Assessing Unfairness in GNN-Based Recommender Systems: A Focus on Metrics for Demographic Sub-Groups

Nikzad Chizari[1][a], Keywan Tajfar[2][b] and María N. Moreno-García[1][c]
[1]Department of Computer Science and Automation, University of Salamanca, Plaza de los Caídos sn, 37008 Salamanca, Spain
[2]College of Science, School of Mathematics, Statistics, and Computer Science, Department of Statistics, University of Tehran, Tehran, Iran
{nikzadchizari, mmg}@usal.es, tajfar@ut.ac.ir

Keywords: Recommender Systems, Bias, Fairness, Graph Neural Networks, Metrics.

Abstract: Recommender Systems (RS) have become a central tool for providing personalized suggestions, yet the growing complexity of modern methods, such as Graph Neural Networks (GNNs), has introduced new challenges related to bias and fairness. While these methods excel at capturing intricate relationships between users and items, they often amplify biases present in the data, leading to discriminatory outcomes especially against protected demographic groups like gender and age. This study evaluates and measures fairness in GNN-based RS by investigating the extent of unfairness towards various groups and su bgroups within these systems. By employing performance metrics like NDCG, this research highlights disparities in recommendation quality across different demographic groups, emphasizing the importance of accurate, group-level measurement. This analysis not only sheds light on how these biases manifest but also lays the groundwork for developing more equitable recommendation systems that ensure fair treatment across all user groups.

1 INTRODUCTION

Recommender systems (RS) are advanced algorithms that suggest relevant items to users by analyzing their preferences and behaviors. By examining user-item interactions, among other data, these systems deliver personalized recommendations, helping users discover new products and services while mitigating information overload, thus enhancing user engagement (Chen et al., 2020; Zheng and Wang, 2022).

GNN-based RS use Graph Neural Networks (GNNs) to improve recommendation quality by capturing complex relationships between users and items in graph-structured data (Zhou et al., 2020; Zhang et al., 2021). GNNs aggregate information from neighboring nodes, learning patterns that lead to more accurate, personalized recommendations, outperforming traditional methods (Steck et al., 2021; Khan et al., 2021; Mu, 2018). However, this approach can unintentionally amplify biases, as the clustering of similar sensitive attributes in social graphs may result in biased representations and recommendations, exacerbating fairness issues (Dai and Wang, 2021).

Sensitive attributes such as race, gender, religion, age, and disability are protected by privacy laws to prevent discrimination, making it essential for RS to consider these factors to avoid biased recommendations under European or US regulations (Di Noia et al., 2022; Floridi et al., 2022). Failing to do so can result in significant economic, legal, ethical, and security risks for both companies and users (Di Noia et al., 2022; Fahse et al., 2021; Wang et al., 2023). In response, international organizations have stressed the need to understand, measure, and mitigate bias, particularly in sensitive areas, and have implemented regulations to address these concerns (Di Noia et al., 2022). In GNN-based RS, bias measurement focuses on assessing disparities between user groups based on protected attributes, with this study concentrating on group fairness and the evaluation of bias toward specific protected groups.

1.1 Group Fairness

Group fairness ensures that algorithms do not produce biased predictions or decisions against individuals in any specific sensitive group. In this section, we dis-

[a] https://orcid.org/0000-0002-7300-6126
[b] https://orcid.org/0000-0001-7624-5328
[c] https://orcid.org/0000-0003-2809-3707

Chizari, N., Tajfar, K. and Moreno-García, M.
Assessing Unfairness in GNN-Based Recommender Systems: A Focus on Metrics for Demographic Sub-Groups.
DOI: 10.5220/0013069400003825

In *Proceedings of the 20th International Conference on Web Information Systems and Technologies (WEBIST 2024)*, pages 433-440
ISBN: 978-989-758-718-4; ISSN: 2184-3252

cuss common fairness concepts under group fairness. Fairness notations under group fairness are described in Figure 1.

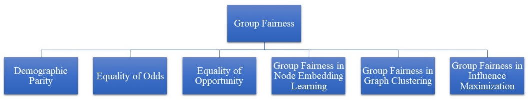


Figure 1: Group fairness notations.

