# Fair Online Dating Recommendations for Sexually Fluid Users via Leveraging Opposite Gender Interaction Ratio

Yuying Zhao
yuying.zhao@vanderbilt.edu
Vanderbilt University

Yu Wang
yu.wang.1@vanderbilt.edu
Vanderbilt University

Yi Zhang
yi.zhang@vanderbilt.edu
Vanderbilt University

Pamela Wisniewski
pam.wisniewski@vanderbilt.edu
Vanderbilt University

Charu Aggarwal
charu@us.ibm.com
IBM T. J. Watson Research Center

Tyler Derr
tyler.derr@vanderbilt.edu
Vanderbilt University

## ABSTRACT

Online dating platforms have gained widespread popularity as a means for individuals to seek potential romantic relationships. While recommender systems have been designed to improve the user experience in dating platforms by providing personalized recommendations, increasing concerns about fairness have encouraged the development of fairness-aware recommender systems from various perspectives (e.g., gender and race). However, sexual orientation, which plays a significant role in finding a satisfying relationship, is under-investigated. To fill this crucial gap, we propose a novel metric, Opposite Gender Interaction Ratio (OGIR), as a way to investigate potential unfairness for users with varying/fluid preferences towards the opposite gender. We empirically analyze a real online dating dataset and observe existing recommender algorithms could suffer from group unfairness according to OGIR. We further investigate the potential causes for such gaps in recommendation quality, which lead to the challenges of group data imbalance and group calibration imbalance. Ultimately, we propose a fair recommender system based on re-weighting and re-ranking strategies to respectively mitigate these associated imbalance challenges. Experimental results demonstrate both strategies improve fairness while their combination achieves the best performance towards maintaining model utility while improving fairness.

**ACM Reference Format:**
Yuying Zhao, Yu Wang, Yi Zhang, Pamela Wisniewski, Charu Aggarwal, and Tyler Derr. 2018. Fair Online Dating Recommendations for Sexually Fluid Users via Leveraging Opposite Gender Interaction Ratio . In *International Workshop on Mining and Learning with Graphs.* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Online dating has grown increasingly popular and is now a leading way of finding romantic partners and even meeting new friends [24]. For example, in 2022 it was estimated that 30% of U.S. adults had used online dating and even upwards of 51% among lesbian, gay or bisexual adults[1]. To accommodate this growing demand, various platforms have emerged, e.g., OkCupid, Tinder, and Grindr. With the booming of users, the challenge of information/choice overload [21] and unawareness [6] have made recommender systems (RS) even more important, which learn user preferences via their interactions/behaviors on the platform. This ultimately provides users with recommended partners that hopefully match their interests and significantly enhance the user experience [30].

However, while RS improve user satisfaction, fairness concerns still exist if systems are solely designed to maximize overall utility. For example, race-related fairness has been investigated to decrease racial homogamy via agent-based model interventions on online dating platforms [9]. Additionally, in online dating, different gender identities have diverse characteristics, motivations, preferences, etc [1]. Thus, if ignored, this generally leads to an inherent distinction in recommendation quality across gender identities, which has motivated past work on gender-aware system modifications to ensure equitable outcomes [34]. Nevertheless, although the aforementioned fairness perspectives are crucial and provide additional consideration beyond utility, another important sensitive user characteristic associated with dating is their sexual orientation, but less commonly discussed in the literature.

In one of the most basic forms, the satisfaction of a recommendation is contingent upon users' sexual orientations and the gender identity of those being recommended to them. Various sexual orientations indicate users' sexual preferences, including but not limited to homosexual individuals who prefer the same gender as their romantic partner, heterosexual individuals who prefer the opposite gender, and bisexual individuals who are attracted to both the same and opposite genders. However, even for a bisexual individual the spectrum as to their preference on dating certain genders can vary, raising further challenges in the recommendation system. To exacerbate this issue, studies have shown that personal experiences with online dating significantly differ by sexual orientation [6, 23].

*With diverse preferences and demands, could users with various sexual orientations be treated similarly?* Unfortunately, unfairness would be likely to exist for the following assumption - heteronormativity. Specifically, heterosexual users are generally the majority of dating applications (if without specific design, such as Grindr, which is designed specifically for the LGBTQ community), and RS inherently tend to perform better for users aligned with the preferences/behaviors of the majority while compromising the performance of the minority; thus, leading to the unfairness. However,


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1] https://www.pewresearch.org/key-findings-about-online-dating-in-the-u-s/

while these minority groups by definition are lower in percentage they are also increasing in size [12] and nearly twice as likely to report using an online dating platform[1]. This indicates despite comprising a smaller proportion of users, minority groups constitute a substantial number of individuals who might have a higher desire for online dating services and deserve quality recommendations.

Although the above discussion strengthens the motivation and the need to investigate the potential unfairness of RS in online dating platforms according to users' sexual orientations, it is nontrivial to study this problem due to the following challenges: (C1) There is a lack of knowledge of accurate sexual orientation. While platforms could allow users to specify their sexual orientation, some users might be reluctant to specify their sexual orientations due to privacy considerations or a lack of suitable selection options on the dating platform; sexual orientation alone is insufficient for a high-quality recommendation, especially in bisexual users (e.g., if a user identifies as bisexual and tends to prefer mostly users of the opposite gender, but the system recommends primarily users of the same gender, it would result in unsatisfying recommendation performance); sexual fluidity is prevalent, and users' sexual orientation might change over time. (C2) Improving fairness without compromising overall utility is a long-standing issue in fairness-related studies and has no established answers till now [16].

