Comprehensive Assessment of Robustness in Fairness of GNN-based Recommender Systems against Attacks

Ludovico Boratto[1], Francesco Fabbri[2], Gianni Fenu[1], Mirko Marras[1] and Giacomo Medda[1,*]

[1]University of Cagliari, Cagliari, Italy
[2]Spotify, Barcelona, Spain

Abstract

The robustness of recommendation models is typically measured by their ability to maintain the original utility when exposed to attacks. In contrast, robustness in fairness pertains to the resilience of fairness levels in the presence of such attacks. Despite its significance, this latter area remains largely underexplored. In this extended abstract, we evaluate the robustness of graph-based recommender systems with respect to fairness from both the consumer and provider perspectives, under attacks involving edge-level perturbations. We analyze the impact of these perturbations on fairness through an experimental protocol involving three datasets and three graph neural networks. Our findings reveal severe fairness issues, particularly on the consumer side, where fairness is compromised to a greater extent than on the provider side. Source code: https://github.com/jackmedda/CPFairRobust.

Keywords

Robustness, Fairness, Recommendation, GNN, Perturbation, Multi-Stakeholder, Provider, Consumer.

1. Introduction

Recommender systems are designed to align with consumers' preferences by suggesting content that matches their interests while also achieving the visibility and engagement goals of content providers. However, this balance can be disrupted by targeted attacks that manipulate recommendations to serve the attacker's objectives, compromising the experience of both consumers and providers [1, 2]. Other research fields, such as computer sensing [3, 4] and code generation [5], have utilized attacks constructively to enhance performance or develop defense mechanisms. In a similar manner, the literature in recommender systems predominantly pursues these goals by focusing on attacks that disrupt model recommendations and, consequently, their utility. This often involves poisoning attacks, which alter training data by adding fake users.

Concerns over fairness in machine learning systems, driven by new regulatory frameworks [6], have intensified efforts to evaluate [7, 8, 9, 10, 11, 12], mitigate [13, 14], and explain [15, 16, 17, 18] fairness issues affecting both consumers and providers. Despite this growing focus, the integration of robustness and fairness objectives remains underexplored, and comprehensive studies on the robustness of fairness in recommender systems against specialized attacks are absent. While other domains have begun to address the interaction between attacks and

IIR2024: The 14th Italian Information Retrieval Workshop, 5th – 6th September 2024, Udine, Italy
*Corresponding author.
giacomo.medda@unica.it (G. Medda)

fairness [19], previous studies that have explored robustness beyond accuracy, such as certified robustness techniques and fairness attacks, have predominantly focused on classification tasks rather than recommender systems. In the recommendation domain, robustness beyond utility has been examined in relation to bias [20] and sparsity [21], yet fairness concerning protected groups remains largely unaddressed in the literature [1].

In this extended abstract, we summarize our prior work [22] on a comprehensive analysis of the robustness of graph-based recommender systems in terms of fairness, a concept we refer to as *robustness in fairness*. Specifically, we investigate the extent to which fairness, from both the consumer and provider perspectives, remains stable under attack scenarios. Given the strong performance of graph neural networks (GNNs) in recommendation tasks and the wide range of attacks targeting graph data [23], our study explores the impact of edge-level perturbations (addition and deletion) on robustness in fairness. We examine a white-box scenario where an attacker aims to compromise the group fairness of a recommender system. Such attacks may have real-world consequences, including damage to a company's public reputation [24].

2. Methodology

To test the robustness in fairness of GNN-based recommender systems, we extended an approach that perturbs a graph at the edge-level to explain the predictions in several downstream tasks [25, 26, 18]. The extended approach iteratively performs poisoning-like attacks and monitors fairness as the user-item interaction graph gets gradually perturbed, encompassing different types of perturbations and fairness operationalizations. To estimate the impact of such an attack on robustness in fairness of recommender systems, we follow [1] and define $(\gamma, \epsilon) - \textit{robustness}$. Given a user-item interaction graph represented by an adjacency matrix A, a fairness metric M, a GNN f parameterized by W, $(\gamma, \epsilon) - \textit{robustness}$ can be formalized as follows:

$$\Delta = M(f(\tilde{A}, W), A) - M(f(A, W), A), \quad \|\Delta\|_2^2 \leq \epsilon, \quad |\tilde{E}| \leq \gamma \tag{1}$$

where $\tilde{A}$ denotes the perturbed adjacency matrix and $\tilde{E}$ the set of candidate edges. A model is $(\gamma, \epsilon) - \textit{robust}$ if an attack bounded by a budget γ causes a change in fairness level lower than ϵ.