Several key metrics are commonly used to measure group fairness in GNN-based RS. Demographic Parity ensures even distribution of recommendations across groups, with Generalized Demographic Parity (GDP) addressing continuous sensitive features (Rahman et al., 2019; Spinelli et al., 2021). Equality of Odds and Equality of Opportunity focus on balancing True Positive and False Positive Rates across groups (Hardt et al., 2016; Spinelli et al., 2021). Distribution-based fairness uses metrics like the Wasserstein distance to maintain fairness in node embeddings, while model-based fairness makes predicting sensitive features difficult (Du et al., 2020; Navarin et al., 2020). Other fairness measures include Balance Score for clustering, Maxmin Fairness for influence maximization (Dong et al., 2023; Rahmattalabi et al., 2021), and Disparate Impact (DI) and Treatment Equality (TE) for comparing outcomes across groups (Purificato et al., 2022). Additionally, NDCG (Normalized Discounted Cumulative Gain) is adapted to assess ranking fairness across demographic groups (Chizari et al., 2023b).

Traditional bias evaluation metrics are often unsuitable for RS due to the unique characteristics and goals of these systems compared to general machine learning models (Chen et al., 2020; Chizari et al., 2023a). While typical bias metrics focus on consistent predictions across demographic groups, RS-specific metrics emphasize identifying biases in recommended items, reflecting the system's focus on predicting user preferences, which naturally vary across groups (Wang et al., 2022). RS bias evaluation also requires more nuanced approaches, such as addressing long-tail effects, popularity bias, and personalization discrepancies. Research by Ekstrand et al. (2018), Chizari et al. (2022), Chen et al. (2023), and Gao et al. (2023) underscores the significance of fairness in enhancing recommendation quality and user experience.

2 LITERATURE REVIEW

This section provides a review of the state-of-the-art literature, focusing on the challenges of bias and unfairness in RS and specifically GNN-based RS. The objective is to thoroughly explore and understand the measurement techniques employed in prior studies, while also identifying their strengths and weaknesses.

GNN-based RS can enhance accuracy but may also exacerbate bias and fairness issues due to their graph structures and message-passing mechanisms (Steck et al., 2021; Khan et al., 2021; Mu, 2018; Chizari et al., 2022; Dai and Wang, 2021). For example, homophily in social networks, where nodes with similar sensitive attributes (e.g., age, gender) are more likely to connect, can lead to biased representations and unfair outcomes (Dai and Wang, 2021).

To measure fairness in these systems, various studies have utilized metrics like statistical parity and classification performance (e.g., NDCG, MRR, AUC) to assess the influence of sensitive attributes on recommendations (Rahman et al., 2019; Wu et al., 2020; Neophytou et al., 2022; Wu et al., 2021). Some approaches focus on measuring performance across multiple groups using variance (Rahman et al., 2019) or the largest gap between groups (Spinelli et al., 2021). However, these techniques can lead to information loss, as variance is sensitive to outliers and does not reveal which groups are most disadvantaged. Additionally, focusing only on the two groups with the highest and lowest accuracy can overlook significant disparities among other groups.

The study by (Boratto et al., 2024) examines the robustness of recommendations in GNN-based RS from both user and provider perspectives, using Demographic Parity (DP) to evaluate group fairness. It also defines consumer preference and satisfaction metrics using NDCG and precision. However, this research is limited by the range of models considered and primarily addresses fairness against poisoning-like attacks based on edge-level perturbations.

Several studies have explored fairness in RS using metrics like NDCG, MRR, AUC, and RBP, but they face challenges such as lack of focus on multiple demographic groups, information loss, and accuracy issues. (Gómez et al., 2022) focus on traditional RS without addressing multiple groups, while (Rahman et al., 2019; Neophytou et al., 2022; Wu et al., 2021) note the absence of sensitive attribute combinations. (Spinelli et al., 2021) highlights fairness gaps but struggles with accuracy, and (Boratto et al., 2024) limit their analysis to Demographic Parity across a small model range. Addressing these issues requires choosing appropriate fairness metrics for GNN-based RS based on the task and data characteristics.

3 METHODOLOGY

This study evaluates group and subgroup unfairness in GNN-based RS, focusing on age and gender. Age is divided into four groups (less than 20, 20 to 40, 40 to 60, and above 60) to assess accuracy consistency across categories. Subgroups like young men, young women, old men, and old women are also analyzed to identify biases at the intersection of these attributes. Various performance metrics are applied to each group and subgroup to assess recommendation quality. Three real-world datasets are used for a comprehensive analysis of potential disparities in accuracy.

3.1 Benchmark Datasets

This study uses three well-known real-world datasets, including MovieLens (Harper and Konstan, 2019) 100K, LastFM 100K (Celma, 2010), and Book Recommendation (Mobius, 2020). These datasets have been the main sources of the majority of research in this section.

