Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact and Mistreatment in Behavioural User Profiling

Erasmo Purificato
Otto-von-Guericke-Universität
Magdeburg, Germany
Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut
Brunswick, Germany
erasmo.purificato@ovgu.de

Ludovico Boratto
Department of Mathematics and Computer Science,
University of Cagliari
Cagliari, Italy
ludovico.boratto@acm.org

Ernesto William De Luca
Otto-von-Guericke-Universität
Magdeburg, Germany
Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut
Brunswick, Germany
ernesto.deluca@ovgu.de

ABSTRACT

Recent approaches to behavioural user profiling employ Graph Neural Networks (GNNs) to turn users' interactions with a platform into actionable knowledge. The effectiveness of an approach is usually assessed with accuracy-based perspectives, where the capability to predict user features (such as gender or age) is evaluated. In this work, we perform a *beyond-accuracy* analysis of the state-of-the-art approaches to assess the presence of disparate impact and disparate mistreatment, meaning that users characterised by a given sensitive feature are unintentionally, but systematically, classified worse than their counterparts. Our analysis on two real-world datasets shows that different user profiling paradigms can impact fairness results. The source code and the preprocessed datasets are available at: https://github.com/erasmopurif/do_gnns_build_fair_models.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Social and professional topics** → **User characteristics**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Fairness, User Models, User Profiling, Graph Neural Networks

ACM Reference Format:
Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 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 and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557584

1 INTRODUCTION

In recent years, due to the huge amount of data provided by web applications and platforms, *user profiling* has become a key topic in many real-world scenarios, mainly social networks [23] and e-commerce [32]. The main goal of user profiling is to infer an individual's interests, personality traits, or behaviours from generated data to create an efficient user representation, i.e. a *user model*, which is exploited by adaptive and personalised systems [12]. Early profiling approaches considered only the analysis of static characteristics (*explicit user profiling*), with data often coming from online forms and surveys [24]. However, these methods have been proven ineffective as users are not concerned about providing their information directly. Therefore, modern systems focus more on profiling users' data implicitly based on individuals' actions and interactions (*implicit user profiling*). This approach is also referred to as **behavioural user profiling** [18].

A natural way to model these behaviours is through graphs, where edges describe the interactions between users, represented by nodes. Graph Neural Networks (GNNs) [15, 19, 29, 37, 38] have demonstrated to be effective in modelling graph data in several domains, such as recommender systems [17, 35], natural language processing [34], text mining [28] and user profiling [6, 7, 26, 33].

State of the art. Li et al. [21] put the first steps towards user profiling on graph data, leveraging a heterogeneous graph built upon "following" and "tweeting" interactions to infer users' location. Rahimi et al. [26] proposed a geolocation model based on Graph Convolutional Networks (GCNs), which makes use of text and network information to detect users' location. Chen et al. [7] introduced a Heterogeneous Graph Attention Network (HGAT) for learning user representations considering the graph structure and the attention mechanism to discern the importance of each node's neighbour. The most recent and promising works in this field were published last year. Chen et al. [6] introduced a GCN-based model which shows the benefits of enhancing the node representation before performing user profiling tasks. Yan et al. [33] proposed a Heterogeneous Graph Network (HGN) to improve prediction performances by considering multiple types of relations and entities for user profiling, in contrast to previous works only based on single types. Generally, existing approaches evaluate user profiling models based on the effectiveness of a classification task at predicting a user's personal characteristics, such as gender or age [7].

Motivation. Despite the success in classifying user profiles, as any machine learning (ML) system trained on historical data, GNNs are prone to learn biases in such data and reveal them in their output. This is mainly due to the topology of graph structures and the typical message-passing process of GNNs, which can amplify discrimination as nodes of the same sensitive attribute are more likely to be linked to each other than those different [27]. *Algorithmic fairness* has recently emerged as a crucial topic alongside the increasing use of automated decision-making systems. Considerable literature has been produced on general methods to detect and

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557584

mitigate bias in ML models [1, 5, 30] and user-related scenarios [25], especially recommenders [11, 20, 22]. Only a few works have been published to evaluate fairness on GNNs (e.g. [8, 9]), but to the best of our knowledge, none of them assesses potential discrimination in state-of-the-art GNN-based models for user profiling tasks.

Unfairness practices hidden in this class of models can be hazardous. Indeed, on one hand, by focusing only on behavioural data, they do not create intentional unfairness (disparate treatment [36]). On the other hand, if modelling is more effective for specific demographic groups, they would inevitably systematically receive less effective services. To provide an example, considering the attributes exploited in this work, when a system provides systematically worse *gender* predictions for a specific *age* group, this group will systematically get worse service (e.g. ads targeted for the opposite gender). Hence, characterising unfairness in behavioural user profiling models becomes a crucial problem in this domain.

