Examining Fairness in Graph-Based Collaborative Filtering: A Consumer and Producer Perspective*

Discussion Paper

Dario Di Palma[1,*], Vito Walter Anelli[1,*], Daniele Malitesta[1,*], Vincenzo Paparella[1,*], Claudio Pomo[1,*], Yashar Deldjoo[1] and Tommaso Di Noia[1]

[1]Polytechnic University of Bari, Bari, Italy

Abstract

To date, graph collaborative filtering (CF) strategies have outperformed pure CF models in generating accurate recommendations. However, concerns about fairness and potential biases in recommendations have emerged, as unfair recommendations may harm the interests of Consumers and Producers (CP). Recognizing the lack of a thorough evaluation of graph CF on CP-aware fairness measures, we initially assessed the effects of eight state-of-the-art graph models and four pure CF recommenders on CP-aware fairness measures. Surprisingly, graph CF solutions do not ensure significant item exposure and user fairness. To unravel this performance puzzle, we propose a taxonomy for graph CF, highlighting differences in node representation and neighborhood exploration. Through this lens, the experimental outcomes become clear and pave the way for a multi-objective CP-fairness analysis (Codes are available at: https://github.com/sisinflab/ECIR2023-Graph-CF).

Keywords

Graph Collaborative Filtering, Fairness, Multi-Objective Analysis

1. Introduction and Motivations

Recommender systems (RSs) have evolved from collaborative filtering (CF) to deep learning (DL) models [2, 3, 4, 5], including graph-based methods that represent users and items as nodes in a user-item bipartite graph. Recent research has focused on improving system accuracy and enhancing explainability [6] and reproducibility [7], while issues surrounding fairness [8] remain. Two core aspects of recommendation fairness are producer fairness (item exposure) and consumer fairness (relevance). Existing graph-based approaches have addressed either consumer or producer fairness, but not both simultaneously.

This work aims to bridge the knowledge gap in the literature by studying the effects of state-of-the-art graph strategies on consumer fairness, producer fairness, and system accuracy. We evaluate these dimensions in terms of overall accuracy, user fairness, and item exposure, referring to their combination as CP-fairness when appropriate.

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

*Extended version [1] published in the 45th European Conference on Information Retrieval, ECIR 2023

d.dipalma2@phd.poliba.it (D. D. Palma); vitowalter.anelli@poliba.it (V. W. Anelli); daniele.malitesta@poliba.it (D. Malitesta); vincenzo.paparella@poliba.it (V. Paparella); claudio.pomo@poliba.it (C. Pomo); yashar.deldjoo@poliba.it (Y. Deldjoo); tommaso.dinoia@poliba.it (T. Di Noia)

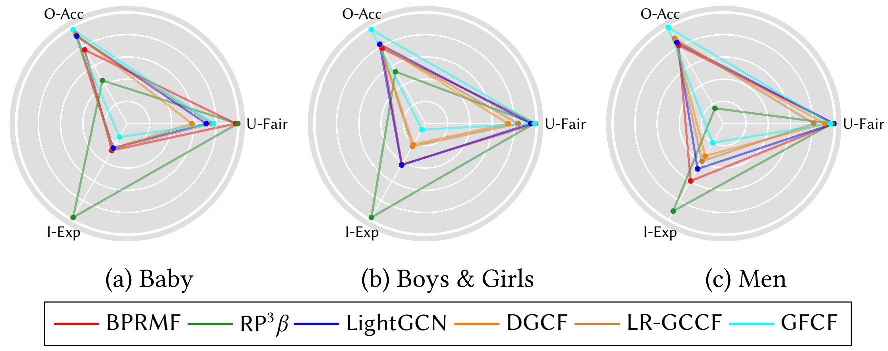

Figure 1: Kiviat diagrams indicating the performance of pure and graph CF recommenders on overall accuracy (i.e., O-Acc, calculated with the *nDCG@20*), item exposure (i.e., I-Exp, calculated with the *APLT@20* [9]), and user fairness (U-Fair, calculated with the *UMADrat@20* [10]). Higher means better.

Motivating Example. We compare leading graph-based models, such as LightGCN, DGCF, LR-GCCF, and GFCF, against classical CF baselines, BPRMF and RP$^3\beta$, on three Amazon datasets. Our evaluation (see Figure 1), considering overall accuracy, user fairness, and item exposure, indicates that graph CF is more accurate, but classical CF shows a better item exposure. No clear winner is found regarding user fairness. We aim to answer two research questions (RQs):
RQ1. Can we explain the variations observed when testing several graph models on overall accuracy, item exposure, and user fairness separately? We analyze the impact of graph CF strategies for nodes representation and neighborhood exploration on accuracy and CP-fairness.
RQ2. How and why do nodes representation and neighborhood exploration algorithms strike a trade-off between overall accuracy, item exposure, and user fairness? Using Pareto optimality [11], we determine the influence of these dimensions in two-objective scenarios, including overall accuracy, item exposure, and user fairness.

