs ranked by score, $\boldsymbol{\pi}$ the original proportion of each sensitive edge group in the graph, and $\hat{\boldsymbol{\pi}}_k$ the distribution of sensitive groups up to the k-th position of $\mathcal{C}$, the NDKL is:

$$\text{NDKL} = \frac{1}{Z} \sum_{k=1}^{|\mathcal{C}|} \frac{1}{\log_2(k+1)} D_{\text{KL}}\left(\hat{\boldsymbol{\pi}}_k \,\|\, \boldsymbol{\pi}\right),$$

where $D_{\text{KL}}(p||q)$ is the KL-Divergence between distributions p and q, and $Z = \sum_{i=1}^{|\mathcal{C}|} \frac{1}{\log_2(i+1)}$ is a normalizer.

Theorem 1. Let $\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2]$ be the target distribution over 3 sensitive groups, with $\sum_i \pi_i = 1$, and let $\hat{\boldsymbol{\pi}}_k$ denote the empirical distribution over the top-k ranked items. Under the constraint that the full ranking satisfies demographic parity (i.e., the overall empirical distribution matches $\boldsymbol{\pi}$), the NDKL score satisfies the following bounds: $0 \leq \text{NDKL} \leq \max_{i\in\{0,1,2\}} \log \frac{1}{\pi_i}$.

Theorem 1 establishes upper and lower bounds for the NDKL score under the assumption of demographic parity at the full-ranking level. To empirically validate this result, we construct controlled scenarios where candidate rankings are manipulated to explore different levels of exposure bias—ranging from completely fair to maximally skewed under a fixed $\boldsymbol{\pi}$. As shown in Section 5.2, the observed NDKL scores in these settings consistently lie within the theoretical bounds, confirming the expected behavior by responding sensitively to violations in exposure fairness.

NDKL satisfies both key properties required for fairness in link prediction. By using KL-Divergence over group distributions, it supports multiple sensitive groups without requiring dyadic aggregation (Property 1). It is also rank-aware, which ensures top-ranked exposures are more influential (Property 2). While NDKL was previously used in fair ranking (Geyik, Ambler, and Kenthapadi 2019) and is our metric of choice, other metrics could be used—provided they quantify divergence across multiple groups and incorporate exposure weighting in ranked outputs.

4.2 MORAL - Multi-Output Ranking Aggregation for Link Fairness

We introduce **MORAL** (Multi-Output Ranking Aggregation for Link fairness), a simple and scalable post-processing framework designed to improve fairness in link prediction. MORAL decouples group-wise predictions and enforces exposure parity through a ranking aggregation mechanism.

Specifically, MORAL trains three distinct link prediction models: $f_{s\text{-}s}$, $f_{s\text{-}s'}$, and $f_{s'\text{-}s'}$, each trained exclusively on edges corresponding to a specific sensitive group interaction. This decoupling mitigates group-specific utility disparities and prevents a single model from exhibiting imbalanced predictive performance across groups. In addition, MORAL remains computationally efficient even for sensitive attributes with larger cardinality, as each model processes only $\frac{|E_{train}|}{|S|\cdot\binom{|S|}{2}\cdot b}$ gradients per epoch, where $|E_{train}|$

denotes the number of training edges, $|S|$ the number of sensitive attribute categories, and b the batch size.

At inference time, MORAL aggregates predictions from the group-specific models into a unified ranking. This is accomplished by maintaining a running estimate of the exposure distribution across the three edge types and greedily selecting, at each rank position, the highest-scoring remaining edge from the model whose inclusion most reduces the cumulative KL divergence from a predefined target distribution $\boldsymbol{\pi}$ (see Algorithm 1 for pseudocode). This exposure-aware ranking procedure ensures that the final output approximates the desired group-wise exposure proportions.

Our greedy strategy solves the following fairness-prioritized objective: $\min_R \mathscr{L}_A(R)$ s.t. $\mathscr{L}_B(R) \leq \min_{R'} \mathscr{L}_B(R')$. Similarly, one could optimize a different objective that balances accuracy and fairness through a hyperparameter λ. A greedy algorithm can optimize this weighted-sum objective by minimizing the combined loss at each step (Celis, Straszak, and Vishnoi 2018). We opt for the constrained formulation above to explicitly prioritize fairness across all datasets, especially in imbalanced settings where exposure risks are more severe.

