Fairness in Link Prediction Beyond Demographic Parity: A Reproducibility Study

Anonymous authors
Paper under double-blind review

Abstract

In fair ranked link prediction, demographic parity (Δ_{DP}) is a common fairness criterion, yet Mattos et al. (2025) argue that it obscures exposure bias induced by ranking order. In this study, we reproduce and validate their claims by confirming (i) the limitations of Δ_{DP}, (ii) the advantages of the rank-aware, exposure-based Normalized Discounted KL-divergence (NDKL) metric, and (iii) the effectiveness of MORAL, a post-processing algorithm for debiasing ranked outputs while maintaining competitive utility. Beyond reproduction, we assess robustness through synthetic asymmetric homophily stress-tests, categorical sensitive attributes, and alternative fairness and utility metrics, including subgroup-pair-adapted Attention-Weighted Rank Fairness (AWRF). Empirical results demonstrate that exposure-based fairness reveals disparities hidden by Δ_{DP} and that MORAL reduces such bias with minimal utility loss across diverse settings and datasets. We release a corrected, reproducible implementation available at `https://github.com/unknown-gitter/reproducing-MORAL`.

1 Introduction

Fairness in link prediction aims to reduce disparities in predicted link probabilities across groups (Chen et al., 2024). In social-network recommendations, link predictors shape who becomes visible and connected (Ferrara et al., 2022) and can induce feedback loops that amplify homophily and disadvantage minority groups (Karimi et al., 2018). Importantly, even small biases can have large downstream impacts on minority groups (Singh & Joachims, 2018). This is especially concerning because link predictors are trained on historical graph data and can therefore inherit and amplify existing structural biases (Dai et al., 2024).

At the model level, such amplification can lead to prediction distortions, including exposure skews in ranked outputs (Gupta et al., 2021), degree-driven popularity effects (Subramonian et al., 2023), and hidden subgroup disparities (Kearns et al., 2018). Mitigating these biases is challenging because candidate links are structurally dependent (non-i.i.d.) and subject to temporal evolution and feedback (Stoica et al., 2018; Zhang et al., 2024). In addition, limited expressivity of graph neural networks (GNNs) can make symmetric neighbourhoods with different sensitive groups indistinguishable (Mattos et al., 2025). These concerns have motivated recent work on defining and enforcing fairness in link prediction (Chen et al., 2024).

In this spirit, Li et al. (2021) propose a *dyadic* definition of fairness, formalising fair link prediction as group fairness by enforcing demographic parity between intra- and inter-group link predictions. However, recent work by Mattos et al. (2025) suggests that aggregating subgroups in dyadic fairness obscures *within*-group disparities, even when demographic parity is satisfied. This implies that, despite the use of dyadic fairness based on demographic parity in recent literature (e.g., Current et al., 2022; Luo et al., 2023), the metric may be insufficient for fair link prediction. To support these claims, Mattos et al. (2025) (i) formalise and empirically validate limitations of dyadic demographic parity, (ii) introduce the rank-aware, exposure-based *Normalized Discounted KL-divergence* (**NDKL**) metric, and (iii) propose **MORAL**, a post-processing algorithm that debiases ranked outputs by exposure-aware aggregation of decoupled link predictions.

In this study, we reproduce and confirm these claims and release a corrected implementation. Beyond reproduction, we extend the original work through (i) an asymmetric homophily stress-test with synthetic graphs, (ii) a metric robustness analysis, and (iii) experiments with categorical sensitive attributes. Overall, our

results support exposure-based, rank-aware fairness in link prediction by showing that dyadic demographic parity can obscure subgroup-pair exposure disparities and that MORAL reduces them across diverse settings.

2 Scope of reproducibility

The objective of this study is to reproduce and extend the key findings of Mattos et al. (2025). The paper makes two main contributions: (i) it introduces the **NDKL**, an exposure-based fairness metric that detects biases missed by demographic parity; and (ii) it proposes **MORAL**, a post-processing algorithm for debiasing outputs ranked outputs. Along with these, the paper makes three key claims, labeled (**C1**–**C3**) for reference:

- **Claim 1 (C1): Demographic parity hides within-group exposure bias.** The authors reason and empirically show how aggregating across subgroup-pair types in dyadic fairness obscures *within*-group disparities, even while group-level fairness and demographic parity are satisfied.
- **Claim 2 (C2): NDKL overcomes the limitations of demographic parity.** Unlike demographic parity, NDKL is a non-dyadic, rank-aware metric. It explicitly treats all subgroup-pairings and detects exposure disparities that demographic parity fails to detect.
- **Claim 3 (C3): MORAL achieves strong fairness-utility trade-offs.** The paper shows how MORAL consistently mitigates previously undetected exposure biases, as measured by NDKL. Meanwhile, MORAL keeps its competitive utility relative to baselines, as measured by Precision@K.

To confirm these claims (**C1**–**C3**), we reproduce the core experimental setup by Mattos et al. (2025) by evaluating MORAL across all datasets using the proposed fairness and utility metrics. Moreover, to further assess their robustness and generality, we extend the original evaluation along variations in homophily, evaluation metrics, and sensitive attribute cardinality:

- **Asymmetric homophily stress-test:** Increasing homophily progressively reduces certain subgroup-pair types, thereby stressing MORAL's ability to reallocate exposure. We evaluate how MORAL's fairness-utility trade-offs evolve as subgroup interactions become more constrained.
- **Metric robustness analysis:** Since MORAL is explicitly optimised w.r.t. NDKL, we additionally evaluate fairness using a metric that differs in how subgroup exposure is aggregated. For utility, we add a rank-aware measure that accounts for the positions of correct predictions.
- **Categorical sensitive attributes:** We evaluate whether the proposed framework generalises beyond binary sensitive attributes to categorical sensitive attributes, assessing the practical validity of the paper's claim that MORAL extends to higher-cardinality settings.

Specifically, the homophily stress-test primarily explores **C3**, the metric robustness analysis targets **C2** & **C3**, and the attribute cardinality extension tests whether **C1**–**C3** generalise beyond binary sensitive attributes.

3 Methodology

This section describes the methodology and experimental setup required to reproduce and evaluate the claims by Mattos et al. (2025). In Section 3.6, we describe the extensions that go beyond reproduction.

3.1 Preliminaries and ranking setup

We first formalise link prediction as a ranking problem, then introduce how sensitive attributes induce subgroup-pair types, and end by motivating the role of exposure in ranked outputs.

Link prediction as ranking. Let $\mathcal{G} = (\mathbb{V}, \mathbb{E}, \mathbf{s})$ be a graph with node set $\mathbb{V}$, edge set $\mathbb{E}$, and sensitive attribute vector $\mathbf{s} \in \{0,1\}^n$ with s_v as the sensitive attribute of node v. Let $\mathbb{C} \subseteq \mathbb{V} \times \mathbb{V}$ denote the set of candidate node pairs. Link prediction is considered as scoring candidate node pairs: each pair $(u, v) \in \mathbb{C}$ is

mapped by a scoring function $f : \mathbb{V} \times \mathbb{V} \to [0, 1]$ whose outputs can be interpreted as probabilities. Sorting all candidate pairs in decreasing score order yields a ranking $\mathbf{R} = (R_1, R_2, \dots)$ where $\mathbf{R}$ is an ordered set and each $R_k \in \mathbb{C}$ denotes the candidate node pair at rank k. We evaluate the models on $\mathbf{R}$'s top-K list.

Sensitive attributes and subgroup-pair types. The sensitive attribute divides candidate node pairs into *subgroup-pair types* based on the attributes of the two nodes involved. For binary sensitive attributes $s \in \{0, 1\}$ with unordered pairs, this yields three pair types: $\mathbb{E}_{0\text{-}0} = \{(u, v) \in \mathbb{C} \mid s_u = 0, s_v = 0\}$, $\mathbb{E}_{1\text{-}1} = \{(u, v) \in \mathbb{C} \mid s_u = 1, s_v = 1\}$, and $\mathbb{E}_{0\text{-}1} = \{(u, v) \in \mathbb{C} \mid s_u \neq s_v\}$. Of these, $\mathbb{E}_{0\text{-}0}$ and $\mathbb{E}_{1\text{-}1}$ belong to the intra-group $\mathbb{E}_{\text{intra}} = \mathbb{E}_{0\text{-}0} \cup \mathbb{E}_{1\text{-}1}$, while $\mathbb{E}_{0\text{-}1}$ belongs to the inter-group $\mathbb{E}_{\text{inter}} = \mathbb{E}_{0\text{-}1}$. This decomposition into pair types is critical (Mattos et al., 2025), since dyadic fairness operates at the coarse group-level and aggregates them by default (Li et al., 2021), thereby obscuring finer-grained *within*-group disparities.

Exposure in rankings. Because link prediction outputs are often used downstream as ranked lists sorted by score, fairness concerns are based on exposure: pairs ranked higher receive disproportionately more attention (Abdollahpouri, 2019). Following the exposure definition of Mattos et al. (2025), we relate each rank k with a monotonically decreasing, logarithmically decaying factor

$$\delta_k = \frac{1}{\log_2(k + 1)}, \tag{1}$$

such that deviations in predictions of higher ranks are penalised more heavily than those at lower ones.

3.2 Fairness and utility metrics

This section formalises the fairness and utility metrics used by Mattos et al. (2025). We start by defining demographic parity in dyadic link prediction and its limitations w.r.t. exposure bias (**C1**). We then present NDKL as a non-dyadic, rank-aware alternative (**C2**), followed by the utility metric.

Demographic parity (Δ_{DP}). Demographic parity (Δ_{DP}) compares predictions $\hat{y} \in \mathbb{Y}$, where $\mathbb{Y}$ is the set of possible classes, across the previously defined intra-group $\mathbb{E}_{\text{intra}}$ and inter-group $\mathbb{E}_{\text{inter}}$, giving

$$\Delta_{\text{DP}} = \left| p(\hat{y} \mid (u, v) \in \mathbb{E}_{\text{intra}}) \; - \; p(\hat{y} \mid (u, v) \in \mathbb{E}_{\text{inter}}) \right|. \tag{2}$$

Equivalently, Δ_{DP} can be interpreted as requiring independence between the predicted links and the sensitive attribute at the level of $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$ (Gajane & Pechenizkiy, 2017). While this dyadic aggregation is used in prior work (e.g., Li et al., 2021; Current et al., 2022; Li et al., 2022; Luo et al., 2023), it can obscure disparities within the aggregated $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$. As a result, exposure imbalances can persist even when Δ_{DP} is satisfied (Mattos et al., 2025), thus providing a theoretical basis for Claim **C1**.

Fairness metric: Normalised Discounted KL-divergence (NDKL). To overcome Δ_{DP}'s limitations, Mattos et al. (2025) propose two properties a metric for fairness in link prediction should have:

- **Property 1: Non-dyadic distribution-preserving fairness.** A fairness metric should treat all subgroup-pair types distinctly and aim to align their predicted distribution $\hat{\boldsymbol{\pi}}$ with their original $\boldsymbol{\pi}$.
- **Property 2: Rank-awareness.** A fairness metric should be sensitive to the proportion and position-dependent exposure of every type of pair. Specifically, it should penalise deviations from $\boldsymbol{\pi}$ according to $\sum_{k=1}^{|\mathbb{C}|} \text{dist}(\hat{\boldsymbol{\pi}}_k(\mathbf{R}), \boldsymbol{\pi})\delta_k$, where $\hat{\boldsymbol{\pi}}_k$ denotes the distribution per pair type in the top-k prefix, $\text{dist}(\cdot, \cdot)$ is some distance function, and δ_k enforces rank-awareness, here by following Eq. 1.

Following these properties, the NDKL (first introduced by Geyik et al. (2019)) is defined as

$$\text{NDKL} = \frac{1}{\delta_i} \sum_{k=1}^{|\mathbb{C}|} \delta_k D_{\text{KL}}(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi}), \tag{3}$$

where the distance function is the KL-divergence and δ_i is a normaliser following Eq. 1. Hence, NDKL satisfies Properties 1 & 2, reflecting the theoretical motivation for Claim **C2** of Mattos et al. (2025).

Utility metric: Precision@1000. We adopt Precision@K, the fraction of true edges among the top-K candidate pairs, using $K = 1000$. Unlike the NDKL, it is insensitive to subgroup composition and exposure.