2. Nodes Representation and Neighborhood Exploration in Graph Collaborative Filtering: A Formal Taxonomy

Let $\mathcal{G} = (\mathcal{U}, \mathcal{I}, \mathbf{R})$ represent a bipartite, undirected graph connecting N users and M items. We denote the node features for users $u \in \mathcal{U}$ and items $i \in \mathcal{I}$ as embeddings $\mathbf{e}_u \in \mathbb{R}^d$ and $\mathbf{e}_i \in \mathbb{R}^d$, where d is much smaller than N and M. Different applications use message-passing schema to update the model using $\mathcal{G}$. Graph convolutional networks (GCNs), neural graph collaborative filtering (NGCF), and LightGCN are examples of collaborative filtering (CF) applications that utilize message-passing. The need to account for the varying importance of user-item interactions has led to attention mechanisms and models such as graph attention networks (GATs) and disentangled graph collaborative filtering (DGCF). The recent UltraGCN and GFCF address over-smoothing by surpassing and simplifying the traditional message-passing concept. To better understand graph CF models, we propose a classification system based on:

Node representation: This aspect concerns the strategies used for representing users' and items' nodes, including the dimensionality of node embeddings and the weighting of contributions from neighboring nodes.

Neighborhood exploration: This aspect involves the methods for exploring multi-hop neighborhoods to update node representations, including the types of node-node connections and the message-passing schema (either explicit or implicit).

Table 1
Categorization of the chosen graph baselines according to the proposed taxonomy. For each model, we refer to the technical description reported in the original paper and try to match it with our taxonomy.

Models	Nodes Representation				Neighborhood Exploration			
	Latent representation		Weighting		Explored nodes		Message passing	
	low	*high*	*weighted*	*unweighted*	*same*	*different*	*implicit*	*explicit*
GCN-CF* [12]		✓		✓	✓			✓
GAT-CF* [13]		✓	✓		✓			✓
NGCF [14]	✓			✓		✓		✓
LightGCN [15]	✓			✓		✓		✓
DGCF [16]	✓		✓			✓		✓
LR-GCCF [17]	✓			✓	✓	✓		✓
UltraGCN [18]	✓				✓	✓	✓	
GFCF [19]						✓	✓	

*The postfix -CF indicates that we re-adapted the original implementations (tailored for the task of node classification) to the task of personalized recommendation.

We propose (see Table 1) a taxonomy to classify the state-of-the-art graph models. In the following two sections, we will assess eight graphs CF models based on this classification system: GCN-CF [12], GAT-CF [13], NGCF [14], LightGCN [15], DGCF [16], LR-GCCF [17], UltraGCN [18], and GFCF [19].

3. Taxonomy-aware evaluation

This section addresses RQ1 (*"Can we explain the variations observed when testing several graph models on overall accuracy, item exposure, and user fairness separately?"*) by examining how the taxonomy of graph strategies can elucidate recommendation evaluation on CP-Fairness and overall accuracy. We test 48 hyper-parameter configurations on the Amazon Men dataset and top-20 lists (Table 2), focusing on *message-passing*, *explored nodes*, *edge weighting*, and *latent representations*. We report the **best** metric result for each <dimension, value> pair.

Message-passing. We study *implicit* and *explicit* message-passing strategies. Both approaches perform similarly on accuracy and user fairness, but *explicit* techniques perform better on item exposure, particularly *Gini* and *APLT*.

Explored nodes. We examine four node exploration methods: *same* and *different*, with 1 and 2 hops. *Same-2* and *different-1* are the most prominent, with *different-1* outperforming *same-2* on overall accuracy and *same-2* being the best strategy for item exposure. User fairness does not offer a reason to choose between *same* and *different.*

Weighted. We investigate *weighted* and *unweighted* graph CF techniques. *Unweighted* methods provide the best performance on almost all CP-fairness metrics, except for *APLT*, where *weighted* GAT-CF performs better.

Latent representations. We compare graph CF techniques with *64*, *128*, and *256* features. Higher latent representations (*128* and *256*) result in better performance on all measurements, with *128* being the turning point for stable performance (see Table 1 as a reference).

4. Trade-off Analysis

This section examines the trade-off, through Pareto optimal solutions, between accuracy, item exposure, and user fairness in graph CF baselines. The analysis is conducted on the Amazon Men dataset and focuses on the message-passing and weighting of graph edges. Three categories

Table 2
Best results (and corresponding graph CF model) for each <dimension, value> pair, on the Amazon Men dataset for top-20 lists. **Bold** indicates the best result in the pairs having a two-valued dimension, while † indicates the best results on *same* and *different* configurations. The symbols ↑ and ↓ indicate if high or low values are better. "*rank*" and "*rat*" stand for *UMADrank@k* and *UMADrat@k*.