MORAL addresses the challenges identified in Section 3 by combining group-specific models with KL-guided ranking. Moreover, MORAL outputs group assignments per rank position, meaning it can be seamlessly paired with any fair ranking metric (like NDKL) that satisfies Properties 1 and 2.

Algorithm 1: MORAL: Multi-Output Ranking Aggregation for Link Fairness

Input:

- Candidate sets $\mathcal{C}_j = \{(u, v, \text{score})\}$ for each group $j \in \{0, 1, 2\}$ (sorted by descending score);
- Target distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$;
- Total output size n.

Output: Ranking list $\mathbf{R}$ of n predicted edges with assigned group labels

Initialize exposure counts: $\boldsymbol{c} \leftarrow (0, 0, 0)$ Initialize output ranking: $\mathbf{R} \leftarrow [\,]$

for $t \leftarrow 1$ **to** n **do**
  Initialize best objective: min_kl $\leftarrow \infty$, selected_group $\leftarrow -1$, selected_edge $\leftarrow$ None
  foreach *group* $j \in \{0, 1, 2\}$ *such that* $\mathcal{C}_j$ *is not empty* **do**
    Let $(u, v, \text{score}) \leftarrow$ top element in $\mathcal{C}_j$
    Temporarily update counts: $c'_j \leftarrow c_j + 1$, $q'_j \leftarrow \frac{c'_j}{t}$, $q_{j' \neq j} \leftarrow \frac{c_{j'}}{t}$
    Compute KL divergence: $D_{\mathrm{KL}}(\mathbf{q}' \| \boldsymbol{\pi})$
    if *this KL is lower than min_kl* **then**
      Update min_kl $\leftarrow D_{\mathrm{KL}}$, selected_group $\leftarrow j$, selected_edge $\leftarrow (u, v)$
    end
  end
  Append (selected_edge, selected_group) to $\mathbf{R}$
  Remove top element from $\mathcal{C}_{\text{selected\_group}}$
  Update $c_{\text{selected\_group}} \leftarrow c_{\text{selected\_group}} + 1$
end
return $\mathbf{R}$

Dataset	facebook	german	nba	pokec_n	pokec_z	credit
$\lvert V\rvert$	1045	1000	403	66569	67796	30000
$\lvert E\rvert$	18726	15220	7435	361934	432572	96165
Feat.	573	27	95	276	265	13
Attr.	Gen.	Age	Nat.	Gen.	Gen.	Age
Topo.	Periph.	Periph.	Periph.	Comm.	Comm.	Periph.

Table 1: Statistics from our six real-world datasets.

5 Experiments

We evaluate link prediction approaches across multiple datasets through fairness and link prediction metrics based on ranking. Considering our previous claims, we formulate three main research questions: *RQ1: To what extent does Δ_{DP} lead to hidden biases in the proportion of each group type in link prediction tasks? RQ2: Does ranking-awareness lead to a more faithful assessment of fairness in link prediction, compared to dyadic or proportion-based measures? RQ3: How does MORAL compare against existing fair link prediction approaches under the ranking metric?*

Baselines To enable a robust comparison for fair link prediction, we evaluate a diverse and competitive set of methods: a node embedding approach (FairWalk (Rahman et al. 2019)); pre-processing methods (EDITS (Dong et al. 2022), FairLP (Li et al. 2022b), FairEGM (Current et al. 2022b)); in-processing methods (GraphAIR (Ling et al. 2023), UGE (Wang et al. 2022a)); post-processing methods (DetConstSort (Geyik, Ambler, and Kenthapadi 2019), MORAL); and a task-specific fair link prediction method (FairAdj (Li et al. 2021a)). For consistency, all approaches use a GCN encoder (f_θ) with a dot-product decoder, following the framework in Section 2. We expect similar trends with other GNNs/S-GNNs (Wang, Yang, and Zhang 2024; Zhang and Chen 2018; Chamberlain et al. 2023). Hyperparameters follow each method's original recommendations.