3.2 Recommendation Approaches

This study uses various approaches with different models including Collaborative Filtering (CF), Matrix Factorization (MF), and GNN-based models:

Table 1: Model Categories and References.

Category	Model	References
CF	ItemKNN (Item K Nearest Neighbour)	(Al-Ghamdi et al., 2021; Airen and Agrawal, 2022)
	NNCF (Neural Network Collaborative Filtering)	(Sang et al., 2021; Girsang et al., 2021)
MF	DMF (Deep Matrix Factorization)	(Xue et al., 2017; Yi et al., 2019; Liang et al., 2022)
	NeuMF (Neural Collaborative Filtering)	(Kuang et al., 2021; Zhang et al., 2016)
GNN-based	LightGCN (Light Graph Convolutional Network)	(Broman, 2021; Ding et al., 2022)
	NGCF (Neural Graph Collaborative Filtering)	(Wang et al., 2021; Sun et al., 2021)
	SGL (Self-Supervised Graph Learning)	(Yang, 2022; Tang et al., 2021)
	DGCF (Disentangled Graph Collaborative Filtering)	(Bourhim et al., 2022; Sha et al., 2021)

3.3 Evaluation Metrics

Two types of metrics, evaluation and fairness, are used to assess model performance. This approach helps evaluate both model accuracy and its behavior toward protected groups. All metrics are applied based on the top-K ranked items in the recommendation list (K represents the list size).

3.3.1 Model Evaluation Metrics

The metrics for evaluating the top K recommendation lists are defined below using the notation described in Table 2.

Table 2: Table of notations.

Notation	Definition
U	Set of users
I	Set of items
u	User
i	Item
$R(u)$	Ground-truth set of items that user u interacted with
$\hat{R}(u)$	Ranked item list produced by a model
K	Length of the recommendation list

- **Recall@K:** Recall is a measure that computes the fraction of relevant items out of all relevant items on the top-K list.
$$Recall@K = \frac{1}{|U|}\sum_{u\in U}\frac{|\hat{R}(u)\cap R(u)|}{|R(u)|} \tag{1}$$
- **Precision@K:** Precision or positive predictive value is a measure that calculates the fraction of relevant items out of all the recommended items on the list. The average is calculated for each user to gather the final result. $|\hat{R}(u)|$ denotes the item count of $\hat{R}(u)$.
$$Precision@K = \frac{1}{|U|}\sum_{u\in U}\frac{|\hat{R}(u)\cap R(u)|}{|\hat{R}(u)|} \tag{2}$$
- **Mean Reciprocal Rank (MRR)@K:** The MRR computes the corresponding rank of the first relevant item found in the top-k list. $Rank_u^*$ represents be the position of that item in the list provided by a given algorithm for the user u.
$$MRR@K = \frac{1}{|U|}\sum_{u\in U}\frac{1}{Rank_u^*} \tag{3}$$
- **Hit Ratio (HR)@K:** HR or Hit calculates how many 'hits' are in a top-K recommendation list. HR requires at least one item that falls in the ground-truth set. $\delta(0)$ is an indicator function. $\delta(b) = 1$ if b is true; otherwise it would be 0. $\emptyset$ denotes the empty set.
$$HR@K = \frac{1}{|U|}\sum_{u\in U}\delta(\hat{R}(u)\cap R(u) \neq \emptyset) \tag{4}$$
- **Normalized Discounted Cumulative Gain (NDCG)@K:** NDCG is a measure of ranking quality, where positions are discounted logarithmically. It accounts for the position of the

hit by assigning higher scores to hits at the top ranks. NDCG is the ratio between DCG and the maximum possible DCG.$\delta(0)$ is an indicator function.

$$NDCG@K = \frac{1}{|U|}\sum_{u\in U}\frac{1}{\sum_{i=1}^{min(|R(u)|,K)}\frac{1}{log_2(i+1)}} \cdot \sum_{i=1}^{K}\delta(i\in R(u))\frac{1}{\log_2(i+1)} \quad (5)$$

4 RESULTS

This section presents the experimental results from three real-world datasets, along with a brief EDA to illustrate user distribution within each group. It first displays performance results for each group, followed by a detailed analysis of the subgroups.