To define M, we draw on recent works that emphasize the importance of the group fairness notion of *demographic parity* from both the consumer [7, 27, 28] and provider perspectives [16, 9, 14]. We formalize M in a binary setting as the absolute difference in impact (e.g., utility for consumers [27] and exposure [29, 16] for providers) across two groups based on the related stakeholder. Specifically, recommendation utility for consumers was operationalized into two approximated metrics of M. Consumer Preference (CP) measures consumer fairness as the disparity across consumer groups in rank-aware top-k recommendation utility, assessed using NDCG@k. On the other hand, Consumer Satisfaction (CS) evaluates consumer fairness based on the disparity across consumer groups in rank-agnostic top-k recommendation utility, measured with Precision@k (P@k). Similarly, exposure objectives for providers were operationalized into two approximated metrics of M. Provider Exposure (PE) quantifies provider fairness as the disparity in rank-aware exposure [9, 29] across provider groups. On a different line, Provider Visibility (PV) assesses provider fairness as the disparity in rank-agnostic exposure, also referred to as visibility [29], across provider groups. For all these operationalizations, 0 indicates fairness and values close to it are better.

3. Experimental Evaluation

Given our objective of comprehensively testing robustness in fairness across intermediary perturbation stages, we do not perform a classic poisoning attack as defined in [1, 2], but we estimated Δ through the perturbed adjacency matrix $\tilde{A}$ generated at the inference stage.

Datasets. We opted for MovieLens-1M [30] (ML1M), Last.FM-1K [31] (LF1K), and Insurance [32] (INS). We assessed consumer fairness on gender and age groups, and provider fairness across short-head and long-tail items [16, 14]. Please refer to our full article [22] for more details.

Models. We selected consolidated baselines in graph collaborative filtering, namely GCMC [33], LightGCN (LGCN) [34], and NGCF [35], which cover different modern GNN architectures.

Results. Figure 1 reports the demographic parity estimates (y-axis, DP) for each fairness metric M (x-axis) under both edge addition (blue) and edge deletion (orange) from the original graph. The data points represent each iteration of the perturbation process, with point sizes increasing as more edges are perturbed (increased budget γ). A model robust in fairness would be represented by points close to the dashed line (original M value of the model without perturbation), while far points indicate that the perturbed edges affected the fairness level.

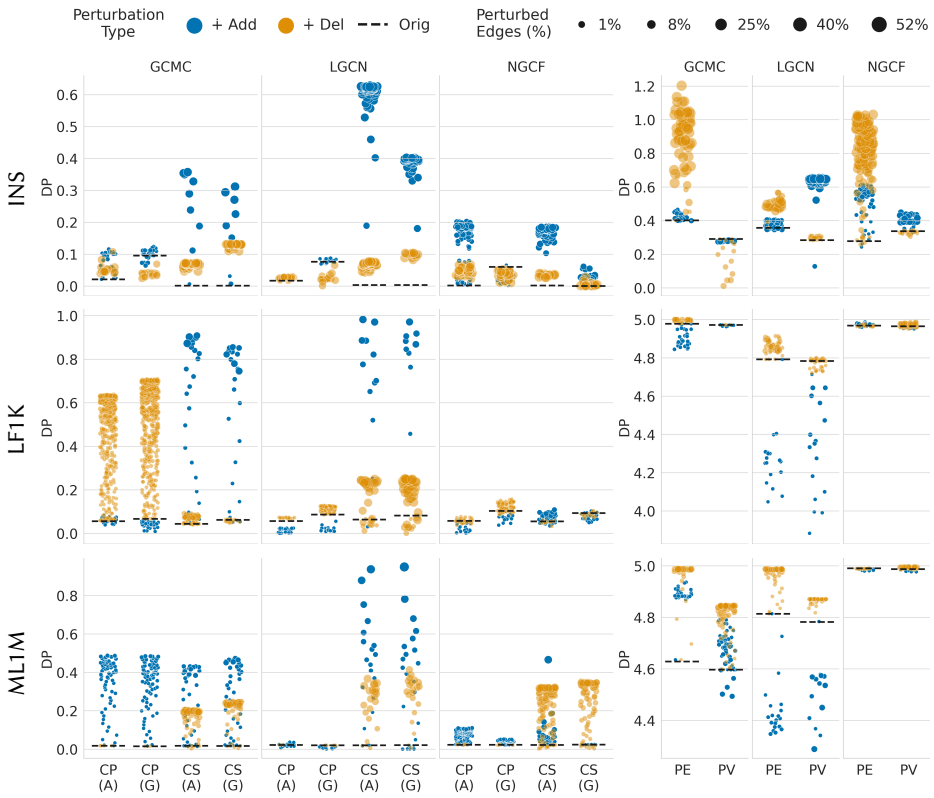

Figure 1: Demographic parity estimates for each fairness metric during edge addition (blue) and edge deletion (orange). Points, sized based on the extent of perturbation, indicate the impact of edge changes on fairness. Points near (far from) the dashed line suggest higher (lower) robustness in fairness.