Datasets. We conduct experiments on six real-world datasets considered in previous works (Dong et al. 2022; Li et al. 2022a; Current et al. 2022a). These datasets represent diverse domains and fairness contexts. The Credit dataset is a credit scoring network where nodes represent individuals and edges represent credit relationships, with age as sensitive attribute. The Facebook dataset is a social network where nodes are users and edges represent friendships, with gender as sensitive attribute. The German dataset is a credit approval network where nodes are individuals and edges link similar individuals, with gender as sensitive attribute. The NBA dataset is a network of NBA basketball players, where edges represent relationships between the athletes on Twitter, and nationality ('US' and 'overseas') is the sensitive attribute. The Pokec-n/z datasets consist of all the data from Pokec, a social network from Slovakia, in 2012, where nodes are users, edges are friendships, and gender is the sensitive attribute. Table **??** presents statistics for each dataset, including the number of nodes, edges, and the distribution of sensitive attributes. We selected datasets from two different graph distributions (community and periphery).

Evaluation Metrics. We consider two metrics, adhering

to the evaluation scheme of ranking tasks. We adopt NDKL as our fairness ranking metric and *prec@k* as our ranking-based performance metric. We also establish comparisons between NDKL and Δ_{DP} in Section 5.2.

Implementation Details We implement all models using PyTorch Geometric (Fey and Lenssen 2019) and the PyGDebias (Dong et al. 2023a). We run experiments using NVIDIA A40 GPUs, fixing random seeds, and run each experiment 3 times. Random edge splits are set as 70/10/20% for training/validation/testing. The sensitive attribute distribution is preserved across all splits. We adopt the Adam optimizer with learning rate 0.0003 for all experiments.

5.1 Hidden Bias in Dyadic Fairness

We first aim to answer RQ1, demonstrating the limitations of dyadic fairness for link prediction. We analyze the distributions of types of pairs in the top-k of each method, independent of the order of the elements. In Figure 3, we compare the proportions of pairs in the top-k ($k = 100$) of each baseline. Our approach obtains the closest approximation to the target distribution, while other methods overestimate one pair type at the detriment of another, despite obtaining low values of Δ_{DP} . This result demonstrates the limitation of Δ_{DP} in detecting underrepresented aggregated groups, and sheds light on the importance of adhering to Property 1.

5.2 Demographic Parity Gap

To answer RQ2, we demonstrate the effect of ranking on fairness in link prediction by fixing the proportion of each pair type. First, we compute the required proportions of each pair type for optimal Δ_{DP} . Then, we fix these proportions, but consider the worst and best possible permutations of these candidate pairs in terms of NDKL. To isolate the fairness component from the utility measurement, we assume in this experiment that the pairs being ranked are all positive, meaning that regardless of the permutation of the pairs, both rankings will obtain $prec@k = 100\%$. We expose the NDKL gap for varying-sized rankings in the extended manuscript. Despite relevant exposure bias between the worst and best rankings, the value of Δ_{DP} is the same. This indicates the necessity of incorporating Property 2 into fair link prediction metrics.

5.3 Baselines Ranking Comparison

We compare our method against UGE, EDITS, FairAdj, GRAPHAIR, FairWalk, DELTR, and DetConstSort. We adopt *NDKL* as the fairness metric and *prec@1000* as the utility metric. For EDITS, we use the dot product between the node embeddings obtained by training a GCN on the fair graph generated as the pair scores. For GRAPHAIR, UGE, and FairWalk, the scores are obtained through the dot product of the fair node embeddings outputted.

We show our results in Table 2. MORAL achieves the fairest rankings while still maintaining high link prediction performance. Both periphery and community graph types are challenging for all methods. We highlight the large fairness improvements obtained by our approach across all datasets, particularly in Credit, German, and NBA.

6 Related Work

We situate MORAL within two major areas: fair graph representation learning and fair ranking. Our approach is an example of a post-processing method, which has proven effective in fair ranking tasks (Xian, Yin, and Zhao 2023; Tifrea et al. 2024; Gorantla, Deshpande, and Louis 2021; Zehlike et al. 2017; Li et al. 2021b), and aligns with decoupling classifiers for group fairness, which has been successfully explored in the past (Dwork et al. 2018).

6.1 Fair Graph Representation Learning

Pre-processing. Methods such as EDITS (Dong et al. 2022) and ALFR (Edwards and Storkey 2016) mitigate bias before training by modifying node features or the graph structure. Despite their effectiveness, these methods can be computationally expensive and face scalability issues.

In-processing. FairGNN (dai 2021), NIFTY (Agarwal, Lakkaraju, and Zitnik 2021), GRAPHAIR (Ling et al. 2023), FairVGNN (Wang et al. 2022b) , and UGE (Wang et al. 2022a) integrate fairness during training using adversarial learning or graph augmentations. These methods assume fairness can be enforced at the node embedding level, which may encounter expressivity power limitations, and not generalize to pairwise tasks like link prediction.