4.1 Exploratory Data Analysis (EDA)

This section presents an exploratory data analysis (EDA) of user distribution (not interaction) across groups, focusing on user count rather than interactions. In Figure 2 the gender distribution for MovieLens and LastFM datasets shows a higher number of male users compared to female users, indicating potential bias in the data that could lead to unfairness.

Figure 3 indicates the age distribution of the users in the groups for the three datasets. Here also the charts show an unequal distribution across groups, with the senior user group being in the minority, at a great distance from the other groups.

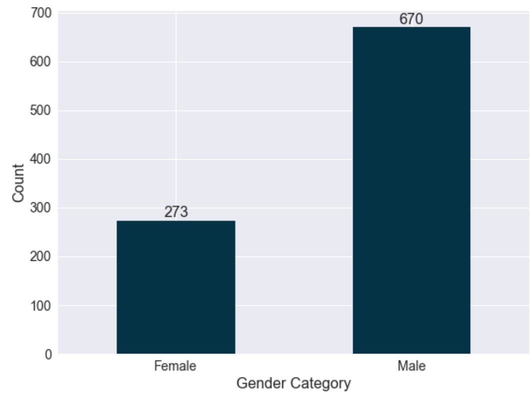
4.2 Results of Overall Performance

In this part, the results of all of the used metrics on all models across the datasets can be seen. All charts present the performance of the models on the @10 recommendation list. Figure 4 shows the overall results without considering any subgroups using different metrics on all datasets. It can be seen that On the LastFM and the Book Recommendation datasets, the overall performance of GNN models is lower.

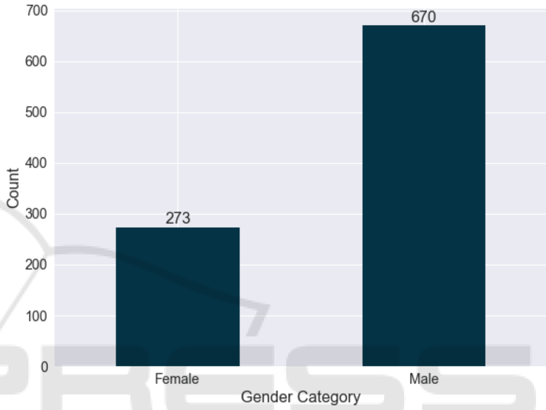
4.3 Results of Each Group

In the first part of the experiment, the results of the NDCG metric are calculated for each group for the three datasets.

Figure 5 shows the NDCG results for gender on the MovieLens and LastFM datasets. In the MovieLens dataset, SGL has the lowest performance and the largest difference between groups, indicating it



(a) MovieLens



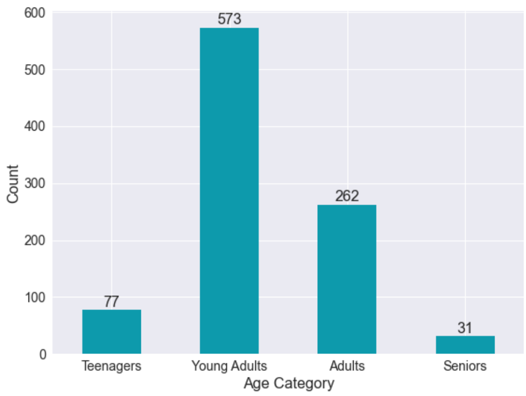
(b) LastFM

Figure 2: Gender distribution for MovieLens and LastFM indicating the bias in data toward men.

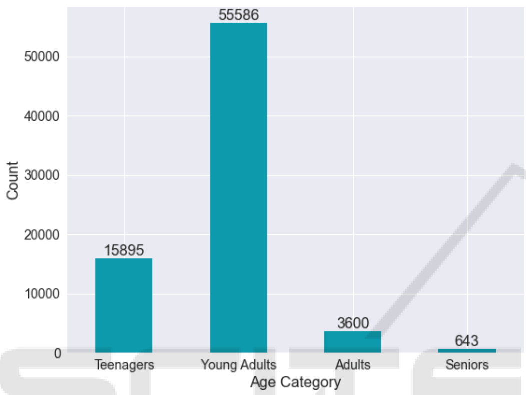
is prone to discrimination. Other GNN models perform better with minor unfairness, though LightGCN shows slight bias. In the LastFM dataset, GNN models generally underperform compared to conventional methods, with ItemKNN also performing poorly. Significant group differences are only observed in SGL and LightGCN.