To address these challenges, this work presents the initial endeavor to investigate fairness of online RS from sexual orientation perspective. To obtain knowledge about sexual orientation, rather than directly classifying users into various categories which are unreliable due to a lack of user profiles in our dataset, we extract an interaction-based metric called Opposite Gender Interaction Ratio (OGIR), which serves as an implicit indicator (i.e., if an individual interacts with both genders, but mostly with the opposite gender, they are likely bisexual but with a stronger preference to the opposite gender). After obtaining OGIR, we divide users into groups where groups have different levels of OGIR, indicating their diverse preferences towards the opposite gender. Given groups, we empirically investigate and verify the existence of group unfairness in existing RS where groups are treated differently in terms of recommendation quality. To mitigate the performance gap among groups, we identify two potential causes: group data imbalance and calibration imbalance [26]. Correspondingly, we propose an in-processing re-weighting strategy and a post-processing re-ranking strategy. Experimental results show that both strategies improve fairness and have their unique advantages. When utilized together, these strategies lead to best performance in improving fairness while maintaining utility performance. Our main contributions are:

- We observe the presence of consistent group unfairness based on Opposite Gender Interaction Ratio (OGIR), which is related to users' sexual orientation, in multiple baseline recommender algorithms for a real-world online dating dataset;
- We identify two potential causes for group unfairness: group data imbalance and calibration imbalance. Correspondingly, we design a re-weighting strategy and a re-ranking strategy;
- Experiments show that both strategies are effective at reducing the recommendation quality gap across groups divided by OGIR. Furthermore, combining the two strategies results in the best performance towards maintaining similar utility performance while improving fairness across user groups.

## 2 RELATED WORK

### 2.1 Recommender Systems in Online Dating

RS serves as an effective solution to tackle information overload by delivering personalized recommendations. There have been numerous works in designing online dating RS, including interaction-based and content-based methods. Most interaction-based methods employ collaborative filtering [3, 13], which generate recommendations according to user similarities. For instance, collaborative filtering methods had been previously used to estimate the attractiveness rating of user pairs according to the ratings of similar users [3]. On the other hand, content-based methods utilize user profiles and features for recommendations [8, 33]. For example, Latent Dirichlet Allocation (LDA) has been previously used to learn user preferences [27]. Additionally, to satisfy user requirements from both ends, reciprocal recommendation methods are proposed [20, 30]. In summary, these approaches effectively capture user preferences and enhance user experience. Nonetheless, few of them take fairness into account during algorithm development.

### 2.2 Fairness in Online Dating

Although fairness in RS has been studied in online dating, there are still relatively few works. The most related stream of work focuses on promoting fairness among groups of users according to their associated sensitive attribute, with race [19, 25], gender [17, 34], and religion [19] being among the most commonly studied. For example, a group fairness metric that not only depends on the ranking results but also on the distribution of user attention was proposed to improve racial fairness [25]. In addition, individual fairness metrics have also been developed, such as calibration-based methods to encourage recommending potential partners that match user preferences focusing on race and religion [19], which shares a similar objective to our research in terms of promoting fairness through calibration, but they focus on conformity to user preferences, while our aim is to mitigate the performance disparity among user groups according to their sexual orientations. Specifically, we also aim to ensure fairness among groups divided based on sensitive attributes, but to the best of our knowledge, this work presents the first endeavor to study fairness from the perspective of sexual orientation and draw connections to imbalanced learning.

## 3 ONLINE DATING DATASET ANALYSIS

In this work, we use a real-world dataset from Líbímseti.cz (which is hosted in the Czech Republic) and is publicly available [3, 14][2]. Unfortunately, many works are unable to make their data public [2, 30, 31], and other available dating datasets pose limitations For example, OkCupid and Lovoo [3], provide user profiles, but without interactions. The Speed Dating dataset[3] was gathered from experimental speed dating events, but therefore smaller scale and not related to online dating. Therefore, this Líbímseti.cz dataset is particularly valuable as it not only contains user interactions, but also the self-identified gender information of the users, and the platform was not exclusively designed for heterosexual users, which enables the investigation presented in this work.

[2]Dataset used in this study: Líbímseti.cz
[3]Other available datasets: OKCupid; Lovoo; Speed dating

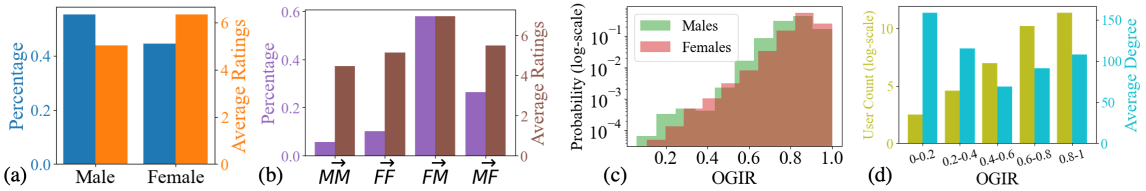


**Figure 1: Dataset analysis (a) gender identity distribution and their average ratings; (b) interaction type distribution and their average ratings; (c) OGIR distribution of female/male users; (d) user counts and average degrees according to OGIR.**

This section presents a detailed analysis of the Líbímseti.cz dataset, providing additional context for interpreting our empirical results. Overall the dataset [14] contains 220,970 users and 17,359,346 interactions in the form of $(u, v, r)$ tuples where user $u$ rates user $v$ with score $r$ according to $u$'s preference. Some users have filled in their (binary[4]) gender information, while others' remain unknown. In this study, we concentrate on users who provide gender identity information. The detailed binary gender identity distribution and their corresponding average ratings given to other users is shown in Fig. 1(a). Among the users with gender information, we further explore the types of interactions where one user rates the other, leading to four types ['Male→Male', 'Female→Female', 'Female→Male', 'Male→Female'] abbreviated as [$\overrightarrow{\text{MM}}$, $\overrightarrow{\text{FF}}$, $\overrightarrow{\text{FM}}$, $\overrightarrow{\text{MF}}$]. The interaction type distribution along with their average ratings is shown in Fig. 1(b). Based on users' interaction, we count the proportion of each user interacting with opposite genders, measured by opposite gender interaction ratio (OGIR).

**Opposite Gender Interaction Ratio (OGIR)** for a user defines the ratio of opposite genders among this user's interaction history, which captures the tendency of a user being sexually attracted by other users of the opposite gender. Suppose user $u$ has rated $N_u$ users among which $\hat{N}_u$ is the number of individuals from opposite gender with user $u$. Formally, it is defined as: $\text{OGIR}_u = \hat{N}_u / N_u$. By definition, OGIR lies in the range [0, 1]. Users with OGIR closer to 0 are more toward homosexual, and users with OGIR closer to 1 are more toward heterosexual.