Fair Embeddings. FairWalk (Rahman et al. 2019) and RELIANT (Dong et al. 2023b) focus on debiasing unsupervised embeddings via neighbor sampling and proxy removal, respectively. However, these methods rely heavily on the structural properties of the graph and are not designed for post-hoc ranking fairness.

6.2 Fair Ranking and Information Retrieval

Post-processing approaches like FA*IR (Zehlike et al. 2017) and DetConstSort (Geyik, Ambler, and Kenthapadi 2019) re-rank outputs to satisfy fairness constraints. Though model-agnostic, they can be computationally intensive and inflexible. In-processing ranking methods, such as DELTR (Zehlike and Castillo 2020), policy learning (Singh and Joachims 2019; Yadav, Du, and Joachims 2021), and constrained optimization via SPOFR (Kotary et al. 2022), jointly optimize fairness and relevance. However, they typically target dyadic ranking tasks in learn-to-rank settings, which is misaligned with the structural nature of the link prediction task considered in this work.

7 Conclusion

Fairness in link prediction is a relevant problem addressed by a diverse set of previous approaches in the literature. In this work, we scrutinize the main assumptions behind how previous works evaluate the fairness of link prediction models. In particular, we shed light on pitfalls related to naively adopting dyadic fairness notions for link prediction and how this approach is prone to a hidden form of bias within aggregated subgroups. Further, we demonstrate how not capturing ranking notions in the fairness metric can potentially prevent a given metric from capturing exposure bias.