3.3 MORAL post-processing framework

This section describes the framework behind *Multi-Output Ranking Aggregation for Link Fairness* (MORAL). MORAL aims to overcome three deficiencies of dyadic fairness in link prediction (Mattos et al., 2025): (i) limited 1-WL expressivity of standard (non-decoupled; single-model) message-passing GNNs (Xu et al., 2018) that can blur subgroup-pair-specific asymmetries, (ii) subgroup-pair imbalances hidden by aggregation of $\mathbb{E}_{\text{intra}}$ or $\mathbb{E}_{\text{inter}}$ under Δ_{DP}, and (iii) Δ_{DP}'s permutation invariance to ranking order that ignores exposure.

Decoupled subgroup-pair predictors. MORAL requires decoupled link predictors per subgroup-pair type; in the binary setting of Section 3.1, that means models $f_{0\text{-}0}, f_{1\text{-}1}, f_{0\text{-}1}$ for $\mathbb{E}_{0\text{-}0}$, $\mathbb{E}_{1\text{-}1}$, and $\mathbb{E}_{0\text{-}1}$, respectively, where each model is trained solely on edges of its type and outputs logits for these edges. Following training, the candidate pairs of each subgroup-pair type $(u_i, v_i) \in \mathbb{C} \cap \mathbb{E}_j$, where $j \in \{0\text{-}0, 1\text{-}1, 0\text{-}1\}$, are sorted using the logits into a decreasingly ordered set $\mathbf{C}_j = ((u_1, v_1, \text{score}_1), (u_2, v_2, \text{score}_2), \dots)$.

Greedy KL-guided aggregation. MORAL's objective is to, given the ordered sets $\mathbf{C}_{0\text{-}0}, \mathbf{C}_{1\text{-}1}, \mathbf{C}_{0\text{-}1}$ and a target exposure distribution $\boldsymbol{\pi}$ over subgroup-pair types, construct a ranking $\mathbf{R}$. $\boldsymbol{\pi}$ is typically set to the observed proportions in the original graph, reflecting **Property 1** (non-dyadic distribution-preserving fairness). Notably, $\boldsymbol{\pi}$ is a design choice that, by choosing distribution-preserving fairness, aims to prevent algorithmic bias, rather than it being a universally correct definition of fairness (Mitchell et al., 2021).

At each rank position, MORAL selects the current top remaining edge from each $\mathbf{C}$ and chooses the one that minimises $D_{\text{KL}}(\boldsymbol{q}' \,\|\, \boldsymbol{\pi})$, where $\boldsymbol{q}'$ is the updated distribution if that subgroup-pair type were chosen (see Appendix A for the full algorithm). This greedy KL-divergence-minimising procedure mirrors the subgroup-aware, exposure-based mechanism of the NDKL (Eq. 3) — a relationship we further detail in Section 3.6.2.

3.4 Datasets

Six graph datasets[1] on social networks are used: Facebook; Credit; German; NBA; Pokec-n; & Pokec-z; with their details listed in Table 1. Pokec-n and Pokec-z are samples from Pokec (Takac & Zabovsky, 2012), a Slovakian social network. The datasets' topologies are of type *Community* and *Periphery*. (In Appendix E, we discuss MORAL's performance on two datasets with a different structure/mixing pattern.) Fairness concerns are particularly relevant for social networks, where factors such as historical inequalities and homophily can be amplified through algorithmic prediction (Stoica et al., 2018; Barocas et al., 2023).

Table 1: Social graph datasets used by Mattos et al. (2025). More details are listed in Appendix C.1.

Dataset	Network type	Node	Edge	Sensitive attribute (2-ary)
Credit	Credit network	Individual	Credit relationship	Age (binned)
Facebook	Online medium	User	Friendship	Gender (male vs. female)
German	Credit approval	Individual	Similarity link	Gender (male vs. female)
NBA	Athletes on Twitter	Athlete	Twitter connection	Nationality (US vs. non-US)
Pokec-n	Online medium	User	Friendship	Gender (male vs. female)
Pokec-z	Online medium	User	Friendship	Gender (male vs. female)

3.5 Experimental setup and code

This section describes the experimental pipeline for reproducing the experiments by Mattos et al. (2025). While the authors provide an open-source repository[2] and report full out-of-the-box reproducibility, we find the codebase omits several elements required for reproduciblity and is inconsistent with the paper's formal definitions. Below, we summarise components that are directly reused, describe necessary corrections and extensions, list reimplemented baselines, and conclude with reproducibility details. Together, these corrections align the implementation with the paper's definitions for reliable evaluation of Claims **C1**–**C3**.

[1]The datasets are retrieved from: `https://github.com/yushundong/PyGDebias` (Dong et al., 2023).
[2]Mattos et al. (2025)'s GitHub code repository: `https://github.com/joaopedromattos/MORAL`.

Reused training pipeline. We adopt the authors' implementation for training the decoupled predictors used by MORAL. Out of the box, the repository provides dataset loaders for all datasets in Table 1 and an extensive model class for training multiple graph neural network architectures (Gori et al., 2005; Wu et al., 2020), including training and evaluation routines (using binary cross-entropy), and the option for different kinds of encoder and decoder models. Following Mattos et al. (2025), we use a GCN (Kipf & Welling, 2016) encoder and a dot-product decoder, where the decoder defines the scoring function f.

Inconsistencies. We identify several discrepancies between the implementation and the paper's formal definitions. Notably, the target distribution is computed from node labels rather than subgroup-pair proportions, giving a fundamentally different distribution. Also, demographic parity is computed by conditioning on the sensitive attribute of one node of an edge, rather than on edge membership in $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$.

For the Pokec datasets, the sensitive attributes alternates between `Gender` and `Region` between the paper text, paper tables, and code; we follow the paper text and adopt `Gender` (as described in Table 1). Finally, MORAL's greedy KL-aggregation step (introduced in Section 3.3) differed from the pseudocode (specified in Appendix A.1): it operates on unnormalised group counts rather than probability distributions, which, for instance, breaks scale-independence and corrupts the KL-divergence computation.

Additional components required for reproduction. Two components required for reproduction are missing: the NDKL and a data-splitting procedure. After correspondence with the authors, we received an NDKL implementation. Since it differs slighty from the formal definition in its normalization and divergence computation, we reimplement NDKL exactly as specified in Section 3.2 with all details in Appendix A.2.

For data splits, the paper specifies 70/20/10% train/validation/test proportions but omits the negative sampling scheme; hence, we implement a stratified split protocol matching these proportions. Importantly, all graphs are treated as undirected by removing self-loops, merging duplicate edges, and treating (u, v) as equivalent to (v, u). Positive edges are grouped by subgroup-pair type and randomly assigned to splits; negative edges are sampled independently per split to ensure (i) no negative edge coincides with a positive one, (ii) no duplicates occur, and (iii) subgroup-pair proportions match those of positive edges.

Baselines. To put MORAL in context of prior work, we evaluate three baselines from Mattos et al. (2025): two in-processing methods and one post-processing re-ranker. For in-processing, we include (i) *FairAdj* (Li et al., 2021), which optimises link prediction under dyadic fairness constraints; and (ii) *FairWalk* (Rahman et al., 2019), a fairness-aware embedding method. For post-processing, we include (iii) *DetConstSort* (Geyik et al., 2019), a deterministic reranker with prefix constraints. As baseline hyperparameters are not fully specified, we use released implementations where possible and report configurations in Appendix B.3.

Reproducibility details. Appendix B summarises the hyperparameters used for all experiments and models. To ensure reproducibility, we average over three random seeds that fix all sources of randomness. We document the Python environments and library versions in our GitHub repository.[3] All computations are executed on NVIDIA A100 Tensor Core GPUs of the Dutch national supercomputer, Snellius. The total computational expense for reproduction equals 17.96 GPU hours with an estimated 1.24 $kgCO_2eq$.[4]

3.6 Extensions beyond original work

The previous sections have outlined the foundational work by Mattos et al. (2025). The following subsections provide a detailed explanation of the extensions (first introduced in Section 2) to the foundational work.

3.6.1 Asymmetric homophily stress-test

Homophily is the tendency to connect to others sharing similar characteristics, such as sensitive group membership. In many types of networks, homophily co-occurs with preferential (i.e. degree-based) attachment, reinforcing within-group connectivity (Subramonian et al., 2023). This, in turn, affects the visibility of mi-

[3] Our implementation is available at: `https://github.com/unknown-gitter/reproducing-MORAL`.

[4] Emissions were estimated using CodeCarbon (Lacoste et al., 2023); the estimated carbon intensity of the Netherlands was 411 $g\,CO_2eq\,kWh^{-1}$ in January 2026 (Nowtricity, 2026). Appendix C.2 gives more details on this study's environmental impact.

nority groups in degree-based rankings and recommender systems (Fabbri et al., 2020) and can exacerbate inequalities, such as the glass ceiling effect and invisibility syndrome (Avin et al., 2015; Karimi et al., 2018).

At high homophily, where homophily is quantified as the fraction of within-group edges h_{edge}

$$h_{\text{edge}}(\mathcal{G}) = \frac{1}{|\mathbb{E}|} \sum_{(u,v)\in\mathbb{E}} \mathbf{1}[s_u = s_v]\,, \tag{4}$$

some subgroup-pair candidate pools become sparse, reducing training signal for decoupled predictors. To study the effect of asymmetric homophily on MORAL's fairness and utility, we generate synthetic graphs using the *Directed Preferential Attachment with Homophily* (DPAH) model by Espín-Noboa et al. (2022). DPAH combines preferential attachment and group-conditional homophily, where parameters h_{MM} and h_{mm} control the within-group attachment probability for newly arriving majority and minority nodes, respectively. For implementation, we adapt the DPAH model by Espín-Noboa et al. (2022) to include a binary sensitive attribute $s \in \{0, 1\}$, where $s = 0$ and $s = 1$ denote majority and minority nodes, respectively, following the setup of Figge et al. (2025). Node features are sampled i.i.d. as $\mathbf{x}_u \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, such that features do not encode group membership. Figure 1 shows examples of different homophily levels by varying h_{MM} and h_{mm}.

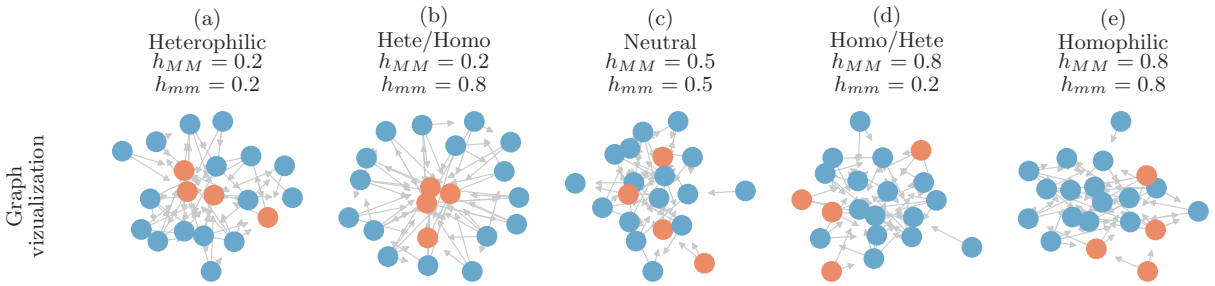

Figure 1: Examples of different homophily levels in graphs. Adapted from Espín-Noboa et al. (2022).

For our experiment, we keep the number of nodes and edge density fixed, while varying h_{MM}, h_{mm}, and minority fraction f_{m}, such that we cover a broad range of edge distributions $\boldsymbol{\pi}$ (see Appendix B.2 for a hyperparameter overview). We expect higher homophily and smaller minority fractions to make subgroup-pair candidate pools sparser, leading to exposure disparities detected by NDKL even when Δ_{DP} remains stable, since Δ_{DP} ignores ranking position (Singh & Joachims, 2018). Across homophily settings, we keep the training and evaluation pipeline fixed and equal to our other experiments, isolating the effect of mixing.[5]

3.6.2 Metric robustness analysis

MORAL's greedy KL-guided aggregation selects edges to minimize divergence to the target distribution $\boldsymbol{\pi}$ at each rank position k of $\mathbf{R}$. NDKL, in turn, evaluates $\mathbf{R}$ by accumulating KL-divergence to the same $\boldsymbol{\pi}$ over prefixes (Eq. 3). Because MORAL's ranking procedure directly optimizes the same divergence that NDKL evaluates post hoc (see Algorithm 1), MORAL's performance under NDKL may be particularly favourable.