The histogram of users' OGIR in Fig. 1(c) shows that most users, regardless of gender, prefer to interact with users of opposite genders. Fig. 1(b) shows that females ($\overrightarrow{\text{FF}}$ and $\overrightarrow{\text{FM}}$) on average tend to rate higher than males ($\overrightarrow{\text{MM}}$ and $\overrightarrow{\text{MF}}$). Additionally, hetero-interactions (i.e., interaction between different genders, $\overrightarrow{\text{FM}}$ and $\overrightarrow{\text{MF}}$) tend to have higher ratings than homo-interactions (i.e., interaction between the same gender, $\overrightarrow{\text{FF}}$ and $\overrightarrow{\text{MM}}$). We also plot the user number and average degree according to OGIR in Fig. 1(d). The user count aligns with the conclusion from Fig. 1(c) where the majority prefer the opposite gender. The degree shows that users with low and high OGIR tend to have more interactions on average.

To summarize, we draw the following observations:

- Males take up a larger proportion than females, but females tend to rate more frequently than males, leading to a larger proportion of $\overrightarrow{\text{FF}}$ and $\overrightarrow{\text{FM}}$ than $\overrightarrow{\text{MM}}$ and $\overrightarrow{\text{MF}}$.
- Most interactions are between different genders (i.e., $\overrightarrow{\text{FM}}$, $\overrightarrow{\text{MF}}$) while those within same gender also exist (i.e., $\overrightarrow{\text{FF}}$, $\overrightarrow{\text{MM}}$), which indicates the interactions are multi-faceted and (on average) users with OGIR 0 to 0.4 have the highest level of engagement/degree.
- Users tend to prefer/ignore the opposite gender at varied levels, which indicates that user sexual preferences toward the opposite gender are complex and diverse.

[4]This work focuses on binary gender, primarily attributed to the limited dataset and does not reflect the authors' opinions on gender identity.

# 4 FAIRNESS CONCERNS IN ONLINE DATING RECOMMENDATIONS

In Sec. 3, we analyzed complex user behaviors in a real-world online dating site with an emphasis on the users' opposite gender interaction ratio (OGIR), which provides insight into user sexual orientations according to their historical interactions. In this section, *we seek to study whether users grouped by OGIR, who have diverse levels of preferences toward the opposite gender, would be treated fairly* if a recommender system was to be applied to improve their user experience. Specifically, we first formally define the group unfairness based on the average performance gap between groups, then we perform an initial empirical evaluation on off-the-shelf recommendation algorithms to simulate whether unfairness was to exist if such a recommender system if deployed in the real world.

## 4.1 User-based Group Unfairness

Following existing literature that fairness can be interpreted as the equality of utility across entities in different groups [7, 15], we define user-based group unfairness as the difference of recommendation performance across users with different levels of OGIR. Intuitively, a larger performance gap indicates higher discrimination/lower fairness. In the following, we define how to divide groups based on OGIR and the corresponding unfairness metrics.

*4.1.1 Group Partition.* To quantify such unfairness, we divide users into multiple groups based on their OGIR. Users in each group are within the same interval of OGIR where the interval range for each group is the same. For this study, we construct a 3-group partition where groups are denoted as $G_1$, $G_2$, and $G_3$, and have users with OGIR in ranges $[0, \frac{1}{3})$, $[\frac{1}{3}, \frac{2}{3})$, and $[\frac{2}{3}, 1]$, respectively. These groups have different levels of OGIR, indicating their diverse preferences toward the opposite gender, and can be seen as partitioning users likely identifying as bisexual into three partitions.

*4.1.2 User-based Group Unfairness Metric.* Our proposed metric measures the discrepancy of recommendation performance among groups $\mathcal{G}$, which is defined as the average performance gap of certain metrics X (e.g., recall, F1, etc) among group pairs:

$$\Delta_X(\mathcal{G}) = \frac{1}{Q_X^{ave}} \mathbb{E}_{(G_1,G_2)\in\mathcal{G}\times\mathcal{G}} |Q_X(G_1) - Q_X(G_2)|, \quad (1)$$

where $(G_1, G_2)$ is a unique group pair (i.e., $G_1 \neq G_2$), and $Q_X(G_i) = (\sum_{u\in G_i} q_X(u))/|G_i|$ is the average recommendation performance

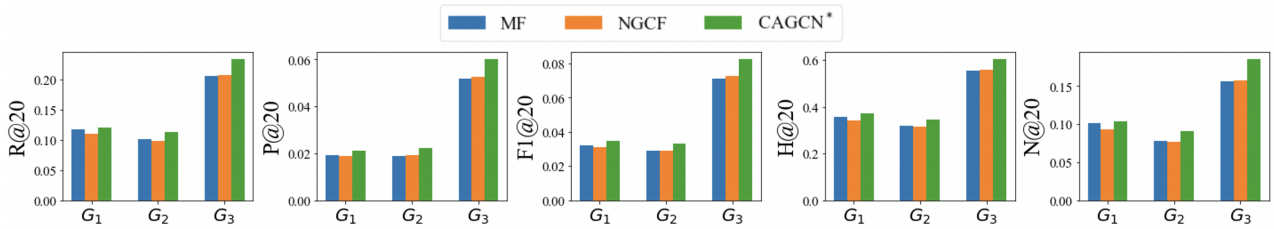


**Figure 2: Utility performance of three models on five metrics, where groups are divided based on even width bins for discretizing OGIR into three groups ($G_1 = \{u|\text{OGIR}_u \in [0, \frac{1}{3})\}$ with $G2$, $G3$ similarly defined). $G_3$ consistently has better performance.**

measured by metric X of users in the group $G_i$ with $q_x(u)$ being user $u$'s performance according to metric X. The denominator normalizes by the average performance to mitigate the impact of performance scale across metrics where $Q_X^{\text{ave}} = (\sum_{G \in \mathcal{G}} Q_X(G))/|\mathcal{G}|$.

## 4.2 Initial Fairness Evaluation

In this section, we evaluate various models to investigate group unfairness issue. We first introduce the pre-processing steps, and evaluation metrics and models. We then report the experimental results, which reveal the consistent presence of group unfairness across algorithms. In the end, we discuss two potential naïve "fixes" and why they could not work, which urges the need for a fair model.

*4.2.1 Pre-processing.* To improve data quality, we perform the following steps: (1) filter accounts lacking self-identified gender information as OGIR needs gender context; (2) filter edges with a rating less than 10 so that the remaining edges show strong preferences; and (3) apply k-core setting iteratively to remove users with interaction number smaller than $k = 5$. After these steps, every user has at least five strong preferences and self-identified gender information. We randomly split the dataset into train/validation/test based on 60%/20%/20% proportions. To further ensure that all genders have the chance to be assigned to train/validation/test, we randomly split for females and males separately.