Our contributions. In this paper, we aim to assess fairness in behavioural user profiling tasks for GNN-based models first in terms of **disparate impact** [2, 31] and then we expand the evaluation to consider **disparate mistreatment** [36] perspective. Our fairness analysis is conducted on two state-of-the-art GNNs (i.e. **CatGCN** [6] and **RHGN** [33]), proven to be the most effective in user profiling. Our contributions can be summarised as follows:

- we perform two user profiling tasks by executing a binary classification on two real-world datasets (Sec. 3.1) by leveraging the most performing GNNs in this context (Sec. 2.1);
- we assess disparate impact and disparate mistreatment (Sec. 3) for GNNs designed for behavioural user profiling, considering four algorithmic fairness metrics (Sec. 2.2);
- from the results of an extensive set of experiments (Sec. 3.3), we derive three observations about the analysed models, correlating their different user profiling paradigms with the fairness metrics scores to create a baseline for future assessment considering GNN-based models for user profiling.

2 PRELIMINARIES

In this section, we first describe the state-of-the-art GNN models analysed in our work. Then we introduce the metrics adopted for fairness assessment, including their mathematical definition.

2.1 Models' description

The analysis performed in this work considers two GNN-based models published in the last year that represent the state of the art in user profiling tasks, i.e. **CatGCN** and **RHGN**.

CatGCN [6] is a Graph Convolutional Network (GCN) model tailored for graph learning on categorical node features. This model enhances the initial node representation by integrating two types of explicit interaction modelling into its learning process: a local multiplication-based interaction on each pair of node features and a global addition-based interaction on an artificial feature graph. The proposed method shows the effectiveness of performing feature interaction modelling before graph convolution.

RHGN [33] is a Relation-aware Heterogeneous Graph Network designed to model multiple relations on a heterogeneous graph between different kinds of entities. The core parts of this model are a transformer-like multi-relation attention, used to learn the node importance and uncover the meta-relation significance on the graph, and a heterogeneous graph propagation network employed to gather information from multiple sources. This approach outperforms several GNN-based models on user profiling tasks.

2.2 Metrics

We define the fairness metrics adopted in our work considering $y \in \{0, 1\}$ as the binary target label and $\hat{y} \in \{0, 1\}$ as the prediction of the user profiling model $f : x \rightarrow y$. The sensitive attribute is denoted with $s \in \{0, 1\}$. In the metrics' descriptions, we also exploit the following notation which relates to classification properties: TP, FP, TN and FN, denoting *true positives*, *false positives*, *true negatives* and *false negatives*, respectively.

Our focus in this paper is the assessment of the fairness of the GNNs introduced in the previous section in terms of **disparate impact**. Also known as *adverse impact*, it refers to a form of indirect and often unintentional discrimination that occurs when practices or systems seem to apparently treat people the same way [14]. It concerns with situations where the model disproportionately discriminates against certain groups, even if the model does not explicitly employ the sensitive attribute to make predictions but rather on some proxy attributes [31]. This is exactly what happens in the analysed GNNs, where the user models are created by aggregating information from neighbours, and the sensitive attribute is not explicitly taken into consideration during classification. The notion of disparate impact is beneficial when there is not a clear linkage in training data between the predicted label and the sensitive attribute, i.e. it is hard to define the validity of a decision for a group member based on the historical data [36].

As reported in several works on fairness in ML, such as [8, 36], we evaluate the disparate impact value of the analysed models through *statistical parity* and *equal opportunity* metrics.

Statistical parity (or *demographic parity*) [10, 13] defines fairness as an equal probability for each group of being assigned to the positive class, i.e. predictions independent with sensitive attributes.

$$P(\hat{y} = 1|s = 0) = P(\hat{y} = 1|s = 1) \tag{1}$$

Equal opportunity [16] requires the probability of a subject in a positive class to be classified with the positive outcome should be equal for each group, i.e. TP should be the same across groups.

$$P(\hat{y} = 1|y = 1, s = 0) = P(\hat{y} = 1|y = 1, s = 1) \tag{2}$$

To extend the disparate impact evaluation conducted in previous works, we measure the *overall accuracy equality* metric to consider both TP and TN and look at relative accuracy across the groups.