Figure 6 also represents the results of NDCG performance for sensitive attribute age including teenagers (less than 20), young adults (20 to 40), adults (40 to 60), and seniors (more than 60) on all datasets.

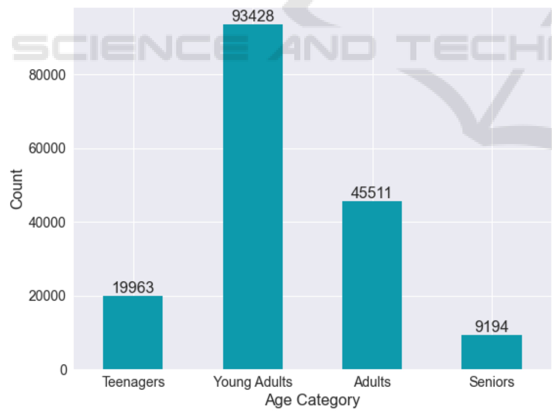
In the MovieLens dataset, SGL performed poorly with significant group differences, while NGCF and DGCF showed strong performance and low discrimination. Traditional models had higher unfairness. In the LastFM, ItemKNN and other GNN models underperformed, with NGCF and DGCF showing more unfairness. Similarly, in the Book Recommendation dataset, GNN models performed poorly, with NeuMF, DGCF, DMF, and NNCF showing higher unfairness, while NGCF had the lowest unfairness.



(a) MovieLens

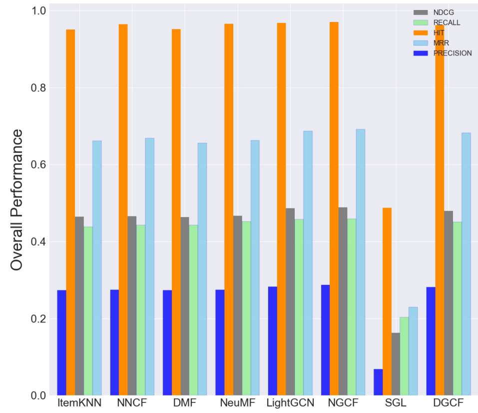


(b) LastFM

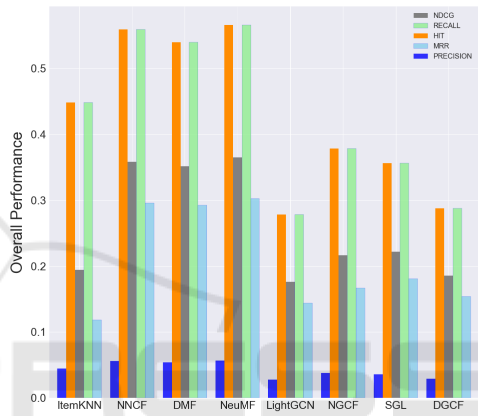


(c) Book Recommendation

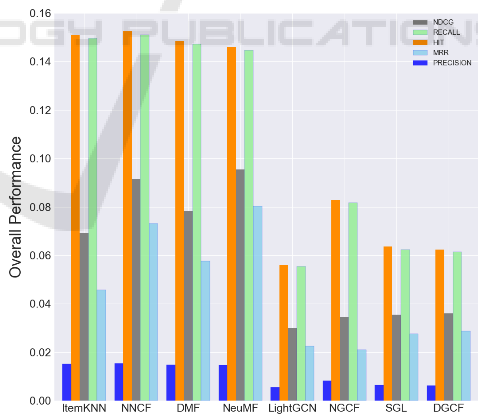
Figure 3: Age distribution for the three datasets showing the bias in data across the groups.



(a) MovieLens



(b) LastFM



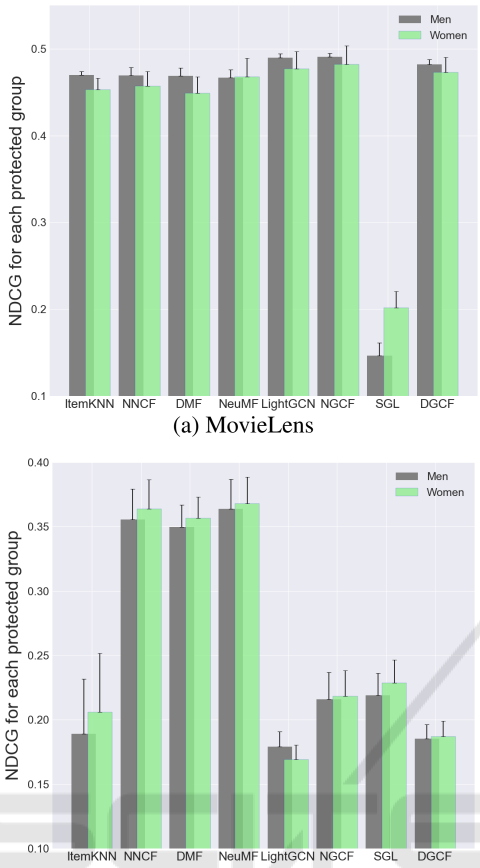
(c) Book Recommendation

Figure 4: Performance results across all datasets based on various metrics.

