What are we missing in algorithmic fairness? Discussing open challenges for fairness analysis in user profiling with Graph Neural Networks

Erasmo Purificato[1,2][0000-0002-5506-3020] and Ernesto William De Luca[1,2][0000-0003-3621-4118]

[1] Otto von Guericke University Magdeburg, Germany
{erasmo.purificato,ernesto.deluca}@ovgu.de
[2] Leibniz Institute for Educational Media | Georg Eckert Institute, Germany
{erasmo.purificato,deluca}@gei.de

Abstract. Due to the rising importance of human-centred perspectives in artificial intelligence and all related fields, *algorithmic fairness* is currently a key topic when dealing with research on novel machine learning models and applications. However, in most cases, in the context of fairness analysis, we are commonly facing situations in which the fairness metrics are applied only in binary classification scenarios, and the capability of a model to produce fair results is evaluated considering the *absolute difference* of the scores of the two sensitive groups considered. In this paper, we aim to discuss these two open challenges and illustrate our position from an ethical perspective. To support our arguments, we present a case study on two recent scientific contributions exploiting Graph Neural Networks models for user profiling, which are considered state-of-the-art technologies in many domains. With the presented work, our goal is also to create a valuable debate in the community about the raised questions.

Keywords: Algorithmic Fairness · AI Ethics · Graph Neural Networks

1 Background and Motivation

As the use of automated decision-making systems has massively increased lately, **algorithmic fairness** [18, 21] has become a crucial research topic, mainly due to the social impact such systems are having on people's life. There is a significant amount of literature on methods to detect and address bias in machine learning (ML) and deep learning (DL) models [2, 4, 30], notably in user-related scenarios [24], information retrieval (IR) [9, 11, 26, 28] and recommendation systems [12, 19, 27]. A number of studies have also been conducted to figure out the potential roots of unfairness in automated systems [20, 22], which are commonly identified in two main categories: (1) biased *data* and (2) *algorithms* receptive to the biases already present in the datasets used for training.

Among the most powerful technologies falling in the latter category, there are **Graph Neural Networks** (GNNs) [13, 17, 29, 33, 34], recently emerged as an effective solution for dealing with graph data structures in many domains, such as recommenders [15], natural language processing [32], and user profiling [5, 31]. Like any ML system, GNNs are susceptible to learning biases from the historical data they are trained on, and this can manifest in their output. This is primarily due to the unique structure of graphs and the GNNs' message-passing procedure, which can exacerbate discrimination as nodes with similar sensitive attributes are more likely to be connected to each other than those with different attributes [25]. In the last couple of years, several works have been published about the analysis and evaluation of fairness in GNNs [1, 6, 7, 20, 23]. Most of them (especially all those cited) show, in their fairness assessment, two crucial characteristics we aim to highlight and argue in this position paper from an ethical perspective:

1. the fairness metrics are applied in classification scenarios where both the target class and the sensitive attribute (e.g. gender, age, race) are *binary*;
2. the capability of a model to produce fair results is evaluated considering the *absolute difference* of the scores of the two sensitive groups considered.

It is worth noting that these aspects are not specific to the fairness analysis of GNN-based models, but they reflect broader issues in bias detection studies for general automated decision-making systems.

To address the open challenges, in the rest of this paper, we first focus on two publications related to GNN-based models for user profiling (i.e. [6, 23]) in order to present the two publications and illustrate how the fairness analysis has been performed in both cases. Finally, we present the results of the experiments carried out on the two analysed contributions to concretely discuss our position. In particular, the case study presented in Section 3 aims to provide quantitative motivations to the above challenges by running two types of analysis on the considered models, re-adapting the experiments conducted in the original publications. In the first one, we focus on the use of the *absolute difference* of the computed fairness metrics, while in the second one, we consider a specific combination of model and dataset in [23] and run the experiment with the original multiclass distribution of the sensitive attribute investigated.

One of the main purposes of the proposed case study is to create a valuable debate in the community about the raised questions.

2 Analysed Contributions

The scientific works we selected for our case study (Section 3) to examine and discuss the posed open challenges are illustrated below and belong to the field of **user profiling**, which primarily aims to generate an efficient user representation, namely a *user model* by gleaning individuals' personal characteristics [16].

Dai and Wang [6] proposed *FairGNN*, a novel framework for fair node classification that employs an adversarial debiasing mechanism for dealing with the shortage of sensitive attributes and producing unbiased results. The authors conducted the experiments in a common binary classification scenario on three

Table 1. Fairness metrics computation without absolute value for Dai and Wang [6] (in particular, we exploited the *FairGCN* version).