*4.2.2 Evaluation Metrics and Models.* We include various utility metrics [10] and their corresponding fairness metrics for a comprehensive comparison, including Recall (R@20), Precision (P@20), F1@20, Hit Ratio (H@20), and Normalized Discounted Cumulative Gain (N@20) and their corresponding fairness metrics, computed according to Eq. 1 ($\Delta_{\text{R}}$@20, $\Delta_{\text{P}}$@20, $\Delta_{\text{F1}}$@20, $\Delta_{\text{H}}$@20, and $\Delta_{\text{N}}$@20). For utility/fairness metrics, the higher/lower the value, the better the performance. We evaluate across three representative recommender models, which include seminal works and current state-of-the-art: MF [22], NGCF [28], and CAGCN* [29]. They are optimized with Bayesian Personalized Ranking (BPR) loss, $\mathcal{L}_{\text{BPR}}$ [22].

*4.2.3 Evaluation Results.* To mitigate the randomness impact for a better comparison, we run the evaluated models 5 times with different seeds and report the average results. Without specification, the group number is set to 3. The model selection is based on the average utility score on validation. The average utility result in Fig. 2 shows that generally, $G_3$ has better performance than $G_1$ and $G_2$, indicating that $G_3$ enjoys better recommendation quality. The performance gap among groups is quantified by the proposed unfairness measurement where these models have more than 0.5 unfairness scores, presenting a consistent unfairness that appears to be algorithm/model-agnostic according to our results.

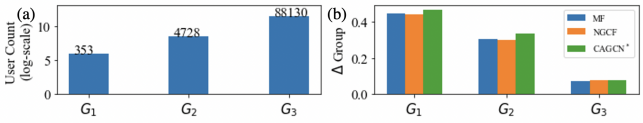


**Figure 3: Two potential causes of unfairness (a) group data imbalance; (b) group calibration imbalance.**

*4.2.4 Potential Naïve "Fix" Towards Fairness.* One potential approach to addressing the unfairness issue could be a fairness-aware model selection. For example, one could use score = Avg Utility – Avg Fairness. The experiment shows no significant fairness improvement compared with baseline models selected based on utility criteria. This indicates that simply considering fairness in model selection is insufficient for a fair model. Another potential solution would be to train the recommender system separately for different groups. However, it presents two challenges. First, it will further exacerbate the data sparsity issue, and such an issue would be more severe for the minority than the majority. Secondly, in the real-world scenario, a user in one group might be interested in a user from another group. Separately training the recommendation system would result in the restricted recommendation and lead to a suboptimal outcome. Therefore, both potential naïve "fixes" cannot fix the problem. This raises the requirement of designing a new fair model, which we present in the next section.

# 5 FAIR RECOMMENDER SYSTEM

In this section, we analyze potential causes of group unfairness, which is related to group data and calibration imbalance. To mitigate them, we introduce re-weighting and re-ranking strategies.

## 5.1 Mitigating Group Data Imbalance: Re-weighting Towards Improved Fairness

The issue of class imbalance, where the number of training instances per class is imbalanced, has been widely investigated across various domains [5, 11]. During the training process, to achieve an overall higher utility performance, the majority class is typically optimized more than the minority class, leading to a performance gap. As shown in Fig. 3(a), the numbers of users in different groups are imbalanced in our setting (i.e., $G_1$ is the majority, $G_2$ and $G_3$ are the minorities). As a consequence, there are performance gaps among majority and minority groups, resulting in unfairness. To mitigate this unfairness, we employ the re-weighting strategy, which has been utilized to address the class imbalance issue. This approach adjusts the focus of training by updating the weights based on the number of users in each group effectively balancing the original loss function accordingly such that equitable emphasis is put on each group when updating the model's parameters.

In traditional $\mathcal{L}_{\text{BPR}}$, each tuple is trained equally without consideration of group size. Generally, as one group (e.g., majority) appears more in the training data during the optimization, the users belonging to this group will achieve better performance as they share common (group-level) user behaviors. To remedy this, we add a weight term for adjustment. The updated loss is as follows:

$$\mathcal{L}_{\text{BPR}}^{\text{re-weighting}} = -\sum_{(u,i,j)\in\mathcal{D}} w_{G(u)} \log \sigma(\hat{y}_{ui} - \hat{y}_{uj}) + \lambda_\Theta \|\Theta\|^2,$$

with training data $\mathcal{D} = \{(u,i,j)|u \in \mathcal{U}, i \in \mathcal{I}_u^+, j \in \mathcal{I}_u^-\}$, total user set $\mathcal{U}$, user sets $u$ did (not) interacted with $\mathcal{I}_u^+$ ($\mathcal{I}_u^-$), predicted preference score $\hat{y}_{ui}$, and user $u$'s weight based on $u$'s group, $w_{G(u)}$. Generally, when $u$ belongs to a group with a larger user number (e.g., $G_3$), the weight will be lower than the case when $u$ belongs to one with a smaller user number (e.g., $G_1$ and $G_2$) to promote the training for the minority. Specifically, we utilize $w_{G,p} = \frac{1}{N_G{}^p}$ where $N_G$ is the number of users in the group $G$ and $p$ is for different weight assignments. Compared with the original objective function with the single goal to improve utility, the updated objective considers utility and fairness simultaneously with $p$ balancing two goals.