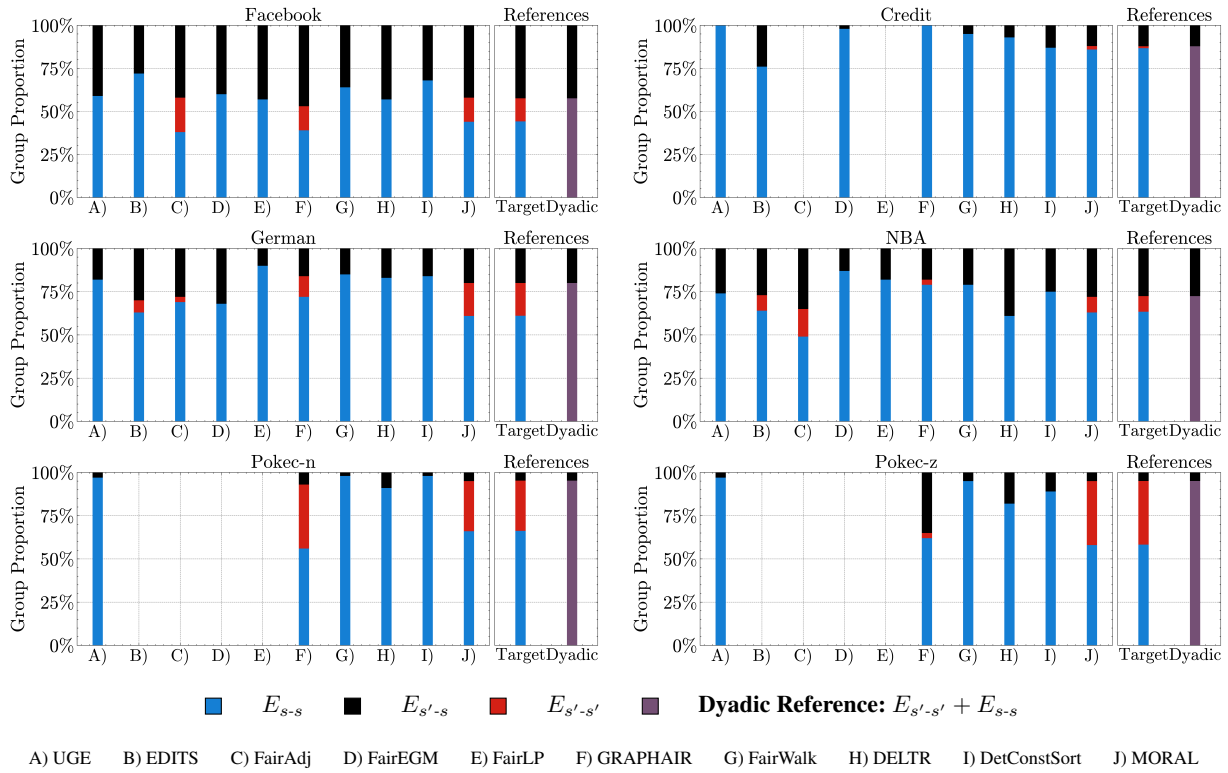

Figure 3: Proportions of pair types in the top-100 predictions by method, compared against the original graph distribution and an optimal dyadic fairness reference. Colors: $E_{s\text{-}s}$ (blue), $E_{s'\text{-}s}$ (black), and $E_{s'\text{-}s'}$ (red). In the dyadic fairness reference, purple represents the combined proportion of $E_{s'\text{-}s'}$ and $E_{s\text{-}s}$ pairs. Missing bars indicate an OOM error.

Method		**Facebook**	**Credit**	**German**	**NBA**	**Pokec-n**	**Pokec-z**
UGE	NDKL	0.05 ± 0.00	0.80 ± 0.07	0.08 ± 0.03	0.07 ± 0.02	0.06 ± 0.00	0.06 ± 0.00
	prec@1000	$\underline{0.97 \pm 0.00}$	$\underline{1.00 \pm 0.00}$	0.69 ± 0.01	0.58 ± 0.01	0.90 ± 0.04	0.91 ± 0.04
EDITS	NDKL	0.21 ± 0.07	0.04 ± 0.02	0.08 ± 0.02	0.08 ± 0.01	OOM	OOM
	prec@1000	0.96 ± 0.00	0.36 ± 0.17	0.42 ± 0.20	0.49 ± 0.00	OOM	OOM
GRAPHAIR	NDKL	0.13 ± 0.03	0.67 ± 0.22	0.07 ± 0.02	0.09 ± 0.01	0.09 ± 0.03	0.26 ± 0.25
	prec@1000	0.96 ± 0.01	$\underline{1.00 \pm 0.00}$	0.73 ± 0.01	0.69 ± 0.01	0.97 ± 0.03	$\underline{1.00 \pm 0.00}$
FairEGM	NDKL	0.09 ± 0.01	0.11 ± 0.00	0.05 ± 0.01	0.07 ± 0.01	OOM	OOM
	prec@1000	0.97 ± 0.00	1.00 ± 0.00	0.62 ± 0.00	0.60 ± 0.01	OOM	OOM
FairLP	NDKL	0.18 ± 0.00	OOM	0.06 ± 0.00	0.20 ± 0.00	OOM	OOM
	prec@1000	0.99 ± 0.00	OOM	0.97 ± 0.00	0.86 ± 0.00	OOM	OOM
FairWalk	NDKL	0.06 ± 0.01	0.06 ± 0.03	0.11 ± 0.02	0.06 ± 0.01	0.07 ± 0.01	0.07 ± 0.00
	prec@1000	0.96 ± 0.00	$\underline{1.00 \pm 0.00}$	0.94 ± 0.00	0.55 ± 0.01	$\underline{1.00 \pm 0.00}$	$\underline{1.00 \pm 0.00}$
FairAdj	NDKL	0.10 ± 0.05	OOM	0.10 ± 0.01	0.11 ± 0.05	OOM	OOM
	prec@1000	0.42 ± 0.01	OOM	0.54 ± 0.01	0.50 ± 0.01	OOM	OOM
DetConstSort	NDKL	0.15 ± 0.00	0.06 ± 0.00	0.04 ± 0.00	0.09 ± 0.00	0.07 ± 0.00	0.23 ± 0.00
	prec@1000	0.00 ± 0.00	0.00 ± 0.00	0.55 ± 0.00	0.21 ± 0.00	0.07 ± 0.00	0.01 ± 0.00
DELTR	NDKL	0.10 ± 0.03	0.03 ± 0.00	0.09 ± 0.06	0.09 ± 0.02	0.23 ± 0.23	0.22 ± 0.20
	prec@1000	0.91 ± 0.05	0.56 ± 0.29	0.31 ± 0.44	0.43 ± 0.24	0.65 ± 0.01	0.48 ± 0.28
MORAL	NDKL	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$
	prec@1000	0.95 ± 0.01	$\underline{1.00 \pm 0.00}$	$\underline{0.96 \pm 0.00}$	$\underline{0.80 \pm 0.00}$	0.98 ± 0.00	$\underline{0.98 \pm 0.00}$

Table 2: Fairness performance comparison of all approaches considered ($k = 1000$). Lower *NDKL* and higher *prec@1000* are better. Best NDKL and *prec@1000* values are in bold and underline, respectively.