4.4 Bias and Fairness Results for Each Subgroup

The second stage analyzes NDCG performance across subgroups, providing insights into model behavior within each subgroup and helping to better understand unfairness toward protected groups.

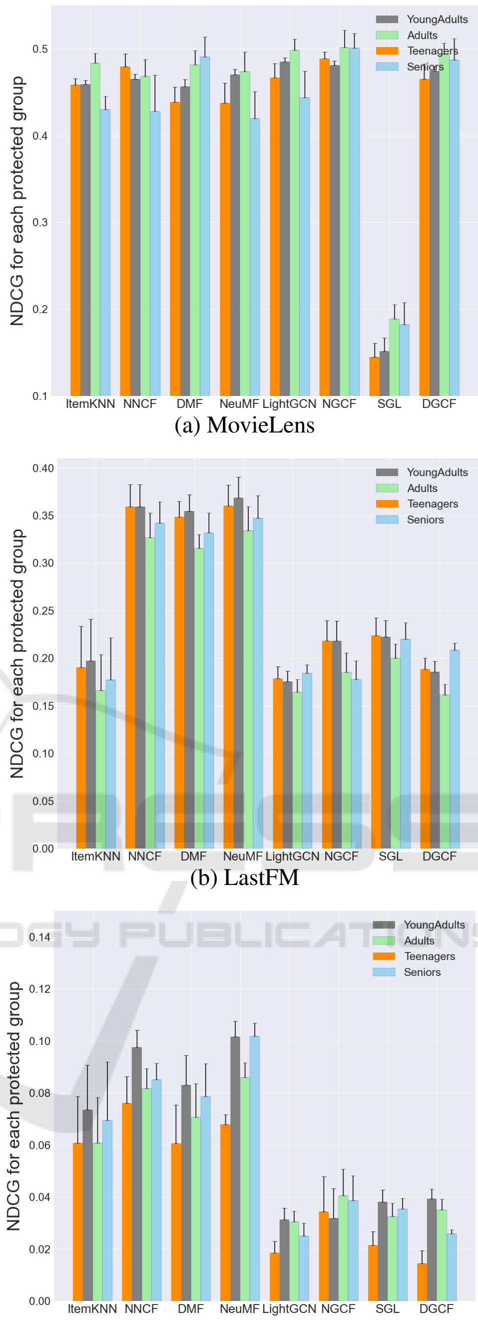
In the MovieLens dataset, NGCF, DGCF, and DMF performed better for a specific subgroup, possibly due to user distribution, while other models showed fair results. In the LastFM dataset, the same subgroup (women seniors) had the poorest results, highlighting that a lack of users in certain subgroups can lead to higher unfairness.



(a) MovieLens

(b) LastFM

Figure 5: Results of NDCG performance for sensitive attribute gender on MovieLens and LastFM.



(a) MovieLens

(b) LastFM

(c) Book Recommendation

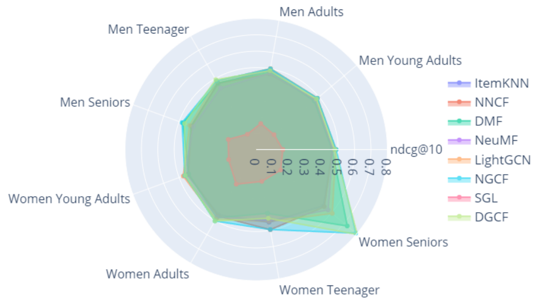
Figure 6: Results of NDCG performance for sensitive attribute age on MovieLens, LastFM, and Book Recommendation.

