Fair-PaperRec: Fair Learning for Bias Mitigation and Quality Optimization in Paper Recommendation

Uttamasha Anjally Oyshi, Susan Gauch

The Seventeenth International Conference on Information, Process, and Knowledge Management
eKNOW2025

Presenter: Uttamasha Anjally Oyshi
University of Arkansas, USA
Email: *uoyshi@uark.edu*

Presenter Biography

- Uttamasha Anjally Oyshi - Ph.D. student, Computer Science, University of Arkansas
- Graduate Research Assistant, EECS Department
- Masters in Computer Science, University of Arkansas at Little Rock
- Bachelors in Electronics and Communication Engineering, Khulna University of Engineering & Technology, Bangladesh
- Research Interests: Fairness in AI, bias mitigation, causal inference, deep learning in recommendation systems

Motivation

- **Persistent demographic gaps** in academia - affects race, gender, nationality, age, disability
- **Lack of diversity affects paper ecosystems -** fewer marginalized voices as authors, reviewers, speakers
- **SIGCHI 2020 set diversity goals**
  - Acknowledged limits of anonymity-based masking
- **Double-blind review is not enough**
  - Reviewers often infer authorship via writing style, citations, preprints
  - Leads to biased acceptance toward familiar demographics
- **Evidence of systemic bias -** reviewers favor similar backgrounds
- **Need for fairness-aware post-review tools**
  - Correct residual bias *without compromising quality*

Research Goals

- **Develop a fairness-aware recommendation model** for post-review conference paper selection
- **Integrate author demographics** into a neural network using paper review metadata
- **Design a custom fairness loss function** that balances paper quality (via h-index, conference tier) with demographic parity
- **Simulate real-world review behavior** by modeling conference-specific tiers (e.g., strong/weak accept) to mimic varying levels of review strictness and selection pressure
- **Ensure demographic parity** across **multiple protected attributes** (e.g., race, country)

Approach Overview

- **Fair-PaperRec**: Fairness-aware MLP model for post-review paper recommendation
- Targets both **individual** (race & country) and **intersectional fairness** (race + country) while preserving paper quality
- Applies **after double-blind review** to correct residual bias in accepted papers

Related Work: Paper Recommendation Approaches

- **Burke (2017)** introduced multi-sided fairness, accounting for fairness across different stakeholders (e.g., users and providers) in recommender systems.
- **Beutel et al. (2017)** proposed adversarial debiasing, using adversarial networks to learn fair representations that remove sensitive attribute information.
- **Alsaffar et al. (2021)** implemented greedy re-ranking, adjusting the output rankings post-hoc to satisfy fairness constraints without retraining the model.
- **Bobadilla et al. (2021)** developed DeepFair, a deep learning framework that integrates fairness constraints directly into the learning process for recommender systems.

Dataset and Feature Overview

- Conference data used: SIGCHI, DIS, IUI
  - SIGCHI as **'strong accept'** (high reputation), DIS as **'weak accept'**, and IUI as 'weak reject'.
  - This structure **reflects real-world paper evaluation** distributions.
  - Incorporated into training for **better utility-fairness tradeoff** modeling.
- Features: h-index, race, country, conference ID
- Target: paper acceptance label
- Author metadata: race, country, gender, career stage
- Data split: 80% training / 20% validation (stratified)

Architecture Design

- Simple Multi Layer Perceptron (MLP) with 2 hidden layers, ReLU and batch normalization
- Inputs: h-index, citation-based features; protected attributes excluded from input
- Outputs: Probability of paper acceptance

Model Architecture

- ❑ Fair-PaperRec Model Overview
  - ❑ Input: Author attributes, h-index, conference ratings
  - ❑ Output: Acceptance probabilities ensuring fairness