## 5.2 Mitigating Group Calibration Imbalance: Re-ranking Towards Improved Fairness

The notion of calibration in recommendation refers to the property that the genre distribution (e.g., Sci-Fi, Romance, etc. in movie recommendation) in the recommendation list should match the distribution in the history interactions [26]. A higher-quality calibration means a lower level of inconsistency between the distributions, which indicates that the model can better preserve users' preferences. In dating recommendation, a good calibration requires the ratios of males/females in training and recommendation to be similar. We quantify the calibration score of a user $u$ with the inconsistency between the ratio of female users that are interacted in the training dataset (i.e., $T^F(u)$) and the ratio of females that are recommended in the recommendation list $\mathcal{R}_u$ (i.e., $R^F(\mathcal{R}_u)$) by the absolute value and quantify the calibration of a group by averaging the users in that group as follows:

$$\Delta_{\text{User}}(u, \mathcal{R}_u) = |T^F(u) - R^F(\mathcal{R}_u)|,$$
$$\Delta_{\text{Group}}(G, \mathcal{R}) = \sum_{u\in G} \Delta_{\text{User}}(u, \mathcal{R}_u).$$

The group calibration results of baseline models in Fig. 3(b), where a low calibration score indicates better calibration, show that the levels of calibration differ among groups. It shows an opposite trend with the performance in Fig. 2 that $G_3$ has the lowest calibration score and the highest performance. We posit that utility performance is negatively correlated with calibration scores. Since the trained model is more towards the majority, the ability to preserve the users' preferences is compromised for the minority. Based on this hypothesis, we aim to mitigate the calibration imbalance issue by reducing the inconsistency between the gender ratio of training interactions and the recommendation list by re-ranking strategy. The minority has poor calibration, which on the other hand, indicates a large space for improvement. Therefore, by ensuring better calibration, it can potentially improve the utility performance of all groups with a larger improvement for the minority group, which will lead to a decrease of utility gap and thus improve fairness.

**Algorithm 1:** Greedy Algorithm for Re-ranking to Mitigate Calibration Imbalance

**Input:** Recommendation number $K$; user id $u$; trade-off parameter $\lambda$, $u's$ top $K'$ baseline recommendations as candidates $C_u$

$\mathcal{R}_u = \{\}$
**while** $|\mathcal{R}_u| \leq K$ **do**
  $i^* = \text{argmax}_{i\in C\setminus\mathcal{R}_u} (1-\lambda)S(\mathcal{R}_u \cup \{i\}) - \lambda\Delta_{\text{User}}(u, \mathcal{R}_u \cup \{i\})$
  $C_u = C_u \setminus \{i^*\}$
  $\mathcal{R}_u = \mathcal{R}_u \cup \{i^*\}$
**return** *User u's re-ranked recommendation list* $\mathcal{R}_u$

The re-ranking strategy is a post-processing mechanism to find a new recommendation based on the original recommendations from baseline models. With the consideration of utility and calibration, we use Maximum Marginal Relevance (MMR) [4, 26, 32] to determine the recommendation list $\mathcal{R}_u^*$ for user $u$, so our objective is formalized as follows:

$$\mathcal{R}_u^* = \mathcal{R}_u, |\mathcal{R}_u|=\text{K} \ (1-\lambda)S(\mathcal{R}_u) - \lambda\Delta_{\text{User}}(u, \mathcal{R}_u) \tag{2}$$

The objective is composed of two terms with trade-off parameter $\lambda \in [0,1]$ (1) the predicted relevance score $\hat{y}_{ui}$ from baseline models related to the utility performance, where $S(\mathcal{R}_u) = \sum_{i\in\mathcal{R}_u} \hat{y}_{ui}$; and (2) the calibration score $\Delta_{\text{User}}(u, \mathcal{R}_u)$. Additionally, as $\Delta_{\text{User}}(u, \mathcal{R}_u) \in [0,1]$, we rescale the relevance scores so that they fall in the same range. Solving Eq. 2 NP-hard [26]. We adopt a greedy algorithm [18] as in Algorithm 1, which finds the approximate solution with $(1-\frac{1}{e})$ optimality guarantee where $e$ is the natural logarithm. To recommend potential partners for a user $u$, Algorithm 1 starts with an empty list with top $K'$ individuals recommended from the original baseline models as the candidate set $C_u$ and then iteratively adds the optimal individual that obtains the largest score. The algorithm ends when the list reaches length $K$.

# 6 EXPERIMENTS

In this section, we conduct experiments to verify the effectiveness of Re-weighting and Re-ranking strategies[5] under the setting of $K = 20$ and $K' = 100$. We aim to answer two main research questions for both of these two strategies.

- **RQ1**: How well can proposed strategies improve fairness while not significantly decreasing utility performance?
- **RQ2**: What are the impacts of the model hyperparameters (i.e., $p$ in re-weighting and $\lambda$ in re-ranking)?

To answer these questions, we first report the re-weighting and re-ranking results. We also report the result of applying them jointly. After analyzing the results, we present a discussion about these strategies in the end.

## 6.1 Experimental Results with Re-weighting

In this section, we report the utility and fairness performance after applying the re-weighting strategy. We then conduct a sensitivity analysis to explore the effect of re-weighting hyperparameter $p$.

*6.1.1 Re-weighting Performance.* We report the test performance for specific $p$ in Table 1. We do not report standard deviations both for space considerations and also because they are always less than 0.02. We tune the hyperparameter $p$ and select the optimal value based on the validation dataset, where we plot the validation curve as shown in Fig. 4(d-f) and select $p$ before the sharp decrease in

[5]Source code is available at: Code link

**Table 1: Performance comparison of baseline model versus re-weighted model ($model_{rw}$). The ↑ represents the larger the better and ↓ represents the opposite. The proportion (+/- %) shows the performance improvement/degradation to the baseline model.**

| Method | Utility Metrics ↑ | | | | | | Fairness Metrics ↓ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@20 | P@20 | F1@20 | H@20 | N@20 | Avg Utility | $\Delta_R$@20 | $\Delta_P$@20 | $\Delta_{F1}$@20 | $\Delta_H$@20 | $\Delta_N$@20 | Avg Fairness |
| MF | 0.2002 | 0.0499 | 0.0690 | 0.5406 | 0.1517 | 0.2023 | 0.4964 | 0.7361 | 0.6397 | 0.3861 | 0.4664 | 0.5449 |
| NGCF | 0.2019 | 0.0508 | 0.0701 | 0.5457 | 0.1527 | 0.2043 | 0.5294 | 0.7577 | 0.6611 | 0.4016 | 0.4961 | 0.5692 |
| CAGCN* | 0.2267 | 0.0580 | 0.0798 | 0.5890 | 0.1802 | 0.2267 | 0.5196 | 0.7534 | 0.6562 | 0.3929 | 0.4955 | 0.5635 |
| $MF_{rw}$ | 0.2003 | 0.0503 | 0.0694 | 0.5421 | 0.1472 | 0.2019 (-0.20%) | 0.3945 | 0.6491 | 0.5447 | 0.3106 | 0.3264 | 0.4450 (+18.33%) |
| $NGCF_{rw}$ | 0.1932 | 0.0482 | 0.0668 | 0.5287 | 0.1430 | 0.1960 (-4.06%) | 0.3832 | 0.6274 | 0.5290 | 0.2924 | 0.3187 | 0.4301 (+24.44%) |
| $CAGCN^*_{rw}$ | 0.2242 | 0.0566 | 0.0781 | 0.5854 | 0.1780 | 0.2244 (-1.01%) | 0.4928 | 0.7222 | 0.6310 | 0.3718 | 0.4577 | 0.5351 (+5.04%) |