5 CONCLUSIONS AND FUTURE WORK

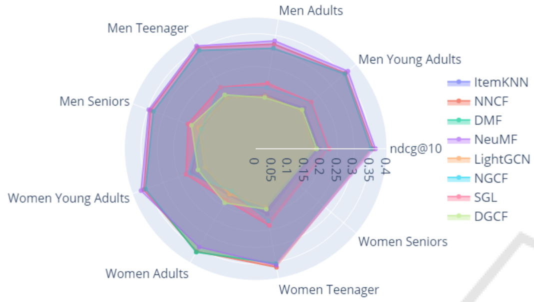
This research highlights the challenges of ensuring fairness in GNN-based recommender systems (RS) for protected demographic groups, particularly concerning gender and age. The exploratory data analysis (EDA) reveals uneven user distributions across groups in the datasets, contributing to biased outcomes and potential discrimination.

Model results indicate that considering gender and age groups leads to lower performance of SGL on the MovieLens, LastFM, and Book Recommendation datasets, exacerbating unfairness toward unprotected groups. In the LastFM dataset, GNN models generally underperformed regarding both gender and age. Analysis of subgroups reveals that specific subgroups suffer from unfairness, with NGCF and DGCF showing higher performance in the MovieLens dataset, while most models exhibit greater unfairness toward certain subgroups in LastFM.

Overall, while GNN models demonstrate strong performance, user group distribution significantly impacts fairness. This necessitates further investigation into data distribution and the application of preprocessing methods to mitigate biases. Future work



(a) MovieLens



(b) LastFM

Figure 7: Results of NDCG performance for subgroups on Movielens and LastFM.

should focus on selecting models that align with data behavior and task requirements, incorporating techniques like counterfactual fairness, adversarial training, and personalized approaches. Additionally, exploring hybrid solutions that integrate graph structures with debiasing mechanisms could enhance both performance and fairness in various recommendation tasks.

REFERENCES

Airen, S. and Agrawal, J. (2022). Movie recommender system using k-nearest neighbors variants. *National Academy Science Letters*, 45(1):75–82.

Al-Ghamdi, M., Elazhary, H., and Mojahed, A. (2021). Evaluation of collaborative filtering for recommender systems. *International Journal of Advanced Computer Science and Applications*, 12(3).

Boratto, L., Fabbri, F., Fenu, G., Marras, M., and Medda, G. (2024). Robustness in fairness against edge-level perturbations in gnn-based recommendation. In *European Conference on Information Retrieval*, pages 38–55. Springer.

Bourhim, S., Benhiba, L., and Idrissi, M. J. (2022). A community-driven deep collaborative approach for recommender systems. *IEEE Access*.

Broman, N. (2021). Comparasion of recommender systems for stock inspiration.

Celma, O. (2010). *Music Recommendation and Discovery in the Long Tail*. Springer.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.

Chizari, N., Shoeibi, N., and Moreno-García, M. N. (2022). A comparative analysis of bias amplification in graph neural network approaches for recommender systems. *Electronics*, 11(20):3301.

Chizari, N., Tajfar, K., and Moreno-García, M. N. (2023a). Bias assessment approaches for addressing user-centered fairness in gnn-based recommender systems. *Information*, 14(2):131.

Chizari, N., Tajfar, K., Shoeibi, N., and Moreno-García, M. N. (2023b). Quantifying fairness disparities in graph-based neural network recommender systems for protected groups. In *WEBIST*, pages 176–187.

Dai, E. and Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688.

Di Noia, T., Tintarev, N., Fatourou, P., and Schedl, M. (2022). Recommender systems under european ai regulations. *Communications of the ACM*, 65(4):69–73.

Ding, S., Feng, F., He, X., Liao, Y., Shi, J., and Zhang, Y. (2022). Causal incremental graph convolution for recommender system retraining. *IEEE Transactions on Neural Networks and Learning Systems*.

Dong, Y., Kose, O. D., Shen, Y., and Li, J. (2023). Fairness in graph machine learning: Recent advances and future prospectives. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5794–5795.

Du, M., Yang, F., Zou, N., and Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34.

Fahse, T., Huber, V., and Giffen, B. v. (2021). Managing bias in machine learning projects. In *International Conference on Wirtschaftsinformatik*, pages 94–109. Springer.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., and Wen, Y. (2022). capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act. *Available at SSRN 4064091*.

Girsang, A. S., Wibowo, A., et al. (2021). Neural collaborative for music recommendation system. In *IOP Conference Series: Materials Science and Engineering*, volume 1071, page 012021. IOP Publishing.