Overall accuracy equality [3] defines fairness as the equal probability of a subject from either positive or negative class to be assigned to its respective class, i.e. each group should have the same prediction accuracy.

$$\begin{aligned} &P(\hat{y} = 0|y = 0, s = 0) + P(\hat{y} = 1|y = 1, s = 0) = \\ &= P(\hat{y} = 0|y = 0, s = 1) + P(\hat{y} = 1|y = 1, s = 1) \end{aligned} \tag{3}$$

In a scenario where it is hard to define the correctness of a prediction related to sensitive attribute values, we argue that a complete fairness assessment should always include the perspective

of **disparate mistreatment**. This concept considers the *misclassification rates* for user groups having different values of the sensitive attribute, instead of considering the corrected predictions [36]. Furthermore, the notion of disparate mistreatment is significant in contexts where misclassification costs depend on the group affected by the error. We select the *treatment equality* metric to evaluate this fairness perspective.

Treatment equality [3] requires the ratio of errors made by the classifier to be equal across different groups, i.e. each group should have the same ratio of *false negatives* (FN) and *false positives* (FP).

$$\frac{P(\hat{y}=1|y=0,s=0)}{P(\hat{y}=0|y=1,s=0)} = \frac{P(\hat{y}=1|y=0,s=1)}{P(\hat{y}=0|y=1,s=1)} \tag{4}$$

According to [4] and [8], to quantitatively evaluate the disparate impact and disparate mistreatment of the analysed models, we operationalise the metrics defined by Eqs. (1)-(4) as follows:

$$\Delta_{SP} = |P(\hat{y}=1|s=0) - P(\hat{y}=1|s=1)|, \tag{5}$$

$$\Delta_{EO} = |P(\hat{y}=1|y=1,s=0) - P(\hat{y}=1|y=1,s=1)|, \tag{6}$$

$$\Delta_{OAE} = |P(\hat{y}=0|y=0,s=0) + P(\hat{y}=1|y=1,s=0) - \\ - P(\hat{y}=0|y=0,s=1) + P(\hat{y}=1|y=1,s=1)|, \tag{7}$$

$$\Delta_{TE} = \left|\frac{P(\hat{y}=1|y=0,s=0)}{P(\hat{y}=0|y=1,s=0)} - \frac{P(\hat{y}=1|y=0,s=1)}{P(\hat{y}=0|y=1,s=1)}\right| \tag{8}$$

3 FAIRNESS ASSESSMENT

To carry out the fairness assessment, we conduct extensive empirical studies to investigate the following research questions:

- **RQ1** How do the different input types of the analysed GNNs and the way the user models are constructed affect fairness?
- **RQ2** To what extent can the user models produced by the analysed state-of-the-art GNNs be defined as *fair*?
- **RQ3** Are disparate impact metrics enough to assess the fairness of GNN-based models in behavioural user profiling tasks or is disparate mistreatment needed to fully assess the presence of unfairness?

We describe below the datasets used in our work, the different experiments executed to answer each research question, and the parameters chosen to set the models before fairness. To conclude the section, we present the results of the assessment.

3.1 Datasets

We choose two public real-world user profiling datasets, namely **Alibaba** and **JD**, from the two popular e-commerce platforms.

Alibaba dataset[1] contains click-through rates data about ads displayed on Alibaba's Taobao platform, and has been adopted in both [6] and [33] for evaluation. According to [6], for CatGCN model, we select the categories of products as the categorical features related to user nodes. In particular, we only consider items clicked at least by two users to establish a *co-click* relationship used as the model's local interaction. We apply the same filtering process to RHGN model to make the datasets consistent before creating the heterogeneous graph. For our experiments, we consider the users' *gender* as the user profiling task label, and their *age* as the sensitive attribute for fairness evaluation. Since the metrics we are focusing on work with binary attributes, we split this feature in two groups (*bin-age*) defining a clear separation between the two groups. In the Alibaba dataset, the age range of each class is not specified and is only characterised by a label.

Table 1: Dataset characteristics

Dataset	Users	Items	Edges	Features
Alibaba	166 958	64 553	427 464	2 820
JD	38 322	49 634	315 970	2 056

Table 2: Distribution of label and sensitive attribute values

Dataset	Label	Count (Percentage)	
		Class 1	*Class 0*
Alibaba	gender	42 192 (25.3%)	124 766 (74.7%)
JD	gender	13 735 (35.8%)	24 587 (64.2%)
Dataset	**Sens. Attr.**	**Count (Percentage)**	
		Class 1	*Class 0*
Alibaba	bin-age	71 583 (42.9%)	95 375 (57.1%)
JD	bin-age	25 717 (67.1%)	12 605 (32.9%)

JD dataset[2] consists of users and items from the retailer company of the same name having *click* and *purchase* relationships, already used in [33]. Since our experiments are not focused on the effectiveness of user profiling, and due to the massive size of the original dataset, we consider a sample of 15% of the items and only the *click* relationship to make experimental settings comparable. As for the previous dataset, CatGCN's local interaction modelling incorporates a *co-click* relationship. To make the experiments consistent, also for this dataset select *gender* as the label for user profiling task and *age* as the sensitive attribute. In this case, we binarised it (*bin-age*) by considering users under and over 35 years old.