Dataset	Δ_{SP}	Δ_{EO}
Pokec-z	0.024 ±0.007	0.012 ±0.003
NBA	-0.021 ±0.007	0.018 ±0.001

Table 2. Fairness metrics computation without absolute value for Purificato et al. [23].

Dataset	**Model**	Δ_{SP}	Δ_{EO}
Alibaba	CatGCN	-0.045 ±0.021	0.139 ±0.074
	RHGN	0.019 ±0.012	-0.133 ±0.086
JD	CatGCN	0.033 ±0.013	-0.052 ±0.016
	RHGN	0.009 ±0.007	-0.042 ±0.017

different datasets[3] and adopted two standard fairness metrics in their analysis: *statistical parity* [8, 10] and *equal opportunity* [14]. For both metrics, they quantitatively evaluated the absolute difference of the probabilities computed for the single sensitive attributes, reported as Δ_{SP} and Δ_{EO}, respectively.

In one of our previous works (hereinafter formally referred to as Purificato et al.) [23], we presented the fairness assessment of two state-of-the-art GNN-based models for user profiling, i.e. *CatGCN* [5] and *RHGN* [31] on two real-world use cases, in order to derive potential correlations between the different profiling paradigms of the analysed architectures and the fairness scores they produce. The authors considered a binary scenario performing a fairness analysis on two datasets and leveraging four metrics: *statistical parity*, *equal opportunity*, *overall accuracy equality* [3] and *disparate mistreatment* [3]. Similar to the previous work, the evaluation is made by exploiting the absolute difference of the probabilities computed for the single sensitive attributes, namely Δ_{SP}, Δ_{EO}, Δ_{OAE} and Δ_{TE}.

3 Case study

We run two types of experiments for the open challenges presented in Section 1. In the first one, we focused on the use of the *absolute difference* of the computed fairness metrics. The setting is straightforward: we remove the absolute value from the fairness computation of the analysed models and execute the same experiments presented in the original papers with the default parameters, computing Δ_{SP} and Δ_{EO}. The results are displayed in Table 1 and Table 2. In the results, it is evident the alternation of positive and negative scores, meaning that for a given combination of model and dataset, the unfairness (regardless of the specific value) might be directed towards one sensitive group or the other.

Concerning the issue related to fairness analysis in binary scenarios, we conducted an experiment only for a specific model and dataset, because the derived

[3] Due to the page limit constraint, the details of the experiments carried out in the original paper are not discussed.

Table 3. Statistical parity scores for binary and multiclass sensitive attribute groups for Purificato et al. [23] (RHGN model and Alibaba dataset).

Binary group	***SP***	Multiclass group	***SP***
A	0.887 ±0.015	s_0	0.81 ±0.02
		s_1	0.91 ±0.02
		s_2	0.91 ±0.01
		s_3	0.92 ±0.01
B	0.797 ±0.055	s_4	0.89 ±0.01
		s_5	0.72 ±0.03
		s_6	0.78 ±0.07

implications can be easily extended. In particular, we focused on RHGN model and Alibaba dataset from Purificato et al. [23] work, adopting the original binary classification task, but with the following setting for the sensitive attribute: on the one hand, we considered its original multiclass distribution (seven groups, named as s_0-s_6) and calculated every single *statistical parity* (SP) probability; on the other hand, we binarised the attribute, as done in the original paper, and again computed the single probabilities for the binary groups. The resulting binary sensitive attribute groups are composed as follows: $A = \{s_0, s_1, s_2, s_3\}$, $B = \{s_4, s_5, s_6\}$. The results are shown in Table 3.

The observation derived from these results is that binarisation can lead to misleading evaluation of a specific subgroup. In this specific experiment, the group s_0 should be treated as a disadvantaged group if considered in the fine-grained assessment, but it would be treated as an advantaged group when included in the binary group A. The opposite applies to group s_4.

4 Ethical Implications of the Open Challenges

From an ethical perspective, there are several implications from the presented results which led us to argue the following positions regarding the challenges we open with this paper:

1. In many of the current works about fairness evaluation of automated systems, the sensitive attributes, that are natively multiclass, are made binary to meet the standard fairness metrics definitions. From our point of view, there are two crucial reasons why it is essential to evaluate fairness by examining the actual distribution of sensitive groups. Firstly, if the system at hand is not as effective for certain groups, they will end up receiving less effective services, such as targeted advertisements or recommendations. Secondly, reducing the different classes and groups into a binary representation can lead to an incorrect evaluation of the fairness of models, potentially distorting the original data conditions.
2. In the same context, considering the absolute difference score in the fairness analysis can be hazardous for other motivations. In particular, from both a system and user perspective, with this practice, we cannot figure out the

disadvantaged groups for every specific combination of model, dataset and fairness metrics, and thus unable to make in place any tailored intervention to mitigate the issue in a real-world scenario.