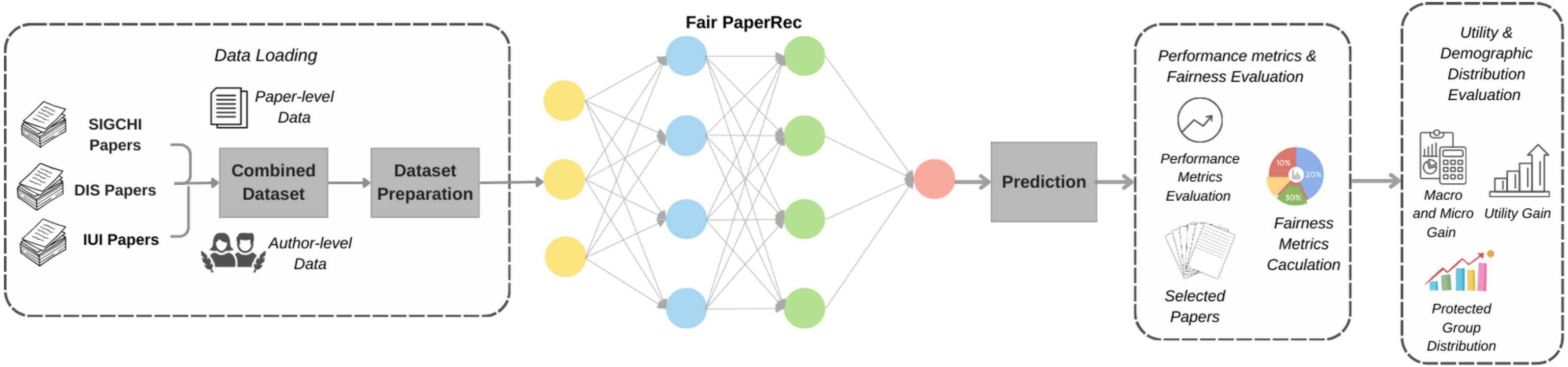


Figure 1: Overview of the Fair-PaperRec Architecture.

Fairness Loss Function

- Total Loss: $L\_total = L\_prediction + \lambda * L\_fairness$
- L_fairness penalizes demographic disparity in acceptance rates
- λ controls trade-off between fairness and accuracy
- Allows post-review adjustment without retraining entire model

Experimental Setup

- PyTorch, NVIDIA GPUs, Adam optimizer
- Metrics: Macro Gain, Micro Gain, Utility Gain
- Early stopping for robustness

Implementation and Training

- Trained with Adam optimizer, early stopping, stratified data split
- λ hyperparameter tuned for fairness–utility tradeoff
- Repeated experiments ensure robustness across conference datasets

Results and Analysis: Impact of Fairness Constraints on Country Representation

- Best trade-off for country fairness observed at λ = 2.5.
- Macro and Micro Gains improve with increasing λ.
- Slight utility drop at high λ shows trade-off with fairness

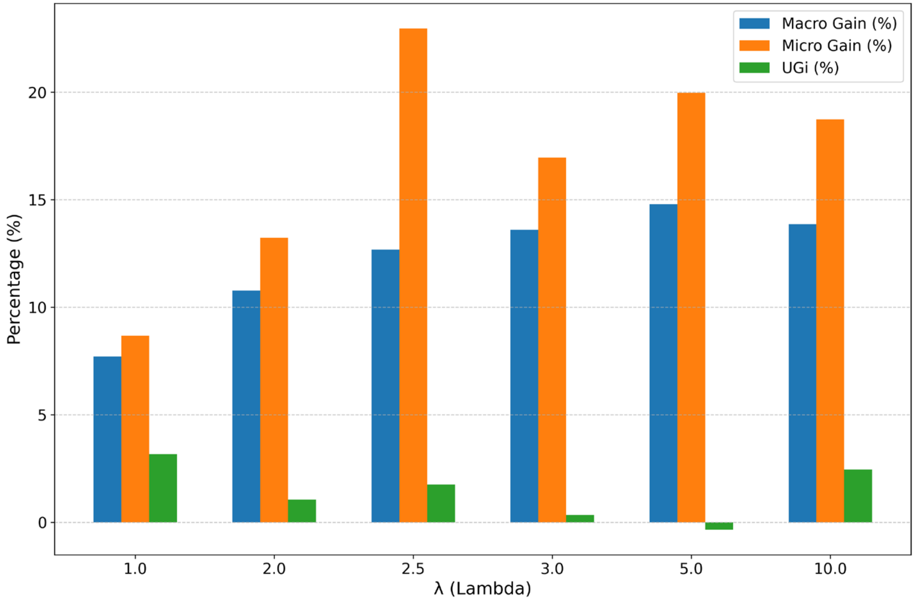


Figure 2: Comparison of Macro and Micro Gains for Country Across Different Fairness Configurations.

Results and Analysis: Impact of Fairness Constraints on Race Representation

- Optimal λ = 3 shows 42.03% macro and 56.48% micro gain.
- Diversity improves steadily with λ for racial fairness.
- Utility remains stable at optimal λ.