To disentangle this coupling and assess the robustness of MORAL's reported fairness-utility trade-offs (Claim **C3**) beyond NDKL, we evaluate fairness with a non-dyadic, rank-aware metric that aggregates exposure differently. In addition, we inspect utility with a position-sensitive metric to complement the Precision@1000.

Fairness metric: AWRF. As described in Section 3.2, Mattos et al. (2025) argue a metric for fairness in link prediction should have **Property 1** (treating subgroup-pair types distinctly) and **Property 2** (rank-awareness). To adhere to these properties while also substituing the aggregation mechanism, we adopt *Attention-Weighted Rank Fairness* (AWRF) as a complementary fairness metric (Sapiezynski et al., 2019; Raj & Ekstrand, 2022). Whereas NDKL accumulates KL-divergence across all ranking pre-fixes, AWRF instead measures exposure in a single, position-weighted distribution over subgroup-pair types. As a result, both satisfy the properties by Mattos et al. (2025), yet differ fundamentally in how exposure is aggregated.

[5]The total computational expense for this extension equals 30.41 GPU hours with an estimated 1.41 $kgCO_2eq$.

To formally adapt Sapiezynski et al. (2019)'s general AWRF definition to our task-specific purposes, we first define normalized position weights

$$\alpha_k = \frac{\delta_k}{\sum_{i=1}^{K} \delta_i}, \tag{5}$$

where δ_k and δ_i follow Eq. 1. Let $g : \mathbb{C} \to \{0\text{-}0,\ 0\text{-}1,\ 1\text{-}1\}$ map each candidate pair to its subgroup-pair type. The attention-weighted exposure distribution induced by $\mathbf{R}$ is then (by notation of Raj & Ekstrand (2022))

$$\epsilon(\mathbf{R})_j = \sum_{k=1}^{K} \alpha_k \, \mathbf{1}[g(R_k) = j] \,, \tag{6}$$

where $\boldsymbol{\epsilon}$ is a probability distribution over subgroup-pair types with $\sum_j \epsilon(\mathbf{R})_j = 1$ and where $\epsilon(\mathbf{R})_j$ denotes component $j \in \{0\text{-}0,\ 0\text{-}1,\ 1\text{-}1\}$ of the distribution. Thus, for every j, $\epsilon(\mathbf{R})_j$ represents the total normalized exposure assigned to subgroup-pair type j. Given the target distribution $\boldsymbol{\pi}$, AWRF is then computed as

$$\mathrm{AWRF}(\mathbf{R}) = \mathrm{dist}(\boldsymbol{\epsilon}(\mathbf{R}),\ \boldsymbol{\pi}) \,, \tag{7}$$

where, as distance function $\mathrm{dist}(\cdot, \cdot)$ we choose the ℓ_1-norm to vary from the NDKL's KL-divergence and to ensure numerical stability, e.g., in cases where subgroup-pair types receive (near-)zero exposure.

In this definition, we adapt AWRF (Sapiezynski et al., 2019) in three ways: (i) we use one-hot subgroup-pair membership, rather than allowing probabilistic group membership; (ii) we fix the position weights α_k using Eq. 1's log-discount, rather than using an attention model; and (iii) we compare exposure to distribution-preserving target distribution $\boldsymbol{\pi}$, rather than using a population estimator (Raj & Ekstrand, 2022).

Utility metric: NDCG@1000. To complement Precision@1000, we additionally report the *Normalized Discounted Cumulative Gain* (NDCG) at $K = 1000$ (Järvelin & Kekäläinen, 2002). Unlike Precision@K, NDCG@K is sensitive to *where* predictions occur. Differences in their scores thus allow us to interpret (to some degree) which part of the ranking is predicted well. Formally, the discounted cumulative gain is

$$\mathrm{DCG@}K = \sum_{k=1}^{K} \delta_k \mathbf{1}[\mathbf{R}_k \in \mathbb{E}_{\text{true}}] \,, \tag{8}$$

where $\mathbb{E}_{\text{true}}$ is the set of true edges and δ_k follows Eq. 1. In turn, $\mathrm{NDCG@}K = \mathrm{DCG@}K/\mathrm{IDCG@}K$, where $\mathrm{IDCG@}K$ is the DCG of an *Ideal* ranking s.t. the range is normalized to the unit interval (Manning, 2008).

3.6.3 Categorical sensitive attributes

In Section 3.1, we described how *binary* sensitive attributes $s \in \{0, 1\}$ partition candidate node pairs into subgroup-pair types. While Mattos et al. (2025) state this framework can be generalized in theory, they do not test this. Thus, to study their claim in practice, we extend MORAL to *categorical* sensitive attributes $s \in \mathbb{S} = \{0, 1, \ldots, m-1\}$ with cardinality $|\mathbb{S}| = m > 2$. We test the updated implementation up to $m = 10$.

Formally, we define the subgroup-pair type $\mathbb{E}_{a\text{-}b} = \{(u, v) \in \mathbb{C} \mid \{s_u, s_v\} = \{a, b\}\}$, where $0 \leq a \leq b \leq m-1$, yielding $T(m) = \frac{m(m+1)}{2}$ subgroup-pair types (the unique entries of a symmetric $m \times m$ matrix). We further define the intra- and inter-sets as $\mathbb{E}_{\text{intra}} = \{(u, v) \in \mathbb{C} \mid s_u = s_v\}$ and $\mathbb{E}_{\text{inter}} = \{(u, v) \in \mathbb{C} \mid s_u \neq s_v\}$. In the binary case, within-group disparites can only arise in the intra-group (since for $m = 2$, $|\mathbb{E}_{\text{intra}}| = 2$ and $|\mathbb{E}_{\text{inter}}| = 1$). In contrast, in the $m > 2$ categorical setting, these disparities can arise in both $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$ (since for, e.g., $m = 3$, $|\mathbb{E}_{\text{intra}}| = |\mathbb{E}_{\text{inter}}| = 3$). More generally, with increasing m, aggregation collapses an increasing number of subgroup-pair interactions into $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$, motivating a granular evaluation across all T types with MORAL and NDKL, rather than at the coarse, aggregate level with Δ_{DP}.

Experimental setup. We test the categorical extension on the Credit dataset by binning the `Age` attribute[6] into m discrete categories. To accomodate any m algorithmically, we modify MORAL to handle the T induced subgroup-pair types dynamically: one decoupled encoder-decoder model is trained per type, and greedy KL-aggregation yields a T-dimensional ranking $\mathbf{R}$. For evaluation, both original and new metrics are adapted; Δ_{DP} (Eq. 2) is generalised using a One-vs-All approach, inspired by Figge et al. (2025).

[6]Here, we use the non-preprocessed version of the Credit dataset by Lichman (2013), downloadable from Hugging Face. We construct the adjacency matrix by connecting each node to its 10 nearest neighbours in the standardised feature space.

4 Results

This section begins by presenting findings on the reproducibility of Claims **C1** through **C3**, introduced in Section 2. Following, extensions that go beyond the original paper are discussed.

4.1 Reproducibility of core claims

Claim C1 & C2. Claim **C1** states that Δ_{DP} hides within-group exposure bias while **C2** states that NDKL explicitly treats all subgroup-pairings and detects exposure disparities that Δ_{DP} fails to detect.

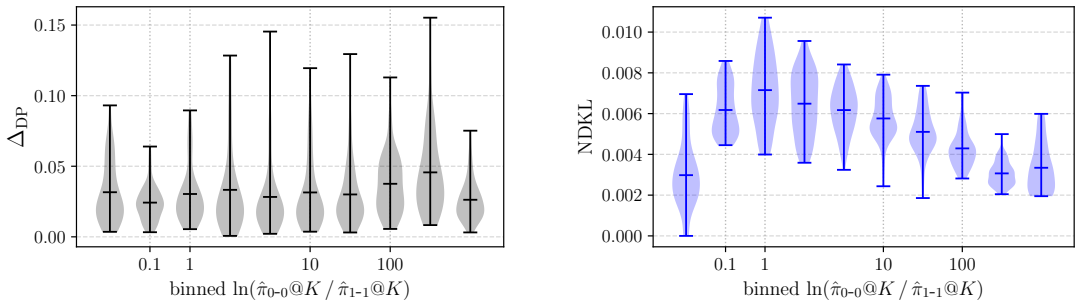

Figure 2: **Left:** permutation-invariant Δ_{DP} (Eq. 2). **Right:** rank-aware NDKL (Eq. 3). Both metrics are plotted versus binned log-ratio $\ln(\hat{\pi}_{0\text{-}0}/\hat{\pi}_{1\text{-}1})$ at K, where $\hat{\pi}_{0\text{-}0}$ and $\hat{\pi}_{1\text{-}1}$ denote the exposure mass assigned to $\mathbb{E}_{0\text{-}0}$ and $\mathbb{E}_{1\text{-}1}$, resp., showing Δ_{DP}'s invariance and NDKL's sensitivity to shifts within $\mathbb{E}_{\text{intra}} = \mathbb{E}_{0\text{-}0} \cup \mathbb{E}_{1\text{-}1}$.

Figure 2 depicts Δ_{DP} and NDKL across a range of experimental configurations with binary sensitive attributes, such that exposure disparities can only arise within $\mathbb{E}_{\text{intra}}$. The configurations are grouped into bins by the log-ratio $\hat{\pi}_{0\text{-}0}/\hat{\pi}_{1\text{-}1}$, thus reflecting variation in exposure allocation within $\mathbb{E}_{\text{intra}} = \mathbb{E}_{0\text{-}0} \cup \mathbb{E}_{1\text{-}1}$. The ratio measures how exposure is redistributed between two edge types while their combined exposure is fixed. While Δ_{DP} remains largely invariant across bins, NDKL varies substantially, revealing its sensitivity to exposure shifts. Since only the relative allocation $\hat{\pi}_{0\text{-}0}/\hat{\pi}_{1\text{-}1}$ is changed, NDKL is maximized where this ratio is, on average, most misaligned with the baseline ratio $\pi_{0\text{-}0}/\pi_{1\text{-}1}$ implied by $\boldsymbol{\pi}$. This seems to occur around $\hat{\pi}_{0\text{-}0}/\hat{\pi}_{1\text{-}1} \approx 1$ in Figure 2. Consistent with **C1** and **C2**, the contrast between the two empirically confirms that Δ_{DP} obscures within-group exposure disparities (**C1**), whereas NDKL does capture them (**C2**).

Claim C3. Claim **C3** states that MORAL consistently mitigates previously undetected exposure biases, as measured by NDKL, while keeping its competitive utility relative to baselines, as measured by Precision.

Tables 2 and 3 report NDKL and Precision at $K = 1000$ on six datasets for MORAL and the baselines (all introduced in Section 3.5). For each method, we compare the reproduced scores with the originally reported results of Mattos et al. (2025) (listed in completion in Appendix F). Relative to the original scores, the NDKL is reproduced or improved, whereas Precision improves in most cases with some exceptions, e.g., for FairWalk. The small standard deviations tell us these fairness-utility patterns are consistent across runs. We also observe anomalies in the original results, e.g., sub-random Precision scores for DetConstSort, and that FairAdj runs out of memory on three datasets, whereas MORAL remains tractable.

Relative to the reproduced scores of the in-processing baselines (FairAdj, FairWalk) and the post-processing baseline (DetConstSort), the bold entries show that MORAL consistently achieves (tied-)lowest, (near-)zero NDKL. At the same time, the reduction in NDKL/exposure bias comes with a minimal trade-off in utility: MORAL achieves the (near-)highest Precision on all datasets. The trade-off in utility is most visible on the NBA dataset, whose smaller scale (both in number of nodes $|\mathbb{V}|$ and number of edges $|\mathbb{E}|$, see Appendix C.1) makes top-K precision more sensitive to exposure-preserving re-rankings because cut-off K spans a larger fraction of the candidate pools. MORAL and DetConstSort are indistinguishable in Precision to two decimals, consistent with both methods post-processing/re-ranking the same base scores and largely preserving the top-K true positives while re-distributing exposure. Overall, MORAL reduces exposure disparities while largely preserving top-K true positives, supporting Claim **C3**.