On the consumer side, NGCF demonstrates the highest robustness, with points generally clustered closer to the dashed line compared to the other models. GCMC and LGCN exhibit similar behavior, particularly regarding Consumer Satisfaction (CS), where edge addition (+ *Add*) substantially impacts demographic parity (DP) more than edge deletion (+ *Del*). Some experiments suggest that dataset characteristics can influence attack outcomes. For example, GCMC was notably affected by edge deletion (+ *Del*) on Consumer Preference (CP) under LF1K, while NGCF's sensitivity to edge addition (+ *Add*) varies for age groups.

On the provider side, the original systems exhibit a high degree of unfairness, which in turn may influence the model robustness in fairness. Indeed, edge perturbations do not remarkably alter such robustness, as observed under LF1K for instance. In other scenarios, edge deletion proves more effective in influencing robustness in provider fairness compared to edge addition (+ *Add*), such as under INS. Notably, NGCF shows increased sensitivity to perturbations under INS, and the differences between GCMC and LGCN are more pronounced. This sensitivity may be due to the GCMC encoder interpreting added edges as noise, while LGCN's linear step views them as new information. Extended results can be found in our original study [22].

4. Conclusions and Future Works

In this extended abstract, we defined robustness in fairness and raised attention towards the issues caused by related attacks in recommendation. Compared to prior work, our analysis aimed to assess the robustness in fairness of GNN-based recommender systems against poisoning-like attacks based on edge-level perturbations, focusing on the models' robustness and not on the attack itself. From our results (in this extended abstract and in the original study [22]), the tested models exhibit a higher sensitivity to attacks tailored for consumer fairness compared with provider one. Specifically, the unfairness level across consumer groups can be increased by a restrained amount of perturbations, whereas the impact on provider fairness is limited by the prior unfairness level. Despite the limited set of considered models, they represent consistent baselines in the literature and good candidates for analyzing a topic still unexplored in recommendation compared with other fields. In future works, we plan to cover a wider set of models, investigate perturbations based on re-wiring, and explore grey- or black-box attacks.

Acknowledgments

We acknowledge financial support from (i) the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.1 - Call for tender No. 3277, published on December 30, 2021, by the Italian Ministry of University and Research (MUR), funded by the European Union – Next Generation EU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – Grant Assignment Decree No. 1056 adopted on June 23, 2022, by the MUR (CUP F53C22000430001) and (ii) the project PHaSE - Promoting Healthy and Sustainable Eating through Interactive and Explainable AI Methods, funded by the MUR under the PRIN 2022 program (CUP H53D23003530006).

References

[1] K. Zhang, Q. Cao, F. Sun, Y. Wu, S. Tao, H. Shen, X. Cheng, Robust recommender system: A survey and future directions, CoRR abs/2309.02057 (2023). arXiv:2309.02057.

[2] V. W. Anelli, Y. Deldjoo, T. D. Noia, F. A. Merra, Adversarial recommender systems: Attack, defense, and advances, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, 2022, pp. 335–379.

[3] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: Proc. of the 37th International Conference on Machine Learning, ICML, volume 119, PMLR, 2020, pp. 2206–2216.

[4] F. Croce, S. Gowal, T. Brunner, E. Shelhamer, M. Hein, A. T. Cemgil, Evaluating the adversarial robustness of adaptive test-time defenses, in: Proc. of the International Conference on Machine Learning, ICML, volume 162, PMLR, 2022, pp. 4421–4435.

[5] A. Mastropaolo, L. Pascarella, E. Guglielmi, M. Ciniselli, S. Scalabrino, R. Oliveto, G. Bavota, On the robustness of code generation techniques: An empirical study on github copilot, in: Proc. of the 45th IEEE/ACM International Conference on Software Engineering, ICSE, IEEE, 2023, pp. 2149–2160.

[6] T. D. Noia, N. Tintarev, P. Fatourou, M. Schedl, Recommender systems under european AI regulations, Commun. ACM 65 (2022) 69–73.

[7] L. Boratto, G. Fenu, M. Marras, G. Medda, Consumer fairness in recommender systems: Contextualizing definitions and mitigations, in: Proc. of the 44th European Conference on IR Research, ECIR, volume 13185 of *LNCS*, Springer, 2022, pp. 552–566.

[8] L. Boratto, G. Fenu, M. Marras, G. Medda, Practical perspectives of consumer fairness in recommendation, Inf. Process. Manag. 60 (2023) 103208.

[9] A. Singh, T. Joachims, Fairness of exposure in rankings, in: Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD, ACM, 2018, pp. 2219–2228.

[10] G. Fenu, M. Marras, G. Medda, G. Meloni, Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition, in: Proc. of the 22nd Annual Conference of the International Speech Communication Association, Interspeech, ISCA, 2021, pp. 1892–1896.