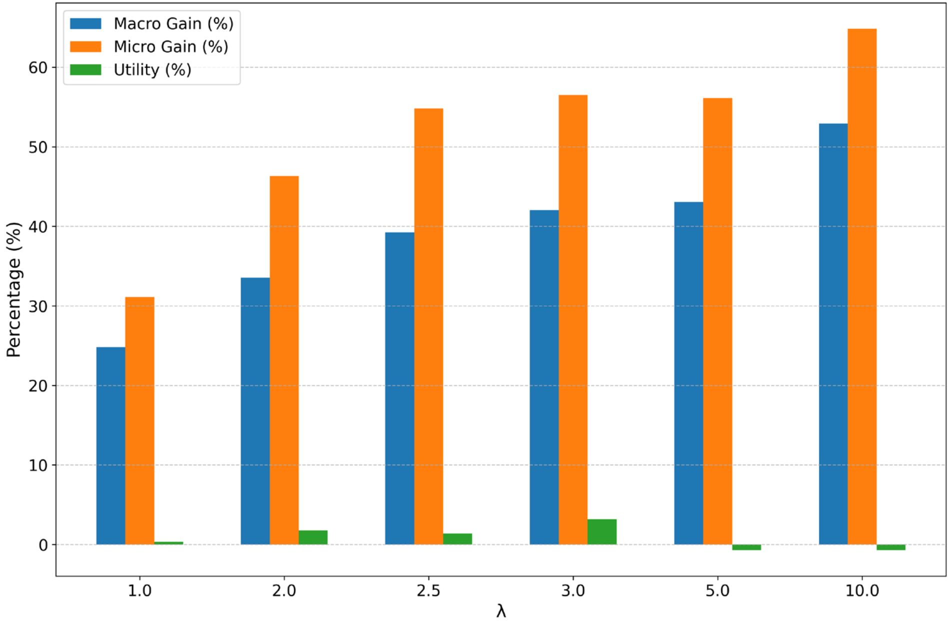
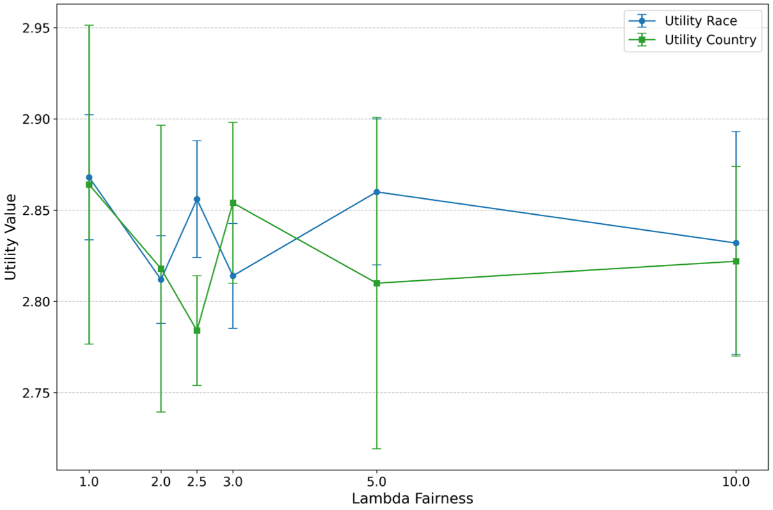


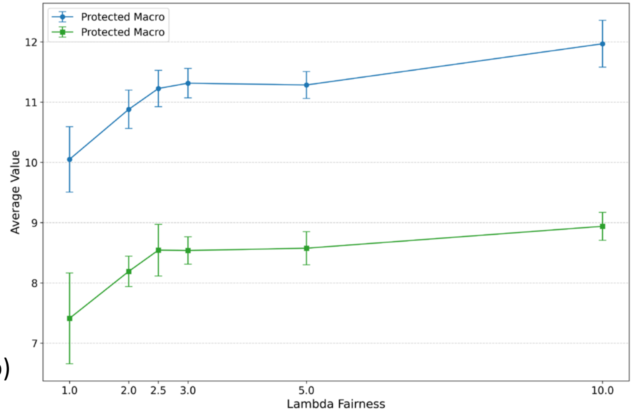
Figure 3: Comparison of Macro and Micro Gains for Race Across Different Fairness Configurations.

Results and Analysis

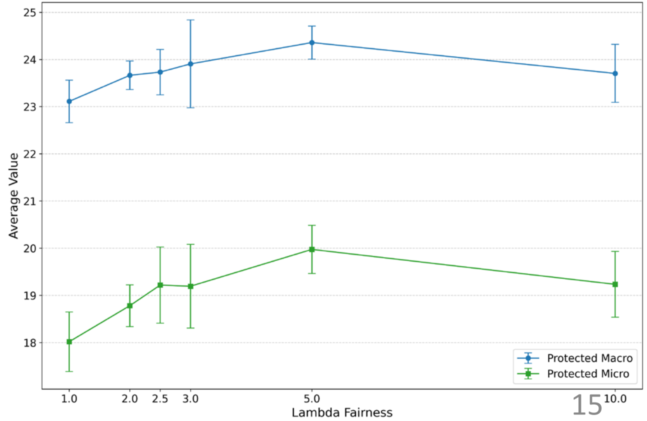
- Optimal λ identified: Race λ=3, Country λ=2.5
- Diversity increases with λ
- Modest impact on quality (Utility Gain: 3.16%)



(a)



(b)



(c)

Figure 4. Comparison of gains across different fairness configurations. (a) Utility Gain. (b) Macro/Micro for Race. (c) Macro/Micro for Country.