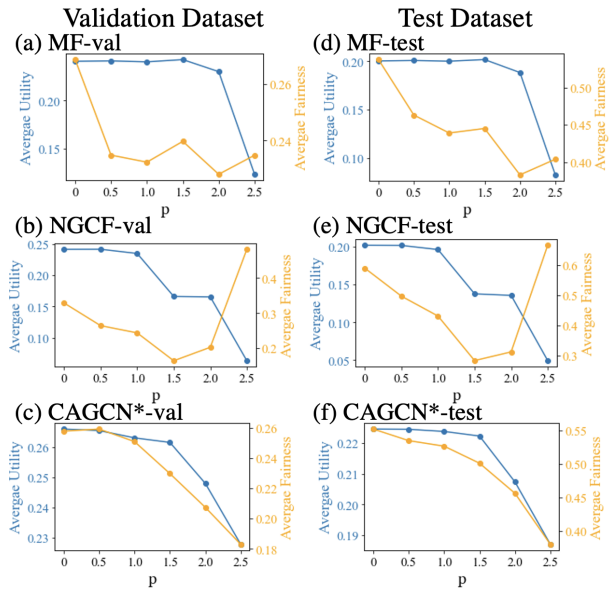


**Figure 4: Analysis on the utility and fairness performance impacts associated with the re-weighting hyperparameter $p$.**

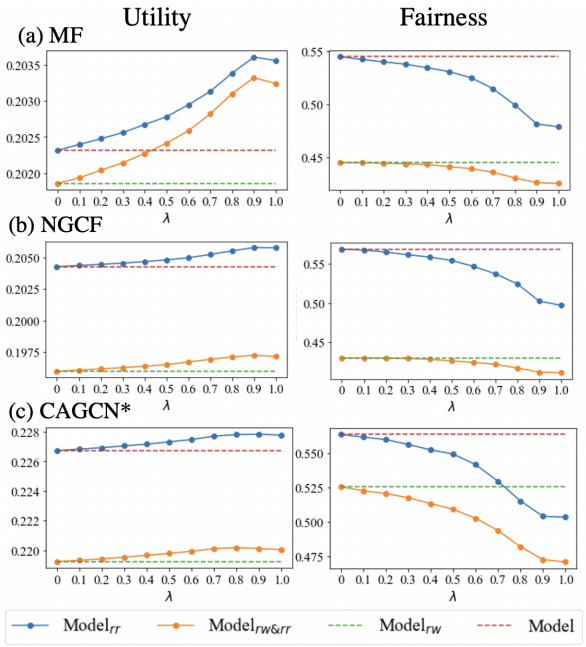


**Figure 5: The utility and fairness performance of variants (1) the baseline model (Model); (2) the re-weighting model ($Model_{rw}$); (3) the re-ranking model ($Model_{rr}$); and (4) the re-ranking model based on re-weighted model ($Model_{rw\&rr}$).**

utility performance to avoid a large compromise in the overall performance (i.e., 1.5 for MF, 1.0 for NGCF, and 0.5 for CAGCN*). Other strategies can be applied to select the best hyperparameter based on the validation curve, where the tradeoff between fairness and utility can be clearly observed. Thus, platforms can pick the hyperparameter based on their demands. In this way, the model selection is more flexible. Compared with the sensitivity analysis in Sec. 6.1.2, we would find that the validation curve generally matches the trend of the test curve, which validates that it is reliable to select the best hyperparameter based on the validation record. From Table 1, we observe that with the re-weighting strategy, for each method, the fairness improves with a little sacrifice of utility performance. NGCF has the best improvement in fairness (i.e., 24.44%), while CAGCN* has the smallest improvement (i.e., 5.04%).

*6.1.2 Sensitivity Analysis of $p$.* In this section, we investigate the effect of hyperparameter $p$, which controls the weight assignment. A larger $p$ means a larger difference in weight assignment among groups. The result in Fig. 4(a-c) shows that the impact of re-weighting on various methods is different, but they align well with the validation result. Therefore, the validation is effective in selecting a hyperparameter that matches the requirement for utility and fairness tradeoff. Generally, when $p$ increases, utility performance decreases while fairness performance increases. MF and NGCF gain a large fairness improvement with a small decrease in utility, but CAGCN* needs a larger sacrifice to obtain a larger improvement in fairness. For NGCF, we also observe an increase in fairness when enforcing a larger $p$. We hypothesize that this will also happen for the other two methods if we further increase $p$ since when utility performance becomes so low, the same quantity of performance gap would lead to larger unfairness according to the unfairness definition in Sec. 4. Another potential reason would be that the relative order of group performance might change at some certain $p$ (i.e., previously, the majority group has better performance, and now the minority might have better performance), resulting in the enlargement of the performance gap when $p$ increases.

## 6.2 Experimental Results with Re-ranking

In this section, we report performance after applying the re-ranking strategy on the baseline models and on the models after applying re-weighting (i.e., re-weighted models). We further conduct a sensitivity analysis on $\lambda$ from different perspectives, which presents the interpretation of the results.