Table 2: NDKL@1000 across reproduced baselines and MORAL. ΔNDKL = ours − original. Best in **bold**. Consistent with Mattos et al. (2025), we round to two decimals. OOM is an Out-Of-Memory error.

Model	Metric	Credit	Facebook	German	NBA	Pokec-n	Pokec-z
FairAdj	$\text{NDKL}_{\text{ours}} \downarrow$	OOM	0.02 ± 0.00	$\mathbf{0.01 \pm 0.00}$	0.04 ± 0.02	OOM	OOM
	$\Delta\text{NDKL} \downarrow$	OOM	(−0.08)	(−0.09)	(−0.07)	OOM	OOM
FairWalk	$\text{NDKL}_{\text{ours}} \downarrow$	0.02 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01
	$\Delta\text{NDKL} \downarrow$	(−0.04)	(−0.04)	(−0.08)	(−0.03)	(−0.04)	(−0.04)
DetConstSort	$\text{NDKL}_{\text{ours}} \downarrow$	0.02 ± 0.01	0.02 ± 0.00	0.02 ± 0.00	$\mathbf{0.01 \pm 0.00}$	0.02 ± 0.00	0.03 ± 0.01
	$\Delta\text{NDKL} \downarrow$	(−0.04)	(−0.13)	(−0.02)	(−0.08)	(−0.05)	(−0.20)
MORAL	$\text{NDKL}_{\text{ours}} \downarrow$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$
	$\Delta\text{NDKL} \downarrow$	(−0.04)	(0.00)	(−0.02)	(−0.01)	(−0.02)	(−0.03)

Table 3: Precision@1000 across reproduced baselines and MORAL. ΔPrecision = ours − original. Best in **bold**. Consistent with Mattos et al. (2025), we round to two decimals. OOM is an Out-Of-Memory error.

Model	Metric	Credit	Facebook	German	NBA	Pokec-n	Pokec-z
FairAdj	$\text{Precision}_{\text{ours}} \uparrow$	OOM	0.94 ± 0.01	$\mathbf{0.99 \pm 0.00}$	0.52 ± 0.03	OOM	OOM
	$\Delta\text{Precision} \uparrow$	OOM	(+0.52)	(+0.45)	(+0.02)	OOM	OOM
FairWalk	$\text{Precision}_{\text{ours}} \uparrow$	$\mathbf{1.00 \pm 0.00}$	0.92 ± 0.01	0.85 ± 0.01	0.53 ± 0.01	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
	$\Delta\text{Precision} \uparrow$	(0.00)	(−0.04)	(−0.09)	(−0.02)	(0.00)	(0.00)
DetConstSort	$\text{Precision}_{\text{ours}} \uparrow$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$	0.95 ± 0.00	$\mathbf{0.75 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
	$\Delta\text{Precision} \uparrow$	(+1.00)	(+0.97)	(+0.40)	(+0.54)	(+0.93)	(+0.99)
MORAL	$\text{Precision}_{\text{ours}} \uparrow$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$	0.95 ± 0.01	$\mathbf{0.75 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
	$\Delta\text{Precision} \uparrow$	(0.00)	(+0.02)	(−0.01)	(−0.05)	(+0.02)	(+0.02)

4.2 Results beyond original paper

We conducted three additional analyses to further evaluate the paper's claims: an asymmetric homophily stress-test, a metric robustness analysis, and testing categorical sensitive attributes.

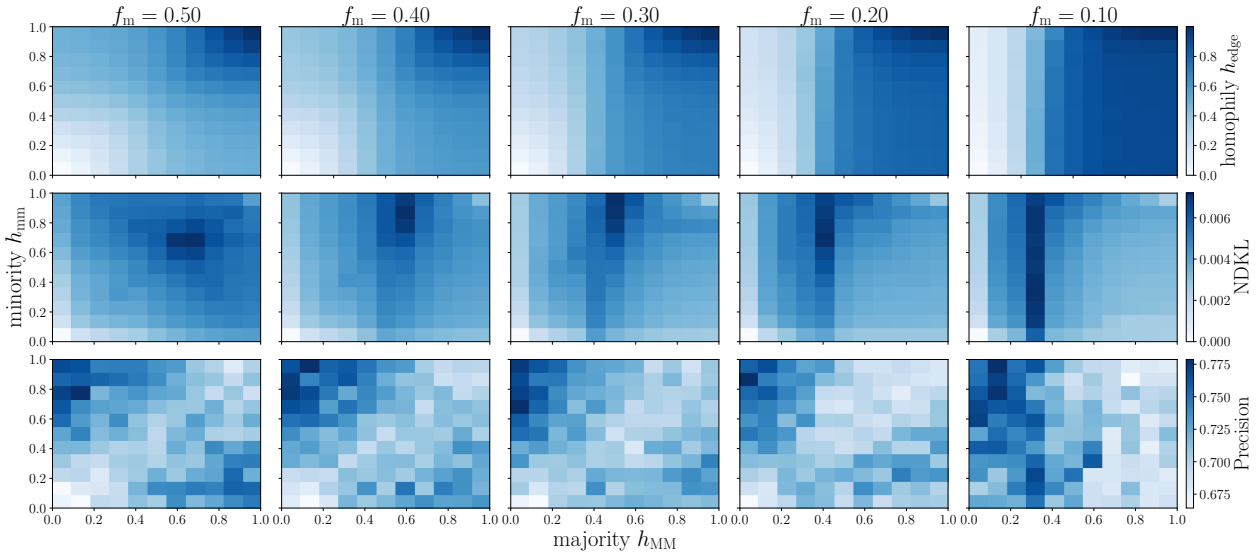

Figure 3: Homophily h_{edge}, NDKL, and Precision on heatmaps of h_{mm} and h_{mm} over five f_{m} values.

Asymmetric homophily stress-test. Figure 3 shows a heatmap for h_{edge}, NDKL, and Precision as function of h_{mm}, h_{MM}, and f_{m}. We first observe that h_{edge} increases with h_{mm} and h_{MM}; as f_{m} decreases, the increase becomes more dominated by h_{MM}. The patterns in the NDKL and Precision show some

consistency over the range of f_{m} values too. For instance, NDKL maxima shift with f_{m}, as changes in the availability of subgroup-pair types change the $(h_{\mathrm{MM}}, f_{\mathrm{m}})$-configuration that maximally misaligns exposure.

To further illuminate these patterns, Figure 4 shows z-scored NDKL, AWRF, and Precision, mean-averaged over minority homophily h_{mm} for each $(h_{\mathrm{MM}}, f_{\mathrm{m}})$-configuration; in other words, it projects the Figure 3 heatmaps onto the h_{MM} axis. Across f_{m} values, NDKL and AWRF peaks coincide with $h_{\mathrm{edge}} \approx 0.6$, even though the peak's location in h_{MM} shifts as f_{m} varies. This medium-homophily setting yields the largest exposure divergence since within-group edges are frequent and rankings are diverse. At higher homophily, the NDKL and AWRF decrease again: subgroup-pair candidate pools become sparser, reducing the degrees of freedom for exposure redistribution and thereby for divergence. For a different reason, Precision tends to drop with increasing h_{edge}: in high homophily, MORAL's decoupled predictors receive less training signal. Finally, Appendix D distinguishes absolute homophily h_{edge} from relative homophily Δh, showing that exposure-based fairness metrics respond primarily to absolute homophily. Together, these results show that MORAL's fairness behaviour and performance respond predictably to the graph's homophily structure.

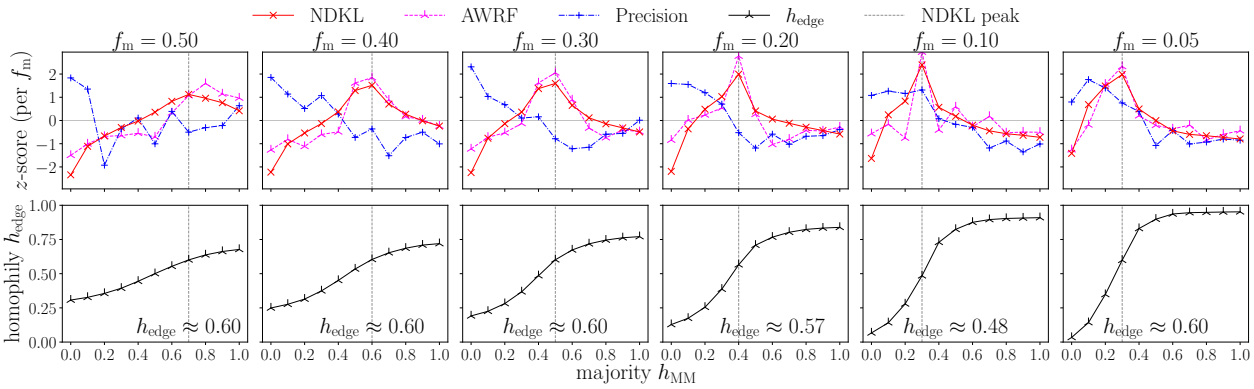

Figure 4: **Top:** z-scored NDKL, AWRF, and Precision versus majority homophily h_{MM} for decreasing minority fraction f_{m}, mean-averaged across minority homophily h_{mm}. **Bottom:** corresponding edge homophily h_{edge}. Vertical dotted lines indicate NDKL peaks, with the corresponding h_{edge} value annotated.

Table 4: AWRF↓ and NDCG↑ at $K = 1000$ across reproduced baselines and MORAL (see Tables 2 and 3 for NDKL@1000 and Precision@1000). Best values are underlined. OOM indicates an Out-Of-Memory error.

Metric	Model	Credit	Facebook	German	NBA	Pokec-n	Pokec-z
AWRF	FairAdj	OOM	0.0421 ± 0.0139	0.0159 ± 0.0015	0.0808 ± 0.0223	OOM	OOM
	FairWalk	0.0111 ± 0.0060	0.0288 ± 0.0113	0.0518 ± 0.0223	0.0837 ± 0.0456	0.0619 ± 0.0234	0.0353 ± 0.0148
	DetConstSort	0.0371 ± 0.0083	0.0111 ± 0.0037	0.0144 ± 0.0046	0.0222 ± 0.0052	0.1225 ± 0.0027	0.1160 ± 0.0039
	MORAL	0.0021 ± 0.0000	0.0032 ± 0.0000	0.0021 ± 0.0000	0.0020 ± 0.0000	0.0023 ± 0.0000	0.0012 ± 0.0000
NDCG	FairAdj	OOM	0.9934 ± 0.0025	0.9993 ± 0.0003	0.8799 ± 0.0362	OOM	OOM
	FairWalk	0.9999 ± 0.0001	0.9907 ± 0.0021	0.9810 ± 0.0003	0.8752 ± 0.0135	0.9999 ± 0.0001	0.9995 ± 0.0004
	DetConstSort	1.0000 ± 0.0000	0.9951 ± 0.0004	0.9942 ± 0.0011	0.9608 ± 0.0025	0.9995 ± 0.0006	0.9995 ± 0.0003
	MORAL	1.0000 ± 0.0000	0.9960 ± 0.0004	0.9942 ± 0.0012	0.9598 ± 0.0035	0.9995 ± 0.0006	0.9995 ± 0.0003

Metric robustness analysis. Table 4 reports the alternative metrics AWRF and NDCG for MORAL and the reproduced baselines. Tables 2 and 3 show the NDKL and Precision scores, respectively. Both NDKL and AWRF remain low across datasets, with the NDKL staying in a 0.00–0.01 range and the AWRF in a 0.001–0.003 range, indicating consistently reduced exposure bias. Figure 4 further shows their coupled reactivity to homophily h_{edge}. Their coupling is expected since both measure ranking-induced exposure relative to $\boldsymbol{\pi}$, though NDKL aggregates position-weighted prefix KL-divergences (Eq. 3) while AWRF applies a single ℓ_1-norm to the top-K exposure distribution (Eq. 7). Although MORAL directly optimizes the divergence underlying NDKL, the aligned behaviour of AWRF suggests that MORAL is not merely "gaming" the metric. In fact, the separation between MORAL and baselines is larger under AWRF. Regarding utility, the gap between NDCG and Precision on the Facebook, German, and NBA dataset suggests that MORAL preserves ranking quality at higher ranks, which is critical for downstream applications. The gap is largest on NBA since its smaller scale leaves more borderline errors around rank K, thus increasing sensitivity to re-rankings.