Acknowledgements

We acknowledge the support by the US Department of Transportation Tier-1 University Transportation Center (UTC) Transportation Cybersecurity Center for Advanced Research and Education (CYBER-CARE) (Grant No. 69A3552348332), and the Rice Ken Kennedy Institute.

References

2021. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *WSDM*.

Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a Unified Framework for Fair and Stable Graph Representation Learning. In *UAI*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Calvo-Armengol, A.; and Jackson, M. O. 2004. The effects of social networks on employment and inequality. *American economic review*, 94(3): 426–454.

Celis, L. E.; Straszak, D.; and Vishnoi, N. K. 2018. Ranking with Fairness Constraints. In *ICALP*.

Chamberlain, B. P.; Shirobokov, S.; Rossi, E.; Frasca, F.; Markovich, T.; Hammerla, N.; Bronstein, M. M.; and Hansmire, M. 2023. Graph Neural Networks for Link Prediction with Subgraph Sketching. In *ICLR*.

Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The echo chamber effect on social media. *PNAS*, 118(9): e2023301118.

Current, S.; He, Y.; Gurukar, S.; and Parthasarathy, S. 2022a. FairEGM: Fair Link Prediction and Recommendation via Emulated Graph Modification. In *EAMO*.

Current, S.; He, Y.; Gurukar, S.; and Parthasarathy, S. 2022b. Fairegm: fair link prediction and recommendation via emulated graph modification. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–14.

Dai, E.; Zhao, T.; Zhu, H.; Xu, J.; Guo, Z.; Liu, H.; Tang, J.; and Wang, S. 2024. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6): 1011–1061.

Dong, Y.; Liu, N.; Jalaian, B.; and Li, J. 2022. EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks. In *WebConf*.

Dong, Y.; Ma, J.; Wang, S.; Chen, C.; and Li, J. 2023a. Fairness in graph mining: A survey. *TKDE*.

Dong, Y.; Zhang, B.; Yuan, Y.; Zou, N.; Wang, Q.; and Li, J. 2023b. RELIANT: Fair Knowledge Distillation for Graph Neural Networks. In *SDM*.

Draws, T.; Tintarev, N.; and Gadiraju, U. 2021. Assessing viewpoint diversity in search results using ranking fairness metrics. *SIGKDD Explorations*, 23(1): 50–58.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *FAccT*.

Edwards, H.; and Storkey, A. 2016. Censoring representations with an adversary. In *ICLR*.

Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR RLGM*.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *SIGKDD*.

Gorantla, S.; Deshpande, A.; and Louis, A. 2021. On the Problem of Underranking in Group-Fair Ranking. In *ICML*.

Gupta, S.; Wang, H.; Lipton, Z.; and Wang, Y. 2021. Correcting exposure bias for link recommendation. In *ICML*.

Han, X.; Jiang, Z.; Jin, H.; Liu, Z.; Zou, N.; Wang, Q.; and Hu, X. 2023. Retiring ΔDP: New Distribution-Level Metrics for Demographic Parity. *TMLR*.

Hofstra, B.; Corten, R.; Van Tubergen, F.; and Ellison, N. B. 2017. Sources of segregation in social networks: A novel approach using Facebook. *American Sociological Review*, 82(3): 625–656.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*.

Karimi, F.; Génois, M.; Wagner, C.; Singer, P.; and Strohmaier, M. 2018. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8(1): 11077.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*.

Kleinberg, J.; Ryu, E.; and Tardos, É. 2024. Calibrated recommendations for users with decaying attention. In *SAGT*.

Kotary, J.; Fioretto, F.; Van Hentenryck, P.; and Zhu, Z. 2022. End-to-end learning for fair ranking systems. In *WebConf*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *NeurIPS*.

Laclau, C.; Redko, I.; Choudhary, M.; and Largeron, C. 2021. All of the fairness for edge prediction with optimal transport. In *AISTATS*.

Li, P.; Wang, Y.; Zhao, H.; Hong, P.; and Liu, H. 2021a. On dyadic fairness: Exploring and mitigating bias in graph connections. In *ICLR*.