Dimensions	Values	Overall Accuracy		Item Exposure			User Fairness	
		Recall↑	*nDCG*↑	*EFD*↑	*Gini*↑	*APLT*↑	*rank*↓	*rat*↓
Message passing	*implicit*	0.1222 (GFCF)	**0.0911 (GFCF)**	**0.2615 (GFCF)**	0.2871 (UltraGCN)	0.1808 (UltraGCN)	0.0123 (UltraGCN)	**0.0022 (UltraGCN)**
	explicit	**0.1223 (LR-GCCF)**	0.0884 (LR-GCCF)	0.2536 (LR-GCCF)	**0.5090 (LR-GCCF)**	**0.3823 (GAT-CF)**	**0.0002 (DGCF)**	0.0169 (LightGCN)
Explored nodes	*same-1*	0.1221† (LR-GCCF)	0.0884† (LR-GCCF)	0.2500† (LR-GCCF)	0.4377 (LR-GCCF)	0.3433 (GAT-CF)	**0.0002† (DGCF)**	**0.0022† (UltraGCN)**
	same-2	0.1184 (LightGCN)	0.0841 (LightGCN)	0.2380 (LightGCN)	**0.5090† (LR-GCCF)**	**0.3823† (GAT-CF)**	**0.0002† (DGCF)**	0.0209 (NGCF)
	different-1	**0.1222† (GFCF)**	**0.0911† (GFCF)**	**0.2615† (GFCF)**	0.4093 (NGCF)	0.3424 (GAT-CF)	**0.0002† (DGCF)**	**0.0022† (UltraGCN)**
	different-2	0.1210 (DGCF)	0.0850 (DGCF)	0.2407 (LightGCN)	0.4934† (LR-GCCF)	0.3438† (LR-GCCF)	**0.0002† (DGCF)**	0.0388 (LightGCN)
Weighting	*weighted*	0.1210 (DGCF)	0.0857 (DGCF)	0.2428 (DGCF)	0.3240 (DGCF)	**0.3823 (GAT-CF)**	**0.0002 (DGCF)**	0.0301 (DGCF)
	unweighted	**0.1223 (LR-GCCF)**	**0.0884 (LR-GCCF)**	**0.2536 (LR-GCCF)**	**0.5090 (LR-GCCF)**	0.3438 (LR-GCCF)	0.0101 (GCN-CF)	**0.0169 (LightGCN)**
Latent representations	*emb-64*	0.1193 (LR-GCCF)	0.0871 (LR-GCCF)	0.2479 (LR-GCCF)	**0.5090 (LR-GCCF)**	0.3627 (GAT-CF)	**0.0002 (DGCF)**	0.0054 (UltraGCN)
	emb-128	0.1221 (LR-GCCF)	0.0883 (LR-GCCF)	**0.2536 (LR-GCCF)**	**0.5090 (LR-GCCF)**	0.3644 (GAT-CF)	**0.0002 (DGCF)**	0.0111 (UltraGCN)
	emb-256	**0.1223 (LR-GCCF)**	**0.0884 (LR-GCCF)**	0.2532 (LR-GCCF)	0.5038 (LR-GCCF)	**0.3823 (GAT-CF)**	**0.0002 (DGCF)**	**0.0022 (UltraGCN)**

are studied: (1) implicit message-passing; (2) explicit message-passing with neighborhood weighting; (3) explicit message-passing without neighborhood weighting. Plots with extensive results are available in the extended work [1].

Accuracy/Item Exposure. Explicit/weighted models exhibit a trade-off, maximizing either accuracy or item exposure. Explicit/unweighted models show a balanced trade-off, not prioritizing a single goal. Implicit models prioritize accuracy, at the expense of niche item exposure.

Accuracy/User Fairness. No graph CF strategy emerges as an absolute winner. Every graph CF strategy is insufficient to guarantee adequate fairness among different user groups.

Item Exposure/User Fairness. Two groups of baselines are observed: those with poor item exposure and those with acceptable exposure for long-tail items. Explicit/unweighted strategies can generally ensure a satisfactory trade-off between user fairness and item exposure.

5. Conclusion and Future work

We assess the performance of graph CF models on Consumer and Producer (CP)-fairness metrics showing that their superior accuracy capabilities is reached at the expense of user fairness, item exposure, and their combination. By recognizing nodes representation and neighborhood exploration as the two main dimensions of a novel graph CF taxonomy, we study their influence on CP-fairness and overall accuracy separately and simultaneously. The outcomes raise concerns about the effective application of recent approaches in graph CF (e.g., implicit message-passing techniques). On such basis, we are performing further investigations on other datasets and algorithms, and we are working on new graph models balancing accuracy and CP-Fairness.