Table 1 displays the characteristics of the two datasets. In particular, *features* represent the dimension of the categorical feature array used as input for CatGCN model. Table 2 shows the distribution within the datasets of target class and sensitive attribute values.

3.2 Experimental setting

We explore **RQ1** and **RQ2** for disparate impact assessment by running a user profiling task (i.e. classification of the *gender* class in both datasets) for each of the two models, CatGCN and RHGN, and computing the related fairness score in terms of Δ_{SP}, Δ_{EO} and Δ_{OAE}, which are defined in Eq. (5), Eq. (6) and Eq. (7), respectively.

In particular, to answer **RQ1**, we consider the three mentioned metrics and compare their scores between the two models to measure the models' discrimination level and evaluate the impact of the different user profiling paradigms on fairness, knowing that the smaller these scores are, the fairer the classifier is.

For **RQ2**, we contextualise the values of Δ_{SP} and Δ_{EO} with the results of **FairGNN** [8], a recent model that focuses on learning fair GNNs for node classification.

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

[2] https://github.com/guyulongcs/IJCAI2019_HGAT

Table 3: Experiment results. We report the best result for each dataset and metric in bold.

Dataset	Label	Model	Performance			Fairness			
			Accuracy	*F1-score*	*ROC AUC*	Δ_{SP}	Δ_{EO}	Δ_{OAE}	Δ_{TE}
Alibaba	gender	CatGCN	0.787 ±0.017	**0.714** ±0.006	**0.714** ±0.008	0.046 ±0.019	0.147 ±0.080	0.175 ±0.109	0.068 ±0.021
		RHGN	**0.812** ±0.005	0.704 ±0.017	0.681 ±0.016	**0.018** ±0.013	**0.133** ±0.086	**0.148** ±0.101	**0.017** ±0.013
JD	gender	CatGCN	0.721 ±0.007	**0.706** ±0.006	**0.712** ±0.006	0.033 ±0.013	0.050 ±0.017	0.062 ±0.020	0.150 ±0.066
		RHGN	**0.735** ±0.005	0.696 ±0.007	0.658 ±0.008	**0.009** ±0.007	**0.041** ±0.017	**0.054** ±0.017	**0.019** ±0.015

Table 4: Variations in fairness scores between CatGCN and RHGN. Differences in averages are considered.

Dataset	Variations in fairness scores			
	Δ_{SP}	Δ_{EO}	Δ_{OAE}	Δ_{TE}
Alibaba	0.028	0.014	0.027	**0.051**
JD	0.024	0.009	0.008	**0.131**

To answer **RQ3**, we extend our fairness evaluation to consider Δ_{TE}, defined in Eq. (8). The aim is to determine to what extent the analysed models discriminate against users from the perspective of disparate mistreatment, in comparison to disparate impact.

Regarding the user profiling tasks, models' hyper-parameters are defined as follows. For CatGCN, the *learning rate* is searched in {0.01, 0.1}, the L_2 regularisation coefficient and the *dropout ratio* are tuned among {1e-5, 1e-4} and {0.1, 0.3, 0.5, 0.7}, respectively, and the aggregation parameter α is searched within {0.1, 0.3, 0.5, 0.7, 0.9}. For RHGN, due to the heavier computational time for each experiment, the hyper-parameter selection is narrower than the other model. The *learning rate* and the L_2 regularisation coefficient are set to 0.01 and 0.001, respectively; the hidden dimension of the two layers of the entity-level aggregation network is searched in {32, 64}, while the number of heads in multi-head attention is tuned among {1, 2}. All other parameters are set according to the original papers. After the grid search, the experiments on fairness are executed 10 times and the probabilities characterised by Eqs. (5)-(8) are evaluated on the test set. The experiments are performed on a GPU Nvidia Quadro RTX 8000 48GB.

3.3 Results

Table 3 shows the results of the presented assessment. For each dataset and model, we first report the *accuracy*, *F1-score*, and *ROC AUC* performance results. Then, we measure the fairness scores.

The results in terms of performance show that, in both datasets, RHGN is more accurate than its counterpart. Anyhow, the F1-score and the ROC AUC show that CatGCN is more effective from these two perspectives. Hence, CatGCN is less impacted by false positives and false negatives. On the other hand, RHGN produces more true positives and true negatives.