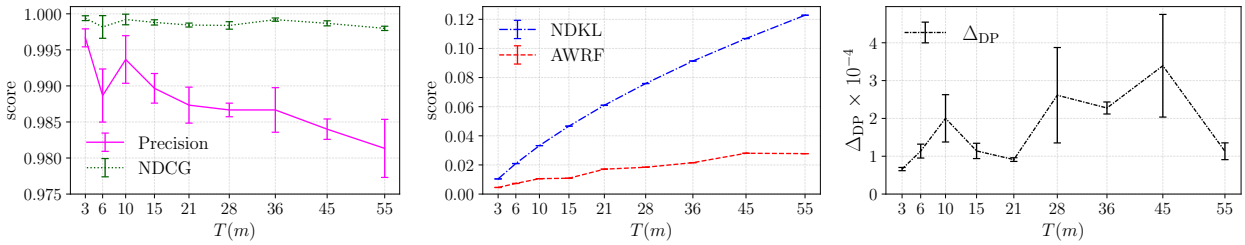

Figure 5: Metrics (mean ± standard deviation) vs. the number of subgroup-pair types $T(m) = \frac{m(m+1)}{2}$ on the Credit dataset. **Left:** Utility (Precision & NDCG). **Middle:** Fairness (NDKL & AWRF). **Right:** Δ_{DP}.

Categorical sensitive attributes. Figure 5 shows MORAL's performance on the Credit dataset up to cardinality $m = 10$, equivalent to $T(10) = 55$ subgroup-pair types. Δ_{DP} remains on the order of 10^{-4} with fluctuations but no systematic dependence on T, as it compares aggregated exposure mass and is invariant to its distribution within $\mathbb{E}_{\text{intra}}$ and $\mathbb{E}_{\text{inter}}$. In contrast, NDKL and AWRF degrade monotonically, as increasing T enlarges the set of subgroup-pair types over which exposure must be balanced, making simultaneous fairness over all types progressively more difficult. NDKL increases more steeply than AWRF because it aggregates KL-divergences for every kth prefix, turning stepwise deviations of $\hat{\boldsymbol{\pi}}_k$ from $\boldsymbol{\pi}$ into an accumulated penalty that grows with T. AWRF instead measures a single ℓ_1-norm at K (i.e., no prefix-by-prefix accumulation) (with the ℓ_1-norm additionally operating on a different scale than KL-divergence (Cover & Thomas, 2006)).

Utility decreases slightly as T increases but remains high (in the range of 0.97–1.00). The widening gap between Precision and NDCG suggests that top-ranked edges remain accurate while lower-ranked positions become less stable with higher T, consistent with the earlier observations on the NBA dataset results. Overall, MORAL maintains utility well at higher cardinality, but the observed NDKL and AWRF degradation indicates that aligning exposure with $\boldsymbol{\pi}$ becomes structurally more challenging as T increases.

5 Discussion

The primary aim of this study was to reproduce and evaluate the three main claims by Mattos et al. (2025), which we introduced in Section 2. In Section 4.1, we found our results support Claims **C1**–**C3**. Across all datasets, we reproduced or even improved upon the originally reported fairness-utility trade-offs. We attribute the discrepancies between the original results and ours primarily to underspecification in the released codebase; in cases of ambiguity, we prioritized optimization, leading to improved fairness and utility. Compared to in- and post-processing baselines, MORAL performed best, suggesting that enforcing distribution-preserving exposure at ranking time is more effective than training-time interventions or heuristic post-processing constraints. Additionally, we confirmed generalizability to additional datasets in Appendix E. Beyond reproducing the main results, our extensions assessed the generality and limits of the framework.

Firstly, the asymmetric homophily stress-test clarified when MORAL's trade-offs are most evident. In Figures 3 and 4, we observe that exposure-based metrics are most strained at intermediate *absolute* homophily (rather than *relative* homophily, see Appendix D): high cross-group connectivity forces allocation of limited exposure over many competing subgroup-pair pools, making disparities easier to surface. At higher homophily, subgroup-pair candidate pools become structurally constrained and the learning signal for some decoupled predictors weakens, consistent with declining utility. Taken together, MORAL performs strongly and consistently across the tested homophily settings, with homophily inducing only mild variation around already low exposure disparity and high utility, supporting Claim **C3**. Appendix D further shows that Δ_{DP} is largely insensitive to homophily, whereas NDKL varies substantially, thus reinforcing Claims **C1** & **C2**.

Secondly, we assessed whether our conclusions are metric-dependent by complementing NDKL and Precision with AWRF and NDCG. The larger separation between MORAL and the baselines under AWRF suggests MORAL's reductions in exposure bias are not NDKL-specific nor a by-product of optimizing a closely-aligned objective, but that it reduces exposure disparity more generally. We note, however, that this robustness is demonstrated within exposure-based fairness metrics and does not extend to fundamentally different fairness definitions. The NDCG results suggest that MORAL maintains ranking quality most near the top of the ranking. This is important for downstream applications such as social networks, where small shifts at the

top of the ranking can result in large attention and visibility differences (Singh & Joachims, 2018; Ferrara et al., 2022). Overall, the results suggest limited metric sensitivity, thus strengthening Claim **C2** & **C3**.

Thirdly, the categorical-attribute extension assessed MORAL in higher-cardinality settings. For fairness, Δ_{DP} remains largely insensitive to growing T, whereas the exposure-based metrics degrade monotonically, supporting Claim **C1** & **C2**. This also suggests finer-grained structure of subgroups makes distribution-preserving exposure alignment increasingly difficult, which, in turn, highlights a structural tension between subgroup granularity and exposure redistribution. Since part of NDKL's degradation may stem from its cumulative prefix-wise aggregation mechanism, subgroup-pair-adapted AWRF provides a complementary single-shot, position-weighted perspective on exposure divergence. Importantly, utility slightly decreases but remains high overall, with the decrease partly due to weaker training signal, similar to what we observed in high homophily. The gap between Precision and NDCG widening with T indicates that errors concentrate lower within the top-K as T grows. Nonetheless, the degradations points to a scalability limitation: as T grows (and subgroup-pair pools become smaller), meeting a distribution-preserving target $\boldsymbol{\pi}$ becomes structurally harder, while also increasing computational cost due to a separate predictor per type. All in all, the results support Claim **C3**, while also highlighting a practical scalability boundary.

In conclusion, dyadic demographic parity is insufficient for ranked link prediction because it can obscure subgroup-pair exposure disparities revealed by NDKL, suggesting that prior work relying on dyadic parity may under-report such effects. More broadly, fairness criteria should match downstream use: when outputs are consumed as rankings, exposure should be the relevant currency of fairness. Despite inconsistencies in the original implementation, MORAL behaves as intended and remains robust across our stress-tests and extensions, supporting its use as a simple post-processing method for debiasing ranked link predictions.

Practical recommendations. We make a few practical recommendations for practitioners: (i) in ranked link prediction, demographic parity is a poor proxy for exposure fairness; prioritize exposure-based metrics; (ii) both NDKL and AWRF are suitable at evaluating for exposure bias; (iii) report NDCG alongside Precision to localize where utility degrades; (iv) use MORAL when outputs are consumed as rankings and top-K exposure is a fairness concern, especially when per-type predictors are feasible; (v) utility trade-offs are largest on smaller graphs, where relative NDCG gains are also strongest, suggesting MORAL is especially competent at preserving top-of-ranking quality; (vi) auditing for exposure bias is most informative at intermediate *absolute* homophily; (vii) beyond binary attributes, MORAL scales to a more realistic, higher-cardinality setting, but granularity should be treated as a fairness-scalability trade-off: exposure alignment becomes harder and more computationally costly, while top-K utility degrades only modestly.

Limitations and future work. A natural direction for future work is to study how compute scales versus fairness and utility. In particular, it remains unclear whether training separate predictors per subgroup-pair type is necessary to achieve strong fairness-utility trade-offs, or whether parameter sharing (e.g., hierarchical grouping or a shared backbone) can reduce combinatorial growth without sacrificing performance. Second, we evaluate rankings at a fixed cut-off; future work should try a range of Ks across settings to better localise where in the ranking MORAL helps most. Third, our adapted AWRF uses an ℓ_1-distance; testing KL-divergence-based variants could further isolate the usability of both NDKL and AWRF. Finally, as mentioned in Section 3.3, preserving $\boldsymbol{\pi}$ is a design choice rather than a universally correct fairness target (Mitchell et al., 2021); future work should compare fairness objectives that reflect different normative commitments.

What was easy. As touched on in Section 3.5, the codebase included dataset loaders and a training pipeline, which was helpful for getting up to speed. Moreover, after contacting the authors for clarification on the NDKL, they shared a reference implementation and offered helpful input on our theoretical questions.

What was difficult. Reproducing the experiments proved more challenging than anticipated: several elements were inconsistent, underspecified, or missing. As many components were dynamic and involved randomness, it was difficult to verify intermediate results. We hope our fixes and extensions make it easier for others to build on this work and support continued progress on fairness in link prediction.

References

Himan Abdollahpouri. Popularity bias in ranking and recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 529–530, 2019.

Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 conference on innovations in theoretical computer science*, pp. 41–50, 2015.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social networks*, 21(4): 375–395, 2000.

April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–23, 2024.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

Sergio Currarini, Matthew O Jackson, and Paolo Pin. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, 107(11): 4857–4861, 2010.

Sean Current, Yuntian He, Saket Gurukar, and Srinivasan Parthasarathy. 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*, pp. 1–14, 2022.

Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6):1011–1061, 2024.

Issa Moussa Diop, Cherif Diallo, Chantal Cherifi, and Hocine Cherifi. Exploring the regional structure of the worldwide air traffic and route networks. *Complexity*, 2023(1):3411799, 2023.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Lisette Espín-Noboa, Claudia Wagner, Markus Strohmaier, and Fariba Karimi. Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific reports*, 12(1):2012, 2022.

Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. The effect of homophily on disparate visibility of minorities in people recommender systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pp. 165–175, 2020.

Antonio Ferrara, Lisette Espín-Noboa, Fariba Karimi, and Claudia Wagner. Link recommendations: Their impact on network structure and minorities. In *Proceedings of the 14th ACM Web Science Conference 2022*, pp. 228–238, 2022.

Ruben Figge, Sjoerd Gunneweg, Aaron Kuin, and Mees Lindeman. Reassessing fairness: A reproducibility study of nifa's impact on gnn models. *Transactions on Machine Learning Research*, 2025.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *Science advances*, 7(12):eabc9800, 2021.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pp. 2221–2231, 2019.

Avijit Ghosh, Ritam Dutt, and Christo Wilson. When fair ranking meets uncertain inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 1033–1043. ACM, July 2021. doi: 10.1145/3404835.3462850. URL `http://dx.doi.org/10.1145/3404835.3462850`.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks.*, volume 2, pp. 729–734. IEEE, 2005.

Roger Guimera and Luis A Nunes Amaral. Modeling the world-wide airport network. *The European Physical Journal B*, 38(2):381–385, 2004.

Shantanu Gupta, Hao Wang, Zachary Lipton, and Yuyang Wang. Correcting exposure bias for link recommendation. In *International Conference on Machine Learning*, pp. 3953–3963. PMLR, 2021.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8(1):11077, 2018.

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

TN Kipf and M Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, 2016.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Codecarbon, 2023. URL `https://github.com/mlco2/codecarbon`.

Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, and Hongfu Liu. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International conference on learning representations*, 2021.

Yanying Li, Xiuling Wang, Yue Ning, and Hui Wang. Fairlp: Towards fair link prediction on social network graphs. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pp. 628–639, 2022.

M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013. Irvine, CA: University of California, School of Information and Computer Science.

Yuankai Luo, Lei Shi, and Xiao-Ming Wu. Classic gnns are strong baselines: Reassessing gnns for node classification. *Advances in Neural Information Processing Systems*, 37:97650–97669, 2024.