Ablation Study: Gain calculations for country and race features with utility gain

- Evaluate Fair-PaperRec's performance when simultaneously enforcing fairness across race and country
- Use different weight combinations and regularization strengths (λ)

λ	Weights	Country Feature		Race Feature		$\mathrm{UG}_i(\%)$	Avg. $\mathrm{D}_G(\%)$	Avg. F (%)
		Macro Gain (%)	Micro Gain (%)	Macro Gain (%)	Micro Gain (%)			
1	$W_r = 0.32, W_c = 0.68$	6.17	6.34	30.51	46.30	3.16	44.66	53.71
	$W_r = 1, W_c = 2$	6.73	9.15	-0.25	0.37	2.81	6.48	13.77
	$W_r = 2, W_c = 1$	7.43	11.43	12.91	16.11	3.16	25.63	40.36
2	$W_r = 0.32, W_c = 0.68$	**13.60**	**24.43**	30.51	42.22	**4.21**	55.38	68.47
	$W_r = 1, W_c = 2$	5.24	6.88	15.45	17.96	0.70	20.69	21.58
	$W_r = 2, W_c = 1$	8.36	12.86	39.49	54.26	1.75	26.31	21.58
2.5	$W_r = 0.32, W_c = 0.68$	8.63	17.33	36.58	50.37	2.46	56.46	66.31
	$W_r = 1, W_c = 2$	9.89	14.00	30.63	46.30	2.81	40.52	62.09
	$W_r = 2, W_c = 1$	9.60	17.11	42.53	56.48	1.40	59.25	69.98
3	$W_r = 0.32, W_c = 0.68$	7.15	11.42	39.49	53.89	1.40	55.98	63.45
	$W_r = 1, W_c = 2$	10.16	21.17	33.29	43.89	0.70	43.45	47.63
	$W_r = 2, W_c = 1$	9.60	18.35	42.53	55.37	2.81	61.90	47.63
5	$W_r = 0.32, W_c = 0.68$	10.80	19.38	**45.82**	**58.52**	0.70	**65.09**	**72.92**
	$W_r = 1, W_c = 2$	4.69	3.88	33.92	40.19	0.35	38.61	15.73
	$W_r = 2, W_c = 1$	7.43	11.90	39.49	52.96	5.26	52.26	15.73
10	$W_r = 0.32, W_c = 0.68$	9.60	18.34	42.53	55.37	1.40	62.92	70.89
	$W_r = 1, W_c = 2$	7.43	13.91	24.94	25.19	4.91	32.37	34.88
	$W_r = 2, W_c = 1$	7.43	11.72	35.44	47.41	-4.21	40.53	34.88

Key Findings from Ablation Study

- Using λ = 2.5 with weights Wr = 0.32, Wc = 0.68 achieved balanced gains:
  - Race: 36.58% Macro, 50.37% Micro
  - Country: 8.63% Macro, 17.33% Micro
  - Utility Gain: 2.46%, Avg. Diversity Gain: 56.46%, Avg. F-measure: 66.31%
- Increasing race weight improves diversity for both race and country.Increasing country weight may harm race fairness
- Higher λ boosts fairness but lowers utility, confirming trade-off
- Use different weight combinations and regularization strengths (λ)

Main Contributions

- Fair-PaperRec: A fairness-aware neural model that reframes paper acceptance as a recommendation problem
- Introduces a custom fairness loss for post-review demographic correction across race and country
- Models' intersectional fairness through weighted loss terms for multiple attributes
- Achieves fairness without modifying the original peer-review pipeline
- Evaluated on real-world conference data (SIGCHI, DIS, IUI), demonstrating fairness–utility trade-offs

Key Findings

- Double-blind review is insufficient: Bias persists due to identity cues and systemic imbalance.
- Fairness-aware post-review correction can significantly improve diversity:
  - Up to 42.53% macro gain in racial representation.
  - Utility remains stable (e.g., 3.16% gain in some settings).
- Optimal fairness is attribute-sensitive:
  - Race requires higher λ due to deeper disparities.
  - Weight calibration is critical when optimizing multiple attributes.

Future Direction

- Causal Inference
  - Explore causal links between demographic traits and paper acceptance
  - Apply models like ATE, Counterfactual Fairness
  - Construct causal DAGs to estimate direct/indirect effects
  - Improves explainability and bias source diagnosis
- Advanced neural models i.e., Variational AutoEncoders (VAE), Graph Neural Networks (GNNs)
- Add more attributes beyond race and geolocation

Acknowledgement

- Supported by National Science Foundation (NSF)
- Award No. OIA-1946391 (DART)

Thank You!