[11] A. Atzori, G. Fenu, M. Marras, Demographic bias in low-resolution deep face recognition in the wild, IEEE J. Sel. Top. Signal Process. 17 (2023) 599–611.

[12] G. Balloccu, L. Boratto, C. Cancedda, G. Fenu, M. Marras, Knowledge is power, understanding is impact: Utility and beyond goals, explanation quality, and fairness in path reasoning recommendation, in: Proc. of the 44th European Conference on IR Research, ECIR, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 3–19.

[13] L. Boratto, F. Fabbri, G. Fenu, M. Marras, G. Medda, Counterfactual graph augmentation for consumer unfairness mitigation in recommender systems, in: Proc. of the 32nd ACM International Conference on Information and Knowledge Management, CIKM, ACM, 2023, pp. 3753–3757.

[14] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, Y. Zhang, Towards long-term fairness in recommendation, in: Proc. of the Fourteenth ACM International Conference on Web Search and Data Mining, WSDM, ACM, 2021, pp. 445–453.

[15] A. Atzori, G. Fenu, M. Marras, Explaining bias in deep face recognition via image characteristics, in: Proc. of the IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2022, pp. 1–10.

[16] Y. Ge, J. Tan, Y. Zhu, Y. Xia, J. Luo, S. Liu, Z. Fu, S. Geng, Z. Li, Y. Zhang, Explainable fairness in recommendation, in: Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, ACM, 2022, pp. 681–691.

[17] A. Ghazimatin, O. Balalau, R. S. Roy, G. Weikum, PRINCE: provider-side interpretability with counterfactual explanations in recommender systems, in: Proc. of the Thirteenth ACM International Conference on Web Search and Data Mining, WSDM, ACM, 2020, pp. 196–204.

[18] G. Medda, F. Fabbri, M. Marras, L. Boratto, G. Fenu, Gnnuers: Fairness explanation in gnns for recommendation via counterfactual reasoning, ACM Trans. Intell. Syst. Technol. (2024). Just Accepted.

[19] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. (2022).

[20] M. Sato, S. Takemori, J. Singh, T. Ohkuma, Unbiased learning for the causal effect of recommendation, in: Proc. of the Fourteenth ACM Conference on Recommender Systems, RecSys, ACM, 2020, pp. 378–387.

[21] J. Zheng, Q. Ma, H. Gu, Z. Zheng, Multi-view denoising graph auto-encoders on heterogeneous information networks for cold-start recommendation, in: Proc. of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD, ACM, 2021, pp. 2338–2348.

[22] L. Boratto, F. Fabbri, G. Fenu, M. Marras, G. Medda, Robustness in fairness against edge-level perturbations in gnn-based recommendation, in: Proc. of the 46th European Conference on Information Retrieval, ECIR, volume 14610 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 38–55.

[23] M. Fang, G. Yang, N. Z. Gong, J. Liu, Poisoning attacks to graph-based recommender systems, in: Proc. of the 34th Annual Computer Security Applications Conference, ACSAC, ACM, 2018, pp. 381–392.

[24] N. Mehrabi, M. Naveed, F. Morstatter, A. Galstyan, Exacerbating algorithmic bias through fairness attacks, in: Proc. of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI , Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI, AAAI Press, 2021, pp. 8930–8938.

[25] A. Lucic, M. A. ter Hoeve, G. Tolomei, M. de Rijke, F. Silvestri, Cf-gnnexplainer: Counterfactual explanations for graph neural networks, in: Proc. of the International Conference on Artificial Intelligence and Statistics, AISTATS, volume 151, PMLR, 2022, pp. 4499–4511.

[26] B. Kang, J. Lijffijt, T. D. Bie, Explanations for network embedding-based link predictions, in: Proc. of the International Workshops of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD, volume 1524, Springer, 2021, pp. 473–488.

[27] H. Wu, C. Ma, B. Mitra, F. Diaz, X. Liu, A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation, ACM Trans. Inf. Syst. (2022). Just Accepted.

[28] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: Proc. of the Web Conference, TheWebConf, ACM / IW3C2, 2021, pp. 624–632.
[29] E. Gómez, C. S. Zhang, L. Boratto, M. Salamó, M. Marras, The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems, in: Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, ACM, 2021, pp. 1808–1812.
[30] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 19:1–19:19.
[31] Ò. Celma, Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space, Springer, 2010.
[32] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, in: Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, ACM, 2021, pp. 1054–1063.
[33] R. V. den Berg", T. N. Kipf, M. Welling, Graph convolutional matrix completion, CoRR abs/1706.02263 (2017). arXiv:1706.02263.
[34] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: Proc. of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR, ACM, 2020, pp. 639–648.
[35] X. Wang, X. He, M. Wang, F. Feng, T. Chua, Neural graph collaborative filtering, in: Proc. of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, ACM, 2019.