Zihan Luo, Hong Huang, Jianxun Lian, Xiran Song, Xing Xie, and Hai Jin. Cross-links matter for link prediction: rethinking the debiased gnn from a data perspective. *Advances in neural information processing systems*, 36:79594–79612, 2023.

Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing, 2008.

João Mattos, Debolina Halder Lina, and Arlei Silva. Breaking the dyadic barrier: Rethinking fairness in link prediction beyond demographic parity. *arXiv preprint arXiv:2511.06568*, 2025.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021.

Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

Nowtricity. CO_2 emissions per kwh in the netherlands, January 2026.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*, 2020.

Tiago de Paula Peixoto. Openflights airport network (global airport routes network). `https://networks.skewed.de/net/openflights`, 2020. A network of regularly occurring flights among airports worldwide, extracted from the Openflights.org dataset; nodes represent airports and directed edges represent regularly scheduled commercial flights.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 36:523–548, 2023a.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023b.

Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: Towards fair graph embedding. 2019.

Amifa Raj and Michael D Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726–736, 2022.

Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. Quantifying the impact of user attention on fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference*, pp. 553–562, 2019.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2219–2228, 2018.

Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pp. 923–932, 2018.

Arjun Subramonian, Levent Sagun, and Yizhou Sun. Networked inequality: Preferential attachment bias in graph neural network link prediction. *arXiv preprint arXiv:2309.17417*, 2023.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

He Zhang, Bang Wu, Xingliang Yuan, Shirui Pan, Hanghang Tong, and Jian Pei. Trustworthy graph neural networks: Aspects, methods, and trends. *Proceedings of the IEEE*, 112(2):97–139, 2024.

A MORAL and NDKL

This Appendix provides the MORAL pseudocode and a brief explanation on our NDKL implementation.

A.1 MORAL pseudocode

Algorithm 1 outlines the MORAL algorithm pseudocode, adapted from Mattos et al. (2025) but adjusted to our notation. In Section 3.6.2, we touch on the algorithm's similarities to the NDKL metric (Eq. 3).

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

Input:

- Candidate ordered sets $\mathbf{C}_j = ((u, v, \text{score}), (u_2, v_2, \text{score}_2), \dots)$ 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 $\mathbf{C}_j$ is not empty **do**

    Let $(u, v, \text{score}) \leftarrow$ top element in $\mathbf{C}_j$

    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: $D_{\mathrm{KL}}(\boldsymbol{q}' \,\|\, \boldsymbol{\pi})$, where $\boldsymbol{q}' = (q'_0, q'_1, q'_2)$

    if $D_{\mathrm{KL}}(\boldsymbol{q}' \,\|\, \boldsymbol{\pi}) <$ min_kl **then**

      Update min_kl $\leftarrow D_{\mathrm{KL}}(\boldsymbol{q}' \,\|\, \boldsymbol{\pi})$, selected_group $\leftarrow j$, selected_edge $\leftarrow (u, v)$

    end

  end

  Append (selected_edge, selected_group) to $\mathbf{R}$

  Remove top element from $\mathbf{C}_{\text{selected_group}}$

  Update $c_{\text{selected_group}} \leftarrow c_{\text{selected_group}} + 1$

end

return $\mathbf{R}$

A.2 NDKL implementation

Another part that required further specification and optimization for reproduction is the NDKL. After correspondence with the authors, they were kind to send their NDKL implementation. We found their implementation could be optimized further and adapted to better reflect their mathematical definition:

- Firstly, KL-divergence is not symmetric w.r.t. the distributions it evaluates: rearranging the order of distributions — for example from $D_{\mathrm{KL}}(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}_k)$ to $D_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi})$, where $\hat{\boldsymbol{\pi}}_k$ is the empirical edge group distribution up to rank k and $\boldsymbol{\pi}$ the global target distribution — changes the resulting NDKL value. We chose to follow the definition defined in the paper text: $D_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi})$.
- Secondly, we removed an extra normalisation term used in their NDKL function, which was not mentioned in the paper text.
- Thirdly, for numerical stability, we smooth the target distribution $\boldsymbol{\pi}$ with clamping and renormalisation, i.e. $\boldsymbol{\pi} = \text{clamp}(\boldsymbol{\pi}, \eta)$, with $\eta = 1 \times 10^{-12}$, followed by $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi} / \sum_j \pi_j$, where j indexes subgroup-pair types. This ensures $\pi_j \geq \eta$ for all j and prevents $D_{\mathrm{KL}}(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi})$ from diverging when $\pi_j \approx 0$ with $\hat{\pi}_{k,j} > 0$.

Our implementation can be found in the following file: `https://github.com/unknown-gitter/reproducing-MORAL/blob/main/scripts/helpers/metrics.py`.

B Hyperparameters

This section first details the hyperparameters used for reproduction, followed by the hyperparameters used for the homophily experiments.

B.1 Hyperparameters for reproduction

The datasets were trained on a graph neural network (Wu et al., 2020) with a GCN (Kipf & Welling, 2016) encoder and a dot-product decoder. Table 5 lists the hyperparameters used for training the model.

Table 5: Model architecture and hyperparameter specifications for the MORAL framework.

General setup		**Encoder & Decoder**		**Training details**	
Hyperparameter	Value	Component	Value	Parameter	Value
Datasets	See Table 1	Encoder	GCN	Optimizer	Adam
Seeds	0, 1, 2	Hidden dim.	128	Learning rate	3×10^{-4}
Sensitive groups	3 (00, 01, 11)	Encoding layers	2	Weight decay	0.0
K	1000	Decoder	Dot product	Batch size	1024
Feature normalisation	Min-max	Activation	ReLU	Max epochs	500
		Dropout	0.2		

B.2 Hyperparameters for homophily experiments

To generate the graphs with differing homophilies, the *Directed Preferential Attachment with Homophily* (DPAH) graph generator by Espín-Noboa et al. (2022) was employed, from whom the hyperparameters in Table 6 were adopted. See Section 3.6.1 for a detailed explanation on the utilisation.

Table 6: Hyperparameter specifications for the DPAH synthetic graph generation.

Global parameters		**Demographic config.**		**Homophily**	
Hyperparameter	Value	Parameter	Range / Value	Parameter	Range / Value
Nodes N	1000	Minority fraction f_m	[0.05, 0.10, 0.20, 0.30, 0.40, 0.50]	Homophily h_{MM}	[0.0, 1.0] in steps of 0.1
Edge density d	0.006	Majority label	0 (M)	Homophily h_{mm}	[0.0, 1.0] in steps of 0.1
Feature dim.	16	Minority label	1 (m)	PL-Outdegree γ_M	2.5
Generation seed	42	Groups (G)	{M, m}	PL-Outdegree γ_m	2.5
Graph type	Directed			Preferential attachment	In-degree

B.3 Baseline experiments

To compare the results, a few baseline models were implemented. See Section 4.1 for the baseline results.

FairAdj. FairAdj is an algorithm proposed by Li et al. (2021), the same paper to which the contributions of Mattos et al. (2025) are a response. It empirically learns a fair adjacency matrix by updating the normalized adjacency matrix while maintaining the graph structure. Li et al. (2021)'s implementation uses a Variational Graph Autoencoder (VGAE) with a two-layer Graph Convolutional Network (GCN) as the inference model. The training procedure consists of two alternating phases: T_1 epochs for optimizing link prediction and T_2 epochs for dyadic fairness. They use projected gradient descent to regulate edge weights. We evaluate with the utility and fairness metrics on the top-k ranking list without applying a post-processing method. We adapted the FairAdj codebase for our implementation and refer to our GitHub repo for further details and references to the original model. We adopted their hyperparameters as seen in Table 7.

Table 7: Hyperparameters and architectural configurations for the FairAdj framework (Li et al., 2021)

Model architecture		**Optimization (Phase 1: T_1)**		**Fairness (Phase 1: T_2)**	
Hyperparameter	Value	Parameter	Value	Parameter	Value
Base model	GCN-VAE	Optimizer	AdaM	Fairness metric	Δ_{DP}
1st hidden layer	32	Learning rate (η_θ)	0.01	Learning rate ($\eta_{\tilde{A}}$)	Dataset-specific
2nd hidden layer	16	Outer epochs	4	T_2 sub-epochs	Dataset-specific
Activation func.	ReLU	T_1 sub-epochs	50	PGD projection	Simplex/right stochastic
Decoder	Dot product	Weight clipping	1.0	Structural constraint	Fixed sparsity
Dropout	0.0	KL-divergence	Gaussian prior	Gradient descent	Projected

FairWalk. FairWalk is a fairness-aware graph embedding algorithm proposed by Rahman et al. (2019). It is an in-processing method, because it incorporates fairness directly into the embedding learning procedure by adjusting the random walk generation. FairWalk is an extension of node2vec (Mikolov et al., 2013). First, neighbours are grouped by sensitive attribute at each step, after which a group is sampled uniformly and a random neighbour is chosen from that group. Then, the standard skip-gram (word2vec) objective learns node embeddings based on the resulting walks. We evaluate on the outputted top-K edge ranking list. Our implementation is inspired by the authors' official GitHub repository and we adopt the node2vec-style hyperparameters reported in their paper, as shown in Table 8. We refer to our codebase for more details.

Table 8: Hyperparameters and configuration for FairWalk.

Walk generation		**Skip-gram training**		**Experimental setup**	
Parameter	Value	Parameter	Value	Parameter	Value
Embedding dimension (d)	128	Skip-gram (sg)	1	Runs	3
Walk length (L)	80	Hierarchical softmax (hs)	0	Random seed	$\{0, 1, 2\}$
Number of walks per node (r)	20	Negative samples (k)	5		
Context window size (w)	10	Weight decay	0.0		
		Learning rate (η)	3e-4		
		Number of training epochs	5		

DetConstSort. DetConstSort is a deterministic constrained re-ranking method proposed by Geyik et al. (2019). It acts as a post-processor and re-ranks the initial list by prioritizing high-scoring items and enforcing group proportion constraints at each rank position. Due to Geyik et al. (2019) not providing an official public implementation, we base our implementation on the repository by Ghosh et al. (2021) and adapt it to the link prediction setting by treating candidate edges as ranked items and edge-types as groups. We apply DetConstSort on the exact same output ranking produced by the three edge-type specific models that were used in the MORAL training pipeline. Thus, for DetConstSort we do not change the training procedure.

C Statistics

This Section provides statistics on the datasets and on the environmental impact of this study.

C.1 Dataset statistics

Table 9 gives more dataset statistics in supplement of Table 1 in Section 3.4, with gradients per epoch as

$$\text{gradients per epoch} = \frac{|\mathbb{E}_{\text{train}}|}{|\mathbb{S}| \cdot \text{batch size} \cdot \binom{|\mathbb{S}|}{2}},$$

where $|\mathbb{S}|$ is the sensitive attribute cardinality and $|\mathbb{E}_{\text{train}}|$ is the number of training edges (Mattos et al., 2025). We replicate Mattos et al. (2025)'s statistics (in their *Table 1*) except for slight deviations in $|E|$ (of maximum 3 edges), probably due to different preprocessing, e.g., (missing) deduplication. Table 10 shows homophily statistics for the datasets and classifies the datasets into homophily levels.

C.2 Environmental impact statistics

The total computational expense for reproducing the original experiments equals 17.96 GPU hours with an estimated 1.24 $kgCO_2eq$. For reference, the estimated carbon intensity of the Netherlands was 411 $g\,CO_2eq\,kWh^{-1}$ in January 2026 (Nowtricity, 2026). Table 11 breaks these numbers down further. The total computational expense for the homophily extension equals 30.41 GPU hours with an estimated 1.41 $kgCO_2eq$. Emissions were estimated using CodeCarbon (Lacoste et al., 2023).

By summing these estimated emissions, we get a total of 2.65 $kgCO_2eq$ for the above experiments, which, by taking the recent estimated carbon intensity per kWh of the Netherlands into our calculation, is equivalent to heating ±70 liters of water from 20 °C to 100 °C.

Table 9: Dataset statistics including edge-type proportions. The upper six datasets are the datasets used by Mattos et al. (2025) and the lower two are added for testing different topologies, see Appendix E. GARN stands for Global Airport Routes Network. GpE denotes the gradients per epoch.