Li, Y.; Chen, H.; Fu, Z.; Ge, Y.; and Zhang, Y. 2021b. User-oriented fairness in recommendation. In *WebConf*.

Li, Y.; Wang, X.; Ning, Y.; and Wang, H. 2022a. FairLP: Towards Fair Link Prediction on Social Network Graphs. In *ICWSM*.

Li, Y.; Wang, X.; Ning, Y.; and Wang, H. 2022b. Fairlp: Towards fair link prediction on social network graphs. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 628–639.

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

Ling, H.; Jiang, Z.; Luo, Y.; Ji, S.; and Zou, N. 2023. Learning Fair Graph Representations via Automated Data Augmentations. In *ICLR*.

Luo, Z.; Huang, H.; Lian, J.; Song, X.; Xie, X.; and Jin, H. 2023. Cross-links matter for link prediction: rethinking the debiased GNN from a data perspective. In *NeurIPS*.

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

Mattos, J.; Huang, Z.; Kosan, M.; Singh, A.; and Silva, A. 2025. Attribute-Enhanced Similarity Ranking for Sparse Link Prediction. In *SIGKDD*.

McDonald, S.; Lin, N.; and Ao, D. 2009. Networks of opportunity: Gender, race, and job leads. *Social Problems*, 56(3): 385–402.

Pan, L.; Shi, C.; and Dokmanić, I. 2022. Neural Link Prediction with Walk Pooling. In *ICLR*.

Rahman, T.; Surma, B.; Backes, M.; and Zhang, Y. 2019. Fairwalk: Towards Fair Graph Embedding. In *IJCAI*.

Singh, A.; and Joachims, T. 2018. Fairness of Exposure in Rankings. In *SIGKDD*.

Singh, A.; and Joachims, T. 2019. Policy learning for fairness in ranking. In *NeurIPS*.

Stoica, A.-A.; Litvak, N.; and Chaintreau, A. 2024. Fairness rising from the ranks: HITS and pagerank on homophilic networks. In *WebConf*.

Stoica, A.-A.; Riederer, C.; and Chaintreau, A. 2018. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *WebConf*.

Tifrea, A.; Lahoti, P.; Packer, B.; Halpern, Y.; Beirami, A.; and Prost, F. 2024. Frappé: A group fairness framework for post-processing everything. In *ICML*.

Tsioutsiouliklis, S.; Pitoura, E.; Semertzidis, K.; and Tsaparas, P. 2022. Link recommendations for PageRank fairness. In *WebConf*.

Tsioutsiouliklis, S.; Pitoura, E.; Tsaparas, P.; Kleftakis, I.; and Mamoulis, N. 2021a. Fairness-aware pagerank. In *WebConf*.

Tsioutsiouliklis, S.; Pitoura, E.; Tsaparas, P.; Kleftakis, I.; and Mamoulis, N. 2021b. Fairness-aware pagerank. In *WebConf*.

Wang, N.; Lin, L.; Li, J.; and Wang, H. 2022a. Unbiased graph embedding with biased graph observations. In *WebConf*.

Wang, X.; Yang, H.; and Zhang, M. 2024. Neural Common Neighbor with Completion for Link Prediction. In *ICLR*.

Wang, Y.; Zhao, Y.; Dong, Y.; Chen, H.; Li, J.; and Derr, T. 2022b. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *SIGKDD*.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and optimal classification via post-processing. In *ICML*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How powerful are graph neural networks? In *ICLR*.

Yadav, H.; Du, Z.; and Joachims, T. 2021. Policy-Gradient Training of Fair and Unbiased Ranking Functions. In *SIGIR*.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM*.

Zehlike, M.; and Castillo, C. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *WebConf*.

Zhang, M.; and Chen, Y. 2018. Link Prediction Based on Graph Neural Networks. In *NeurIPS*.

Zhu, Y.; Li, J.; Bian, Y.; Zheng, Z.; and Chen, L. 2024a. One fits all: Learning fair graph neural networks for various sensitive attributes. In *SIGKDD*.

Zhu, Y.; Li, J.; Chen, L.; and Zheng, Z. 2024b. The devil is in the data: Learning fair graph neural networks via partial knowledge distillation. In *WSDM*.

Zhu, Z.; Zhang, Z.; Xhonneux, L.-P.; and Tang, J. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *NeurIPS*.