Gómez, E., Boratto, L., and Salamó, M. (2022). Provider fairness across continents in collaborative recommender systems. *Information Processing & Management*, 59(1):102719.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Harper, F. M. and Konstan, J. A. (2019). Movie-

Lens 25m dataset. https://grouplens.org/datasets/movielens/25m/.

Khan, Z. Y., Niu, Z., Sandiwarno, S., and Prince, R. (2021). Deep learning techniques for rating prediction: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54(1):95–135.

Kuang, H., Xia, W., Ma, X., and Liu, X. (2021). Deep matrix factorization for cross-domain recommendation. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 2171–2175. IEEE.

Liang, G., Sun, C., Zhou, J., Luo, F., Wen, J., and Li, X. (2022). A general matrix factorization framework for recommender systems in multi-access edge computing network. *Mobile Networks and Applications*, pages 1–13.

Mobius, A. (2020). Book recommendation dataset.

Mu, R. (2018). A survey of recommender systems based on deep learning. *Ieee Access*, 6:69009–69022.

Navarin, N., Van Tran, D., and Sperduti, A. (2020). Learning kernel-based embeddings in graph neural networks. In *ECAI 2020*, pages 1387–1394. IOS Press.

Neophytou, N., Mitra, B., and Stinson, C. (2022). Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In *European Conference on Information Retrieval*, pages 641–654. Springer.

Purificato, E., Boratto, L., and De Luca, E. W. (2022). Do graph neural networks build fair user models? assessing disparate impact and mistreatment in behavioural user profiling. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4399–4403.

Rahman, T., Surma, B., Backes, M., and Zhang, Y. (2019). Fairwalk: Towards fair graph embedding.

Rahmattalabi, A., Jabbari, S., Lakkaraju, H., Vayanos, P., Izenberg, M., Brown, R., Rice, E., and Tambe, M. (2021). Fair influence maximization: A welfare optimization approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11630–11638.

Sang, L., Xu, M., Qian, S., and Wu, X. (2021). Knowledge graph enhanced neural collaborative filtering with residual recurrent network. *Neurocomputing*, 454:417–429.

Sha, X., Sun, Z., and Zhang, J. (2021). Disentangling multi-facet social relations for recommendation. *IEEE Transactions on Computational Social Systems*.

Spinelli, I., Scardapane, S., Hussain, A., and Uncini, A. (2021). Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, 3(3):344–354.

Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., and Basilico, J. (2021). Deep learning for recommender systems: A netflix case study. *AI Magazine*, 42(3):7–18.

Sun, W., Chang, K., Zhang, L., and Meng, K. (2021). Ingcf: An improved recommendation algorithm based on ngcf. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 116–129. Springer.

Tang, H., Zhao, G., Wu, Y., and Qian, X. (2021). Multisample-based contrastive loss for top-k recommendation. *IEEE Transactions on Multimedia*.

Wang, S., Hu, L., Wang, Y., He, X., Sheng, Q. Z., Orgun, M. A., Cao, L., Ricci, F., and Yu, P. S. (2021). Graph learning based recommender systems: A review. *arXiv preprint arXiv:2105.06339*.

Wang, Y., Ma, W., Zhang*, M., Liu, Y., and Ma, S. (2022). A survey on the fairness of recommender systems. *ACM Journal of the ACM (JACM)*.

Wang, Y., Ma, W., Zhang, M., Liu, Y., and Ma, S. (2023). A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43.

Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., and Wang, M. (2021). Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*, pages 2198–2208.

Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. (2020). Graph neural networks in recommender systems: a survey. *ACM Computing Surveys (CSUR)*.

Xue, H.-J., Dai, X., Zhang, J., Huang, S., and Chen, J. (2017). Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia.

Yang, C. (2022). Supervised contrastive learning for recommendation. *arXiv preprint arXiv:2201.03144*.

Yi, B., Shen, X., Liu, H., Zhang, Z., Zhang, W., Liu, S., and Xiong, N. (2019). Deep matrix factorization with implicit feedback embedding for recommendation system. *IEEE Transactions on Industrial Informatics*, 15(8):4591–4601.

Zhang, Q., Wipf, D., Gan, Q., and Song, L. (2021). A biased graph neural network sampler with near-optimal regret. *Advances in Neural Information Processing Systems*, 34:8833–8844.

Zhang, Z., Liu, Y., Xu, G., and Luo, G. (2016). Recommendation using dmf-based fine tuning method. *Journal of Intelligent Information Systems*, 47(2):233–246.

Zheng, Y. and Wang, D. X. (2022). A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.