Dataset	$\lvert\mathbb{V}\rvert$	$\lvert\mathbb{E}\rvert$	**Feature dim.**	**Attribute**	$\mathbb{E}_{0\text{-}1}$	$\mathbb{E}_{1\text{-}1}$	$\mathbb{E}_{0\text{-}0}$	**Topology**	**GpE**
Credit	30000	96162	13	Age	12%	87%	1%	Periphery	47
Facebook	1045	18723	573	Gender	42%	44%	13%	Periphery	10
German	1000	15218	27	Gender	20%	61%	19%	Periphery	8
NBA	403	7434	95	Nationality	28%	63%	9%	Periphery	4
Pokec-n	66569	361932	276	Gender	54%	23%	22%	Community	177
Pokec-z	67796	432569	265	Gender	5%	58%	37%	Community	212
GARN	7698	18858	5	Hub/non-hub	27%	68%	5%	Core-periphery	7
Chameleon	890	8854	2326	See Section E	50%	14%	35%	Flat	4

Table 10: Edge homophily relative to random mixing across datasets. We report the observed edge homophily h_{edge} (Eq. 4), the expected homophily under random mixing $h_{\text{edge}}^{\text{rand}}$ (Eq. 9), and the excess homophily $\Delta h = h_{\text{edge}} - h_{\text{edge}}^{\text{rand}}$ (Eq. 10). We classify Δh as homophilic ($\Delta h > 0$) or heterophilic ($\Delta h < 0$) relative to random mixing, with low/medium/high categories based on $|\Delta h|$.

Dataset	$h_{\text{edge}}^{\text{rand}}$	h_{edge}	Δh	**Homophily type**
Credit	0.8370	0.8790	0.0420	Low homophily
Facebook	0.5502	0.5757	0.0255	Low homophily
German	0.5722	0.8048	0.2326	High homophily
NBA	0.6100	0.7237	0.1137	Medium homophily
Pokec-n	0.5003	0.4560	-0.0443	Low heterophily
Pokec-z	0.5001	0.4507	-0.0494	Low heterophily
Global Airport Routes Network	0.8162	0.7301	-0.0861	Medium heterophily
Chameleon	0.5730	0.4955	-0.0775	Medium heterophily

Table 11: Environmental impact metrics for reproducing the core link prediction models, each trained on 500 epochs, on six different datasets. Mean and standard deviation over 3 seeds are reported.

Dataset	$kgCO_2eq$	**Total** $kgCO_2eq$	kWh	**Total** kWh	**Duration** (min)
Credit	0.0200 ± 0.0059	0.1198	0.0746 ± 0.0221	0.4477	22.39 ± 0.04
Facebook	0.0021 ± 0.0000	0.0064	0.0080 ± 0.0000	0.0240	3.24 ± 0.00
German	0.0029 ± 0.0001	0.0087	0.0109 ± 0.0003	0.0326	2.67 ± 0.00
NBA	0.0023 ± 0.0001	0.0070	0.0087 ± 0.0002	0.0261	2.22 ± 0.00
Pokec-n	0.1476 ± 0.0223	0.4428	0.5515 ± 0.0833	1.6544	134.66 ± 0.35
Pokec-z	0.2190 ± 0.0006	0.6569	0.8182 ± 0.0022	2.4545	171.63 ± 0.63

D Additional homophily results

In this section, we support Section 4.2 with additional results. Figure 6 shows a heatmap for h_{edge}, $h_{\text{edge}}^{\text{rand}}$, Δh_{edge}, Δ_{DP}, NDKL, Precision, AWRF, and NDCG, as a function of h_{mm}, h_{MM}, and f_{m}. Here, by Newman (2003)'s random-mixing method (and substituting with f_{m}), we define the expected homophily under random mixing as

$$h_{\text{edge}}^{\text{rand}} = f_{\text{m}}^2 + (1 - f_{\text{m}})^2. \tag{9}$$

Consequently, the relative/excess homophily is

$$\Delta h_{\text{edge}} = h_{\text{edge}} - h_{\text{edge}}^{\text{rand}}. \tag{10}$$

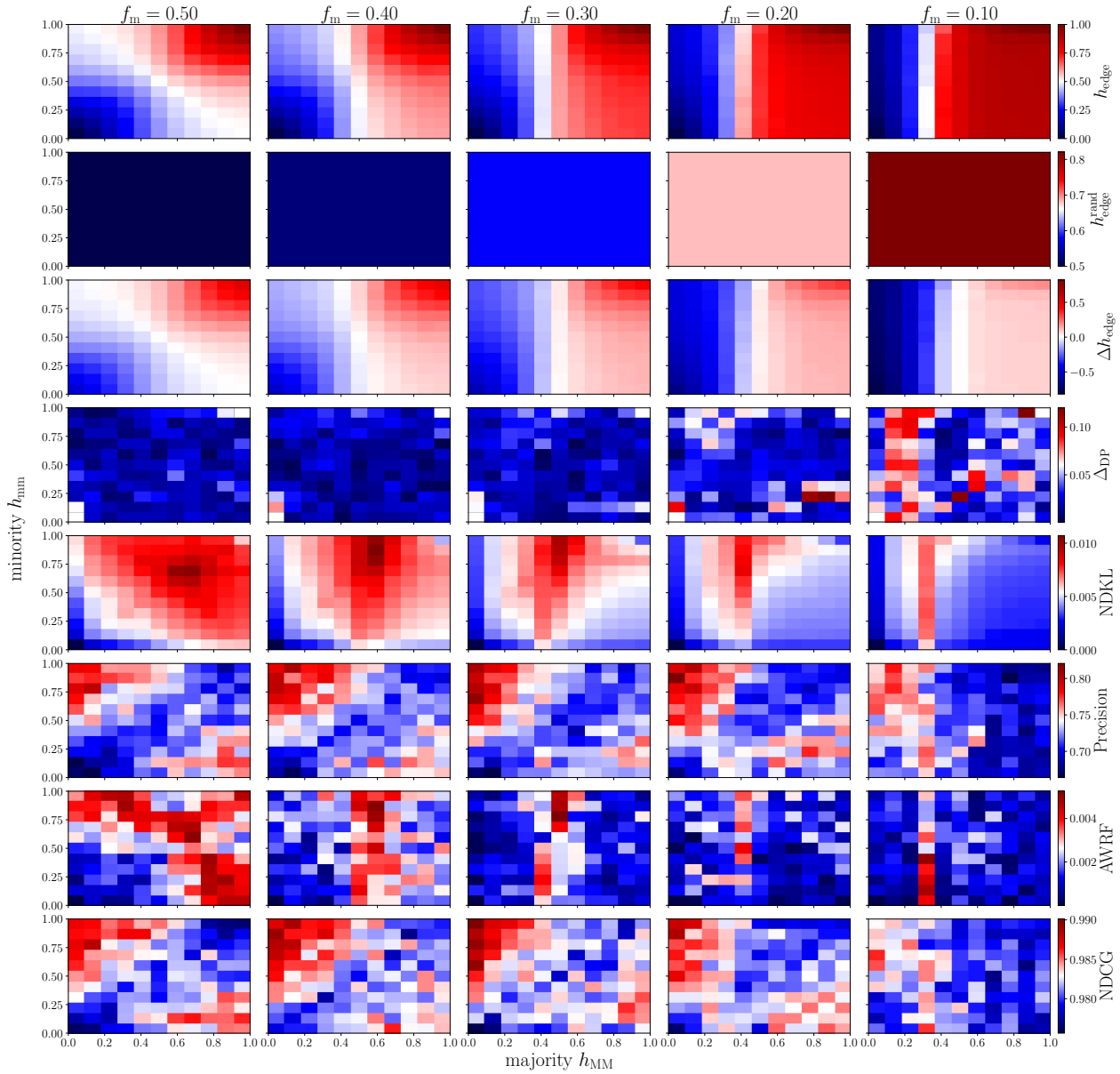

Figure 6: Edge homophily h_{edge} (Eq. 4), random homophily baseline $h_{\text{edge}}^{\text{rand}}$ dependent on minority fraction f_{m}, homophily w.r.t baseline $\Delta h_{\text{edge}} = h_{\text{edge}} - h_{\text{edge}}^{\text{rand}}$, Δ_{DP} (Eq. 2), NDKL (Eq. 3), Precision, AWRF (Eq. 7), and NDCG as a function of minority parameter f_{mm} and majority parameter f_{MM}, for different values of f_{m}, all at $K = 1000$. As seen by the top row, absolute homophily generally increases with f_{m} and h_{MM}.

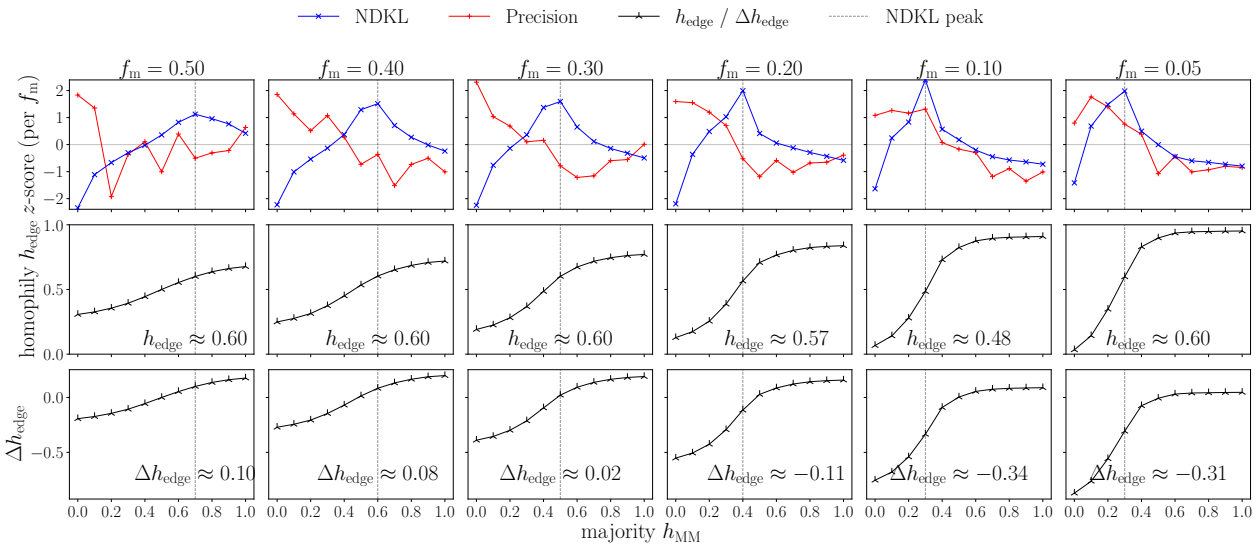

Figure 7: **Top:** z-scored NDKL and Precision versus majority homophily h_{MM} for decreasing minority fraction f_{m}, mean-averaged across minority homophily h_{mm}. **Middle:** corresponding edge homophily h_{edge}. Vertical dotted lines indicate NDKL peaks, with the corresponding h_{edge} value annotated. **Bottom:** corresponding relative edge homophily Δh_{edge}.