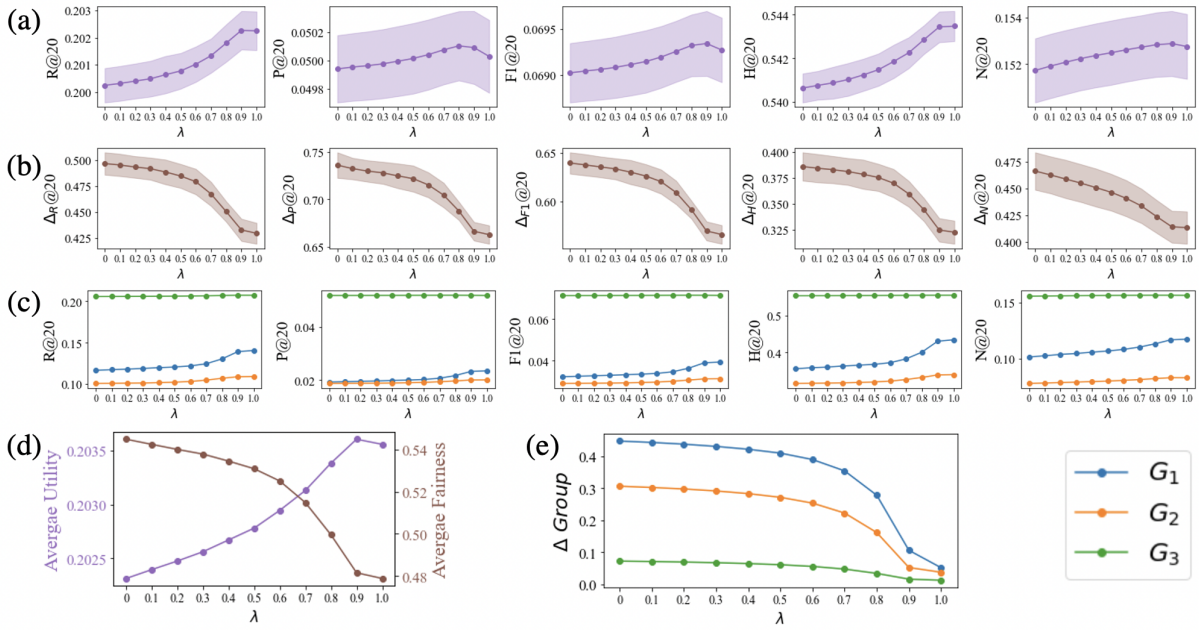


**Figure 6: Re-ranking results of different $\lambda$s on MF (a) utility performance; (b) fairness performance; (c) group utility performance; (d) average utility and fairness performance; (e) group calibration. The group results in (c) and (e) are for three groups divided based on their opposite gender interaction ratio (i.e., $G_1 = \{u|\mathbf{OGIR}_u \in [0, \frac{1}{3})\}$ with $G2$, $G3$ similarly defined).**

*6.2.1 Re-ranking Based on the Baseline Models.* The dashed red line and the solid blue line in Fig 5 correspond to the performance of baselines and re-ranking models. When $\lambda$ increases, utility and fairness performance both improve for all the baselines. For utility performance, MF has the largest improvement, while NGCF and CAGCN* show smaller improvements. For fairness performance, all of them experience an improvement. Surprisingly, the traditional utility-fairness trade-off (i.e., fairness usually improves at the cost of utility) does not occur. We give an in-depth analysis in Sec 6.2.3.

*6.2.2 Re-ranking Based on Re-weighted Models.* The dashed green line in Fig. 5 corresponds to the performance of the re-weighted model where the same hyperparameter is selected as in Sec. 6.1, and the solid orange line shows the performance of the re-ranking models based on the re-weighted models. A similar trend is observed. Both utility and fairness improve for all the methods after re-weighting. When comparing with the same $\lambda$ without re-weighting, the utility performance of $\text{Model}_{\text{rr}}$ is lower than $\text{Model}_{\text{rw\&rr}}$ since the base re-weighted model sacrifice a little utility performance as reported in Sec. 6.1. On the other hand, the re-weighted model has improved fairness, providing a good basis for re-ranking. Therefore, with the same $\lambda$, $\text{Model}_{\text{rw\&rr}}$ has better fairness than $\text{Model}_{\text{rr}}$. This result shows that the re-ranking strategy is effective irrespective of being applied to the baseline model or the re-weighted model.

*6.2.3 Sensitivity Analysis of $\lambda$.* We further investigate the reason why re-ranking strategy improves both fairness and utility. We conduct a comprehensive analysis on the hyperparameter $\lambda$. The result for MF-based model is shown in Fig. 6 (a similar trend is observed for other baseline and the re-weighted models). Fig. 6(a)-(b) suggest that generally, both utility and fairness improve along with the increase of $\lambda$ in every metric. The average metrics in Fig. 6(d)

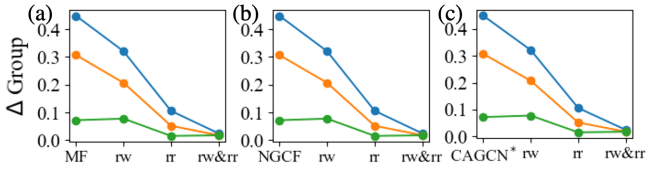


**Figure 7: Analysis of group calibration based on different model variants with re-weighting, re-ranking, and their combination denoted as rw, rr, rw&rr, respectively.**

gives us the same observation. A closer look at the performance per group in Fig. 6(c) shows that the performance change for $G_3$ is much more stable than $G_1$ and $G_2$, as $G_1$ and $G_2$ have a greater improvement when $\lambda$ increases. Since originally $G_3$ has better performance than $G_1$ and $G_2$ and now $G_1$ and $G_2$ have larger improvement than $G_3$, the performance gap between them decreases and leads to a better fairness score. The underlying reason for this imbalance of performance improvement is revealed in Fig. 6(e) where groups' calibration scores decrease when $\lambda$ increases, but due to the fact that $G_3$ has a small calibration score at first, its improvement is the weakest and thus the benefit is the least.

## 6.3 Discussion of Re-weighting and Re-ranking

The above results show that both re-weighting and re-ranking strategies improve fairness while the former improves fairness at the cost of utility, and the latter can even improve them simultaneously. We compare the performance in Fig. 5 and also the calibration in Fig. 7 of four variants: (1) the baseline model; (2) the re-weighting model ($\text{Model}_{\text{rw}}$); (3) the re-ranking model ($\text{Model}_{\text{rr}}$); (4) the re-ranking model based on re-weighted model ($\text{Model}_{\text{rw\&rr}}$). From the figures, we draw the following observations:

- **Effect on fairness performance**: re-weighting achieves better fairness than re-ranking on MF and NGCF, which indicates that the in-processing method might be more effective since it can change the training process and has more flexibility in providing a fairer recommendation. The combination $\text{Model}_{\text{rw\&rr}}$ achieves best fairness performance, which has the lowest calibration score.
- **Effect on utility performance**: re-weighting generally decreases utility performance, and re-ranking, on the other hand, can improve utility performance in addition to improving fairness.
- **Discussion on calibration**: re-weighting, although not designed to improve calibration, reduces the inconsistency when comparing Model and $\text{Model}_{\text{rw}}$, which gives another interpretation of its effectiveness in terms of fairness.

