# Fairness–Utility Trade-offs in Graph Learning (Reformatted for Imperial/DSI)

Prepared by: Fan Xuejiao · 2025-08-29 ·

## Abstract

This short, reproducible study examines how a demographic-parity (DP) penalty changes the behaviour of graph neural networks (GNNs) on a controlled synthetic graph and what accuracy we trade off. For *node classification*, increasing the penalty drives DP and equalised odds (EO) towards ~0 while accuracy drops from 1.00 to 0.71 ($\lambda$: 0 $\rightarrow$ 2.0). For *link prediction*, the effect becomes visible only when the DP objective is applied to *both positive and negative edges* and its weight is increased: DP falls from ≈0.08 ($\lambda$=0) to 0 ($\lambda$=10) while AUC remains ≈0.71–0.72 and AP shifts modestly. The practical lesson is that the *definition and population of the fairness penalty must match the reported metric* for results to be interpretable and reproducible.

## Introduction

Fairness in graph learning is often framed as a trade-off with utility; whether that trade-off appears depends on details. Two guiding questions: (1) will a simple DP penalty produce a visible trade-off on a basic synthetic network? (2) which design choices make the effect visible *and* reproducible?

## Setup

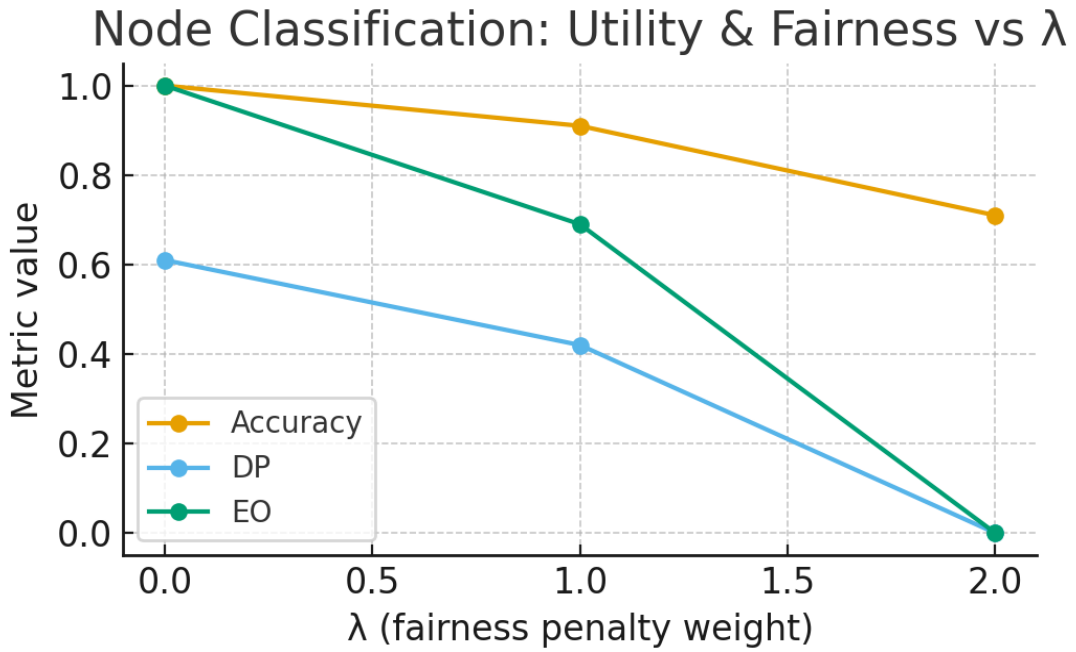
**Data.** Homophilous LocalSBM (70/30 split), $p_{in}$=0.08, $p_{out}$=0.005; moderately separable node features. **Sensitive proxy.** Eigenvector centrality (median split). **Models.** Two-layer GNN (hidden=64). For link prediction, candidate edges are scored using observed positives and sampled negatives. **Training objective.** Binary cross-entropy + $\lambda$·DP penalty. For node classification, the penalty uses predicted class probabilities. For link prediction, the penalty is computed on probabilities for both positive and negative edges so it matches evaluation.

## Results

**Node classification.** $\lambda$=0 $\rightarrow$ Acc 1.00, DP 0.61, EO 1.00; $\lambda$=1.0 $\rightarrow$ Acc 0.91, DP 0.42, EO 0.69; $\lambda$=2.0 $\rightarrow$ Acc 0.71, DP 0.00, EO 0.00. This is the textbook trade-off.