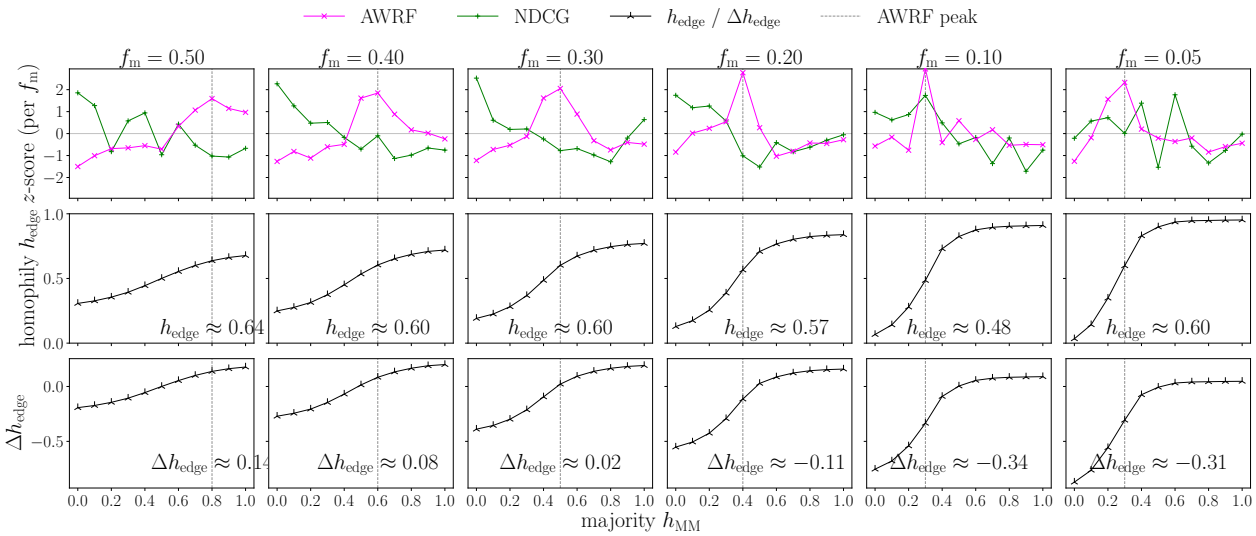

Figure 8: **Top:** z-scored AWRF and NDCG versus majority homophily h_{MM} for decreasing minority fraction f_{m}, mean-averaged across minority homophily h_{mm}. **Middle:** corresponding edge homophily h_{edge}. Vertical dotted lines indicate NDKL peaks, with the corresponding h_{edge} value annotated. **Bottom:** corresponding relative edge homophily Δh_{edge}.

In the heatmaps, we first observe that the absolute edge homophily h_{edge} increases as h_{MM} and h_{mm} increase. As Eq. 9 takes just f_{m}, the random-mixing baseline $h_{\text{edge}}^{\text{rand}}$ heatmap is constant for each minority fraction f_{m}. Therefore, Δh_{edge} highlights where the observed homophily is above or below the random mixing baseline. We see that Δ_{DP} reacts comparatively little and inconsistently to increases of h_{edge}.

Figures 7 and 8 show a similar pattern. NDKL and AWRF increase initially (indicating more unfair ranking predictions), but drop once absolute homophily reaches around $h_{\text{edge}} \approx 0.60$ (in most cases). Importantly,

this turning point does not align with a fixed value of the relative homophily Δh_{edge}. This implies that these exposure-based fairness metrics are more closely tied to absolute homophily than to relative homophily.

A plausible explanation is that absolute homophily directly shapes the composition of the top of the ranking. When edges concentrate more strongly within groups, a majority-with-majority structure can dominate the highest scores and positions, which is exactly where exposure metrics are most sensitive (Singh & Joachims, 2018) (for example by following the logarithmic decay of Eq. 1). A normalized measure such as Δh_{edge} can miss this, because it corrects for group size and can take similar values under completely different absolute mixing patterns.

Although in the literature relative homophily is used to compare mixing patterns across group sizes (e.g., the assortativity perspective and inbreeding homophily) (McPherson et al., 2001; Newman, 2003; Currarini et al., 2010; Platonov et al., 2023a), our results suggest that exposure-based fairness metrics in ranked link prediction are primarily sensitive to absolute homophily h_{edge} rather than to relative homophily Δh_{edge}. In particular, the aforementioned NDKL and AWRF peaks occur at similar h_{edge}-levels across settings while occuring at different Δh_{edge}-levels, thus implying that comparable Δh_{edge} can yield different exposure (disparities) and visibility at the top of the ranking. Thus, absolute homophily gives a more informative signal than relative homophily for explaining exposure-based fairness in rankings.

E Verifying generalizability to additional datasets

In this Appendix Section, we verify the generalizability of MORAL to two additional real-world datasets. Whereas the original datasets, see Section 3.4, had *Community* and *Periphery* topologies, the Global Airport Routes Network (GARN) dataset has a *Core-periphery* topology such that we can test generalizability to a different topology. With the second dataset, Chameleon, we complement our synthetic homophily experiments with a realistic alternative and show MORAL's fairness there is not an artefact of the generator.

Designing a binary sensitive attribute for GARN. For GARN (Peixoto, 2020), the goal is to design the sensitive attribute such that several cores will form with peripheral nodes around it. Airport networks naturally resemble a core-periphery topology (Gallagher et al., 2021; Diop et al., 2023). One can argue for the need of fair link prediction in airport networks by economic reasons, as, e.g. through "rich-get-richer" effects (Subramonian et al., 2023), smaller airports could be disadvantaged in route planning recommendations.

By sorting the airports in descending order by degree, we get a power-law decaying pattern (Guimera & Amaral, 2004). As Guimera & Amaral (2004) point out, the most connected airports are typically not the most central ones. Hence, proper classification of the nodes into hub and non-hub requires a method that goes beyond just degree. One method for this is the Borgatti-Everett (BE) core-periphery detection algorithm by Borgatti & Everett (2000). BE finds a partition of nodes into a dense core and sparse periphery by comparing the connectivity in the observed graph to an ideal core-periphery pattern, and assigns each node without the need for hyperparameters. We use the results of the algorithm as binary attributes.

Designing a binary sensitive attribute for Chameleon. The Chameleon dataset is a Wikipedia page-page network, first introduced by Pei et al. (2020) and later refined by Platonov et al. (2023b), whose version we use. It is often used as a heterophilic benchmark, as, e.g., in Luo et al. (2024).

The Chameleon dataset has 5-ary sensitive attributes. To obtain binary sensitive attributes while keeping the heterophilic mixing, we try all possible binary partitions and select the binary partition that minimizes the fraction of within-group edges. To keep group sizes roughly balanced, we divide the groups into one group of two and one group of three, giving $\binom{5}{3} = 10$ possible groups $\mathbb{S}$. To choose the most heterophilic group $\mathbb{S}^\star$, we find the group with minimal absolute edge homophily h_{edge} (Eq. 4), yielding $h_{\text{edge}}(\mathbb{S}^\star) = 0.4955$. By Eq. 9, we get $h_{\text{edge}}^{\text{rand}} \approx 0.573$ as random baseline. Since $h_{\text{edge}}(\mathbb{S}^\star) < h_{\text{edge}}^{\text{rand}}$, the found attribute distribution exhibits heterophily relative to chance, which we qualify as "Medium heterophily" in Table 10.

Results of additional datasets. Table 12 shows the results of training on the dataset. MORAL predicts GARN with near-perfect scores, similar to the scores on other datasets described in Section 4.1.

MORAL performs even better on Chameleon. By comparing Chameleon's $h_{\text{edge}} \approx 0.5$ to results of synthetic homophily experiments with a similar absolute edge homophily, e.g. those in Figure 3, we see NDKL is

comparable between the synthetic and real-data setting, though Precision is in a significantly lower range in the synthetic setting. Most plausibly, this can be explained by the difference in features between the scenarios: the DPAH features are non-informative with Gaussian i.i.d. features ($d = 16$), whereas Chameleon has high-dimensional ($d = 2326$) real features, likely making prediction substantially easier.

Overall, the GARN results suggest that MORAL's fairness-utility behaviour generalizes to a different topology. On Chameleon, MORAL exhibits comparable exposure-based fairness behaviour in a heterophilous real-world setting, supporting transfer of the homophily stress-test conclusions beyond synthetic graphs.

Table 12: NDKL, AWRF, Precision, and NDCG at $K = 1000$ for the additional datasets.

Model	**Metric**	Global Airport Routes Network	Chameleon
MORAL	Δ_{DP}@1000	0.0164 ± 0.0015	0.0011 ± 0.0009
	NDKL@1000	0.0053 ± 0.0000	0.0074 ± 0.0000
	AWRF@1000	0.0009 ± 0.0000	0.0015 ± 0.0000
	Precision@1000	0.9587 ± 0.0090	0.9860 ± 0.0065
	NDCG@1000	0.9967 ± 0.0010	0.9994 ± 0.0003

F Results from the original paper

In this Appendix, we include the main results from the original paper (Mattos et al., 2025) for reference. Figure 9 shows the proportions of edge types in the top-100 per method they used compared to the original graph distribution. Table 13 compares performance of all their approaches by Precision@1000 and NDKL.

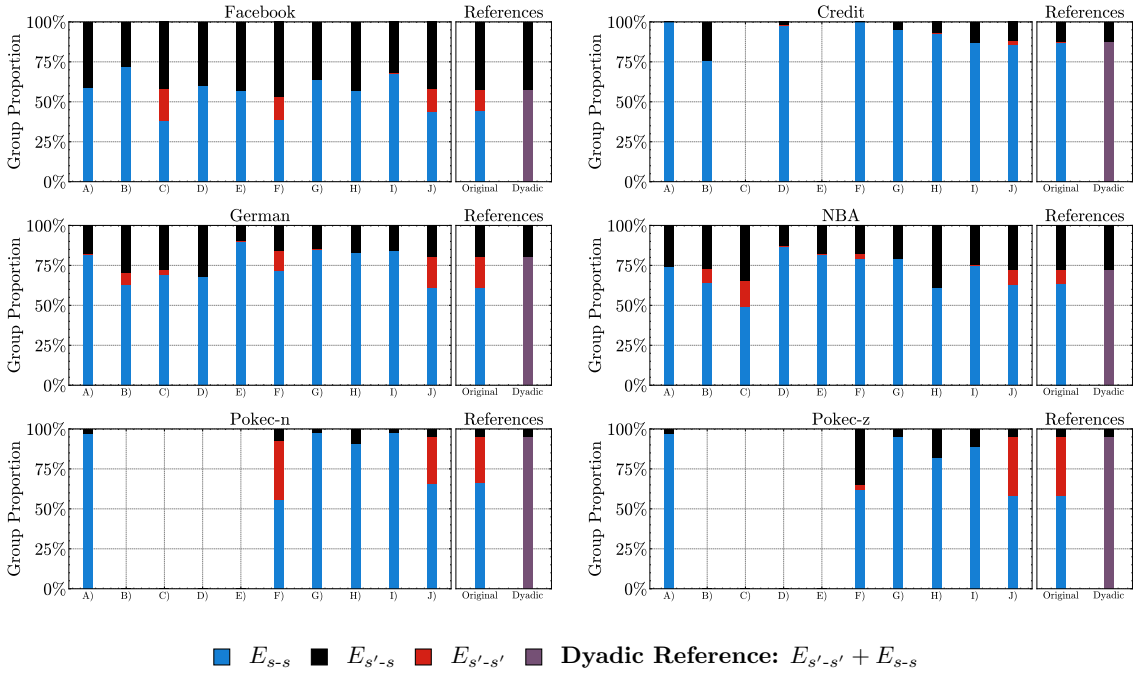

Figure 9: Adopted from Mattos et al. (2025). Proportions of pair types in the top-100 predictions by method, compared against the original graph distribution and an optimal dyadic fairness reference. Colors: $E_{s\text{-}s}$ (**blue**), $E_{s'\text{-}s}$ (**black**), and $E_{s'\text{-}s'}$ (**red**). 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.

Table 13: Adopted from Mattos et al. (2025). Fairness performance comparison of all approaches considered ($k = 1000$). Lower NDKL and higher Precision@1000 are better. Best NDKL and Precision@1000 values are in **bold** and underline, respectively.

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
	Precision@1000	0.97 ± 0.00	1.00 ± 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
	Precision@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
	Precision@1000	0.96 ± 0.01	1.00 ± 0.00	0.73 ± 0.01	0.69 ± 0.01	0.97 ± 0.03	1.00 ± 0.00
FairEGM	NDKL	0.09 ± 0.01	0.11 ± 0.00	0.05 ± 0.01	0.07 ± 0.01	OOM	OOM
	Precision@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
	Precision@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
	Precision@1000	0.96 ± 0.00	1.00 ± 0.00	0.94 ± 0.00	0.55 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
FairAdj	NDKL	0.10 ± 0.05	OOM	0.10 ± 0.01	0.11 ± 0.05	OOM	OOM
	Precision@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
	Precision@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
	Precision@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	**0.04 ± 0.00**	**0.01 ± 0.00**	**0.03 ± 0.00**	**0.02 ± 0.00**	**0.03 ± 0.00**	**0.04 ± 0.00**
	Precision@1000	0.95 ± 0.01	1.00 ± 0.00	0.96 ± 0.00	0.80 ± 0.00	0.98 ± 0.00	0.98 ± 0.00