Moving to the analysis of the fairness values, RHGN can produce less discrimination in its outcome than CatGCN (**RQ1**), and this result is confirmed for all the considered metrics.

Observation 1. *The ability of RHGN to represent users through multiple interaction modelling gains better values in terms of fairness than a model only relying on binary associations between users and items, as CatGCN, which also amplifies discrimination by modelling users' local interactions (e.g. co-click relationship).*

Comparing the fairness scores with the results of the baseline model, FairGNN (**RQ2**), we observe different cases: knowing that the values must be close to 0 for a model to be considered *fair* for a specific metrics, only RHGN is effective *w.r.t.* Δ_{SP} in both experimental settings, while in all other cases none of the two analysed models can be deemed *fair*.

Observation 2. *Even though RHGN demonstrates to be a fairer model than CatGCN, a debiasing process is equally needed in order to exploit the user models produced by both GNNs as fair.*

The extended fairness analysis (**RQ3**) is presented in Table 4 by providing the variations in fairness metrics scores between CatGCN and RHGN. The results show that the widest difference is registered for Δ_{TE}, denoting the need to consider both disparate impact and disparate mistreatment metrics to have a complete picture of the fairness situation.

Observation 3. *In scenarios where the correctness of a decision on the target label w.r.t. the sensitive attributes is not well defined, or where there is a high cost for misclassified instances, a complete fairness assessment should always take into account disparate mistreatment evaluation, since disparate impact results could be misleading for these specific contexts.*

4 CONCLUSIONS AND FUTURE WORK

Recent GNN-based behavioural user profiling models monitor the interactions between the users and a platform to build a user representation that characterises their preferences. While the rest of the literature has analysed how effective are these models at predicting certain characteristics of the users, in this paper, we analysed possible disparities emerging from how users belonging to different demographic groups are classified (*unfairness*).

Our analysis on two state-of-the-art models, and real-world datasets, covering four fairness metrics, showed that directly modelling raw user interactions with a platform hurts a demographic group, who gets misclassified more than its counterpart.

In future work, we will explore these phenomena more in-depth, also considering different attributes, datasets, and models. Moreover, we will provide interventions to mitigate unfairness.

REFERENCES

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[5] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[6] Weijian Chen, Fuli Feng, Qifan Wang, Xiangnan He, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. CatGCN: Graph Convolutional Networks with Categorical Node Features. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[7] Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019. Semi-supervised user profiling with heterogeneous graph attention networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2116–2122.

[8] Enyan Dai and Suhang Wang. 2021. 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*. 680–688.

[9] Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li. 2021. Individual fairness for graph neural networks: A ranking based approach. In *Proceed. of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 300–310.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[11] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2020. FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9, 2 (2020), 197–213.

[12] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2019. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access* 7 (2019), 144907–144924.

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.

[14] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2125–2126.

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[17] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. 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*. 639–648.

[18] Sumitkumar Kanoje, Sheetal Girase, and Debajyoti Mukhopadhyay. 2015. User profiling trends, techniques and applications. *arXiv preprint arXiv:1503.07474* (2015).

[19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*.

[20] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proc. of The Web Conference 2018*. 101–102.

[21] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1023–1031.

[22] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2654–2657.

[23] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (2018), 2257–2270.

[24] Danny Poo, Brian Chng, and Jie-Mein Goh. 2003. A hybrid approach for user profiling. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 9–13.

[25] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2022. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *International Journal of Human–Computer Interaction* (2022), 1–20.

[26] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. *arXiv preprint arXiv:1804.08049* (2018).

[27] Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. 2019. Fairwalk: towards fair graph embedding. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3289–3295.

[28] Zhiqing Sun, Jian Tang, Pan Du, Zhi-Hong Deng, and Jian-Yun Nie. 2019. Divgraphpointer: A graph pointer network for extracting diverse keyphrases. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 755–764.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[30] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness (FairWare 2018)*. IEEE, 1–7.

[31] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2021. Modeling Techniques for Machine Learning Fairness: A Survey. *arXiv preprint arXiv:2111.03015* (2021).

[32] Chuhan Wu, Fangzhao Wu, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. 2019. Neural demographic prediction using search query. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. 654–662.

[33] Qilong Yan, Yufeng Zhang, Qiang Liu, Shu Wu, and Liang Wang. 2021. Relation-aware Heterogeneous Graph for User Profiling. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 3573–3577.

[34] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7370–7377.

[35] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 974–983.

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[37] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 793–803.

[38] Ziwei Zhang, Peng Cui, and Wenwu Zhu. 2022. Deep Learning on Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2022), 249–270.