References

[1] V. W. Anelli, Y. Deldjoo, T. D. Noia, D. Malitesta, V. Paparella, C. Pomo, Auditing consumer- and producer-fairness in graph collaborative filtering, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I, volume 13980 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 33–48. URL: https://doi.org/10.1007/978-3-031-28244-7_3. doi:10.1007/978-3-031-28244-7_3.

[2] F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, 2022. URL: https://doi.org/10.1007/978-1-0716-2197-4. doi:10.1007/978-1-0716-2197-4.

[3] V. W. Anelli, Y. Deldjoo, T. D. Noia, A. Ferrara, F. Narducci, How to put users in control of their data in federated top-n recommendation with learning to rank, in: C. Hung, J. Hong, A. Bechini, E. Song (Eds.), SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021, ACM, 2021, pp. 1359–1362. URL: https://doi.org/10.1145/3412841.3442010. doi:10.1145/3412841.3442010.

[4] V. W. Anelli, T. D. Noia, P. Lops, E. D. Sciascio, Feature factorization for top-n recommendation: From item rating to features relevance, in: Y. Zheng, W. Pan, S. S. Sahebi, I. Fernández (Eds.), Proceedings of the 1st Workshop on Intelligent Recommender Systems by Knowledge Transfer & Learning co-located with ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 27, 2017, volume 1887 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 16–21. URL: https://ceur-ws.org/Vol-1887/paper3.pdf.

[5] V. W. Anelli, G. M. Biancofiore, A. D. Bellis, T. D. Noia, E. D. Sciascio, Interpretability of BERT latent space through knowledge graphs, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 3806–3810. URL: https://doi.org/10.1145/3511808.3557617. doi:10.1145/3511808.3557617.

[6] A. Ferrara, V. Anelli, A. Mancino, T. Di Noia, E. Di Sciascio, Kgflex: Efficient recommendation with sparse feature factorization and knowledge graphs, ACM Transaction on Recommender Systems 1 (2023). URL: https://doi.org/10.1145/3588901. doi:10.1145/3588901.

[7] V. W. Anelli, A. Bellogín, T. D. Noia, C. Pomo, Reenvisioning the comparison between neural collaborative filtering and matrix factorization, in: H. J. C. Pampín, M. A. Larson, M. C. Willemsen, J. A. Konstan, J. J. McAuley, J. Garcia-Gathright, B. Huurnink, E. Oldridge (Eds.), RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021, ACM, 2021, pp. 521–529. URL: https://doi.org/10.1145/3460231.3475944. doi:10.1145/3460231.3475944.

[8] G. Cornacchia, V. W. Anelli, G. M. Biancofiore, F. Narducci, C. Pomo, A. Ragone, E. D. Sciascio, Auditing fairness under unawareness through counterfactual reasoning, Inf. Process. Manag. 60 (2023) 103224. URL: https://doi.org/10.1016/j.ipm.2022.103224. doi:10.1016/j.ipm.2022.103224.

[9] H. Abdollahpouri, R. Burke, B. Mobasher, Controlling popularity bias in learning-to-rank recommendation, in: RecSys, ACM, 2017, pp. 42–46.
[10] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogín, T. D. Noia, A flexible framework for evaluating user and item fairness in recommender systems, User Model. User Adapt. Interact. 31 (2021) 457–511.
[11] V. Paparella, Pursuing optimal trade-off solutions in multi-objective recommender systems, in: RecSys, ACM, 2022, pp. 727–729.
[12] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR (Poster), OpenReview.net, 2017.
[13] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: ICLR (Poster), OpenReview.net, 2018.
[14] X. Wang, X. He, M. Wang, F. Feng, T. Chua, Neural graph collaborative filtering, in: SIGIR, ACM, 2019, pp. 165–174.
[15] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: SIGIR, ACM, 2020, pp. 639–648.
[16] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, T. Chua, Disentangled graph collaborative filtering, in: SIGIR, ACM, 2020, pp. 1001–1010.
[17] L. Chen, L. Wu, R. Hong, K. Zhang, M. Wang, Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach, in: AAAI, AAAI Press, 2020, pp. 27–34.
[18] K. Mao, J. Zhu, X. Xiao, B. Lu, Z. Wang, X. He, Ultragcn: Ultra simplification of graph convolutional networks for recommendation, in: CIKM, ACM, 2021, pp. 1253–1262.
[19] Y. Shen, Y. Wu, Y. Zhang, C. Shan, J. Zhang, K. B. Letaief, D. Li, How powerful is graph convolution for recommendation?, in: CIKM, ACM, 2021, pp. 1619–1629.