In summary, re-weighting and re-ranking strategies both have unique advantages. Re-weighting improves more on fairness, while re-ranking can improve utility and better calibration. Combining them leads to even better performance.

## 7 CONCLUSION

Sexual orientation, which is a significant factor for individuals to find a satisfying romantic relationship, is under-investigated in online dating recommender systems. In this paper, to investigate whether users with varying/fluid preferences for the opposite gender are treated fairly by recommender systems, we leverage our proposed metric, Opposite Gender Interaction Ratio (OGIR). The empirical experiments on a real-world online dating dataset show consistent unfairness among user groups based on OGIR across algorithms, which provide better recommendations for the majority groups than the minority groups (i.e., $G_3$ that has a higher level of preference toward the opposite gender, $G_1$ and $G_2$ that have a lower preference, respectively). Then, based on our validated hypothesis that bias/unfairness is associated with group data and group calibration imbalances, we propose a fair recommender system based on re-weighting and re-ranking strategies designed to alleviate the two imbalance challenges. Experimental results show that both strategies independently help improve fairness, but when combined they lead to the best overall performance in terms of maintaining utility while significantly improving fairness.

In the studied dataset, some users do not fill in gender identity, and one potential reason besides privacy concerns could be that the platform only provides binary options and these users do not identify themselves as male/female. Valuing the importance of these users, we will look into their characteristics and interaction patterns in the future. We also advocate dating platforms offer more gender identity options and explicitly collect information on sexual orientation to better serve users.

## REFERENCES


[1] Olga Abramova, Annika Baumann, Hanna Krasnova, and Peter Buxmann. 2016. Gender differences in online dating: What do we know so far? A systematic literature review. In *HICSS*. IEEE, 3858–3867.

[2] Mohammed Al-Zeyadi, Frans Coenen, and Alexei Lisitsa. 2017. User-to-user recommendation using the concept of movement patterns: A study using a dating social network. In *IC3K*.

[3] Lukas Brozovsky and Vaclav Petricek. 2007. Recommender System for Online Dating Service. In *Znalosti* (Ostrava, Czech Republic). VSB, Ostrava.

[4] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*. 335–336.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *JAIR* 16 (2002), 321–357.

[6] Eli J Finkel, Paul W Eastwick, Benjamin R Karney, Harry T Reis, and Susan Sprecher. 2012. Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public interest* 13, 1 (2012), 3–66.

[7] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *SIGIR*. 69–78.

[8] Günter J Hitsch, Al Hortaçsu, and Dan Ariely. 2010. Matching and sorting in online dating. *American Economic Review* 100, 1 (2010), 130–163.

[9] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy. In *FAccT*.

[10] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal* 16, 3 (2015), 261–273.

[11] Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (2019), 1–54.

[12] Jeffrey M Jones. 2021. LGBT identification rises to 5.6% in latest US estimate. *Gallup News* 24 (2021).

[13] Alfred Krzywicki, Wayne Wobcke, Xiongcai Cai, Ashesh Mahidadia, Michael Bain, Paul Compton, and Yang Sok Kim. 2010. Interaction-based collaborative filtering methods for recommendation in online dating. In *WISE*. Springer, 342–356.

[14] Jérôme Kunegis, Gerd Gröner, and Thomas Gottron. 2012. Online dating recommender systems: The split-complex number approach. In *RecSys workshop*.

[15] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Web Conference*.

[16] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in Recommendation: A Survey. *arXiv:2205.13619* (2022).

[17] Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

[18] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14, 1 (1978), 265–294.

[19] Dimitris Paraschakis and Bengt J Nilsson. 2020. Matchmaking under fairness constraints: a speed dating case study. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 43–57.

[20] Luiz Pizzato, Tomek Rej, Thomas Chung, Irena Koprinska, and Judy Kay. 2010. RECON: a reciprocal recommender for online dating. In *RecSys*. 207–214.

[21] Tila M Pronk and Jaap JA Denissen. 2020. A rejection mind-set: Choice overload in online dating. *Social Psychological and Personality Science* 11, 3 (2020), 388–396.

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv:1205.2618* (2012).

[23] Michael J Rosenfeld and Reuben J Thomas. 2010. Meeting online: The rise of the Internet as a social intermediary. *Paper session presented at the Population Association of America Meetings, Dallas, TX* (2010).

[24] Michael J Rosenfeld, Reuben J Thomas, and Sonia Hausen. 2019. Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *ProcNAS* 116, 36 (2019), 17753–17758.

[25] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *WWW*. 553–562.

[26] Harald Steck. 2018. Calibrated recommendations. In *RecSys*.

[27] Kun Tu, Bruno Ribeiro, David Jensen, Don Towsley, Benyuan Liu, Hua Jiang, and Xiaodong Wang. 2014. Online dating recommendations: matching markets and learning preferences. In *WWW*.

[28] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*. 165–174.

[29] Yu Wang, Yuying Zhao, Yi Zhang, and Tyler Derr. 2023. Collaboration-Aware Graph Convolutional Network for Recommender Systems. In *ACM Web Conference*. 91–101.

[30] Peng Xia, Benyuan Liu, Yizhou Sun, and Cindy Chen. 2015. Reciprocal recommendation system for online dating. In *ASONAM*. IEEE, 234–241.

[31] Kang Zhao, Xi Wang, Mo Yu, and Bo Gao. 2013. User Recommendations in Reciprocal and Bipartite Social Networks–An Online Dating Case Study. *IEEE intelligent systems* 29, 2 (2013), 27–35.

[32] Xing Zhao, Ziwei Zhu, and James Caverlee. 2021. Rabbit holes and taste distortion: Distribution-aware recommendation with evolving interests. In *WWW*. 888–899.

[33] Xuanzhi Zheng, Guoshuai Zhao, Li Zhu, Jihua Zhu, and Xueming Qian. 2022. What you like, what i am: online dating recommendation via matching individual preferences with features. *IEEE TKDE* (2022).

[34] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. 2018. Fairness in reciprocal recommendations: a speed-dating study. In *UMAP*. 29–34.