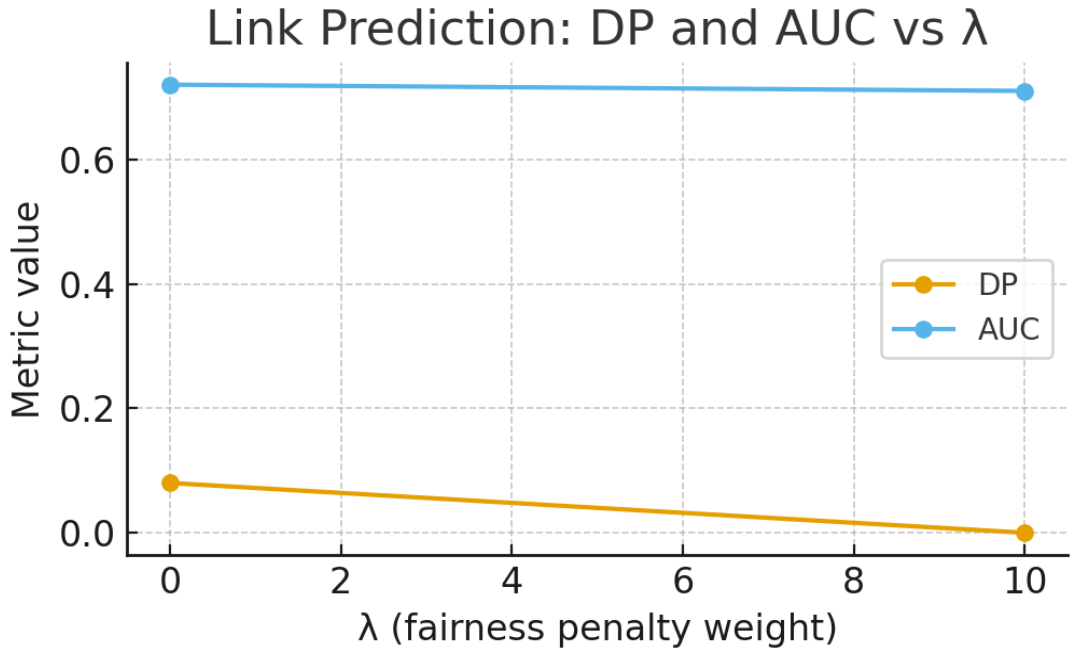
| $\lambda$ | Accuracy | DP | EO |
|---|---|---|---|
| 0.00 | 1.00 | 0.61 | 1.00 |
| 1.00 | 0.91 | 0.42 | 0.69 |
| 2.00 | 0.71 | 0.00 | 0.00 |

Figure 1. Node Classification — Utility & Fairness vs $\lambda$




**Link prediction.** When the DP penalty is applied to both positives and negatives and its weight is increased, DP reduces from ≈0.08 (λ=0) to 0 (λ=10) while AUC remains ≈0.71–0.72 and AP changes modestly.

Figure 2. Link Prediction — DP and AUC vs λ (two-point summary)




## Discussion

**What matters.** (i) Align the fairness objective with the evaluated population (here: include both positives and negatives). (ii) Adjust task difficulty: moderate feature separation and adequate negatives make the trade-off visible at realistic λ. (iii) Report exact values for fairness and utility, not only trends, and keep logs/figures for reproducibility.

## Reproducibility

Commands (sketch). Node classification: $\lambda \in \{0.0, 1.0, 2.0\}$; link prediction: $\lambda \in \{0, 5, 10\}$; 150 epochs; hidden=64; LocalSBM 70/30, p_in=0.08, p_out=0.005; sensitive=eigen; figures via make_figures.py. JSON logs are stored alongside plots to ensure reproducibility.

**Imperial/DSI Alignment.** This work pattern—careful measurement, fair/utility trade-offs, and auditable artefacts—translates naturally to *privacy & memorization audits* and *auditable ML pipelines* (data lineage, confidential computing, model/risk cards) under GDPR/PIPL. It aligns with Dr Yve-Alexandre de Montjoye's focus on privacy and memorization and can benefit from co-supervision with Prof Peter Pietzuch (systems/TEEs) or Dr Thomas Heinis / Dr Holger Pirk (data lineage/indexing).