5 Conclusion

In this paper, we posed and discussed two potential open challenges in recent studies on algorithmic fairness, namely the common practices of performing the assessment only in classification scenarios where both the target class and the sensitive attribute are binary, and the use of the absolute difference of the fairness metrics scores in the evaluation to deem a model as fair or not. With a case study on GNN-based models for user profiling, we presented our position arguing in favour of a multiclass assessment with a clear understanding of the disadvantaged groups, exposing also some ethical implications which derive from the experimental results displayed. Our aim is to foster discussion in the community around these topics and continue to deepen into them with even more detailed future analysis.

References

1. Agarwal, C., Lakkaraju, H., Zitnik, M.: Towards a unified framework for fair and stable graph representation learning. In: Uncertainty in Artificial Intelligence. pp. 2114–2124. PMLR (2021)
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021)
4. Caton, S., Haas, C.: Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053 (2020)
5. Chen, W., Feng, F., Wang, Q., He, X., Song, C., Ling, G., Zhang, Y.: Catgcn: Graph convolutional networks with categorical node features. IEEE Transactions on Knowledge and Data Engineering (2021)
6. Dai, E., Wang, S.: Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In: Proceed. of the 14th ACM International Conference on Web Search and Data Mining. pp. 680–688 (2021)
7. Dong, Y., Kang, J., Tong, H., Li, J.: Individual fairness for graph neural networks: A ranking based approach. In: Proceed. of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 300–310 (2021)
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
9. Ekstrand, M.D., Das, A., Burke, R., Diaz, F., et al.: Fairness in information access systems. Foundations and Trends® in Information Retrieval **16**(1-2), 1–177 (2022)
10. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268 (2015)

11. Gao, R., Shah, C.: How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In: Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval. pp. 229–236 (2019)
12. Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Ramos, G.: Enabling cross-continent provider fairness in educational recommender systems. Future Gener. Comput. Syst. **127**, 435–447 (2022). https://doi.org/10.1016/j.future.2021.08.025
13. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)
14. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
15. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 639–648 (2020)
16. Kanoje, S., Girase, S., Mukhopadhyay, D.: User profiling trends, techniques and applications. arXiv preprint arXiv:1503.07474 (2015)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings (2017)
18. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. In: Aea papers and proceedings. vol. 108, pp. 22–27 (2018)
19. Leonhardt, J., Anand, A., Khosla, M.: User fairness in recommender systems. In: Companion Proc. of The Web Conference 2018. pp. 101–102 (2018)
20. Loveland, D., Pan, J., Bhathena, A.F., Lu, Y.: Fairedit: Preserving fairness in graph neural networks through greedy graph editing. arXiv preprint arXiv:2201.03681 (2022)
21. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application **8**, 141–163 (2021)
22. Pessach, D., Shmueli, E.: Algorithmic fairness. arXiv preprint arXiv:2001.09784 (2020)
23. Purificato, E., Boratto, L., De Luca, E.W.: 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. pp. 4399–4403 (2022)
24. Purificato, E., Lorenzo, F., Fallucchi, F., De Luca, E.W.: The use of responsible artificial intelligence techniques in the context of loan approval processes. International Journal of Human–Computer Interaction pp. 1–20 (2022)
25. Rahman, T., Surma, B., Backes, M., Zhang, Y.: Fairwalk: towards fair graph embedding. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 3289–3295 (2019)
26. Ramos, G., Boratto, L.: Reputation (in)dependence in ranking systems: Demographics influence over output disparities. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020. pp. 2061–2064. ACM (2020). https://doi.org/10.1145/3397271.3401278
27. Ramos, G., Boratto, L., Caleiro, C.: On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. Inf. Process. Manag. **57**(2), 102058 (2020). https://doi.org/10.1016/j.ipm.2019.102058
28. Saúde, J., Ramos, G., Boratto, L., Caleiro, C.: A robust reputation-based group ranking system and its resistance to bribery. ACM Trans. Knowl. Discov. Data **16**(2), 26:1–26:35 (2022). https://doi.org/10.1145/3462210

29. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
30. Verma, S., Rubin, J.: Fairness definitions explained. In: IEEE/ACM International Workshop on Software Fairness (FairWare 2018). pp. 1–7. IEEE (2018)
31. Yan, Q., Zhang, Y., Liu, Q., Wu, S., Wang, L.: Relation-aware heterogeneous graph for user profiling. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3573–3577. Association for Computing Machinery, New York, NY, USA (Oct 2021)
32. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7370–7377 (2019)
33. Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803 (2019)
34. Zhang, Z., Cui, P., Zhu, W.: Deep learning on graphs: A survey. IEEE Transactions on Knowledge and Data Engineering **34**(1), 249–270 (2022)