R-Fairness: Assessing Fairness of Ranking in Subjective Data

Lorenzo Balzotti[1]*, Donatella Firmani[1], Jerin George Mathew[1]
Riccardo Torlone[2], Sihem Amer-Yahia[3]
[1]Sapienza Università di Roma, Rome, Italy
[2]Università Roma Tre, Rome, Italy
[3]CNRS, University Grenoble Alpes, Grenoble, France
{lorenzo.balzotti, donatella.firmani, jeringeorge.mathew}@uniroma1.it
riccardo.torlone@uniroma3.it, sihem.amer-yahia@cnrs.fr

Abstract

Subjective data, reflecting individual opinions, permeate collaborative rating platforms like Yelp and Amazon, influencing everyday decisions. Despite the prevalence of such platforms, little attention has been given to fairness in their context, where groups of reviewers writing best-ranked reviews for best-ranked items have more influence on users' behavior. In this paper, we design and evaluate a new framework for the assessment of fairness of rankings for different reviewer groups in collaborative rating platforms. The key contributions are evaluating group exposure for different queries and platforms and comparing how various fairness definitions behave in different settings. Experiments on real datasets reveal insights into the impact of item ranking on fairness computation and the varying robustness of these measures.

1 Introduction

Subjective data represents people and their opinions (Tan, 2020; Li et al., 2019). Many subjective datasets are available nowadays and are used in decision-making, studying online behavior, or providing recommendations. *Collaborative rating* platforms such as Amazon for products, Yelp for restaurants, and Booking for hotels, are a special case of subjective data that are permeating our everyday decisions. Upon a user query, these platforms return a collection of items ranked in an order that is often not transparent to the users. Then, each item is presented with a collection of reviews in an order that typically is, again, rather opaque. This mechanism intrinsically favors the opinions of certain groups of reviewers over others.

In what follows, we refer to individuals writing reviews as *reviewers* and those submitting queries as *users* to distinguish their roles in this context.

Running Example. Suppose, as shown in Figure 1, that a user submits the query "Italian restaurants near me" on a collaborative rating platform like Yelp. The system returns a ranked list of restaurants ordered on relevance, rating, or other platform-specific factors. Each restaurant is associated with a ranked list of reviews from different reviewers, each of whom may belong to a distinct demographic (e.g. young reviewers) or opinion-based group (e.g. reviewers who mainly wrote negative reviews). The combination of the two levels of rankings (restaurants and reviews thereof) determines which opinions about restaurants are most visible, potentially amplifying the influence of certain groups on user decisions, while diminishing that of others. For example, this can occur for young reviewers if their reviews, highlighted in green in Figure 1, are ranked low Ⓐ, the restaurants they reviewed are ranked low Ⓑ, or both.

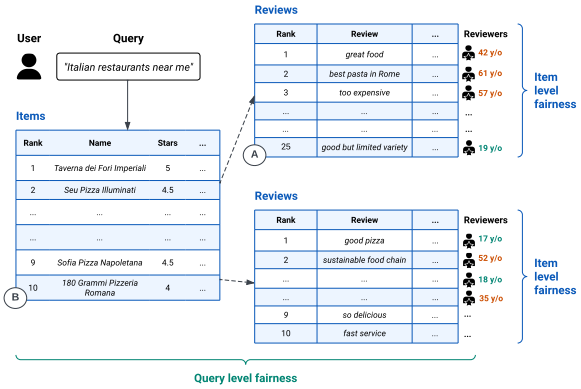


Figure 1: Ranking of items and reviews.

R-fairness. In this work, we address the novel problem of *r-fairness* in collaborative rating platforms, that is, how fairly these systems behave toward groups of reviewers. We consider reviewers' groups based on the traditional notion of demographics such as gender or age, as well as opinions, such as reviewers that are mainly positive or negative towards items on those platforms.

Fairness of ranking has been trending in research for the last few years as we increasingly rely on

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32187–32199

algorithms for decision-making (e.g., (Kirkpatrick, 2016; Tramèr et al., 2015; Zehlike et al., 2017; Biega et al., 2018a; Celis et al., 2017; Yang and Stoyanovich, 2017)). Most of these works focus on *group fairness*, stating that individuals in protected groups with comparable utilities should have equal probability of being ranked at the same position. Fairness in ranking has been applied in various scenarios, from search engines (Singh and Joachims, 2018) to recommendation systems (Wang et al., 2023). Surprisingly, little attention has been paid to the collaborative rating scenarios, despite their ubiquitous presence in our everyday lives.

Contribution. In this paper we develop a framework[1] to assess r-fairness in subjective data, that is the fairness of ranking in collaborative rating systems. We also provide insights from its application on real-world platforms such as Yelp and Amazon. We build on the fairness measures proposed in the literature that formulate fairness constraints on rankings in terms of *exposure* allocation (Singh and Joachims, 2018; Kendall, 1938). The main idea of exposure is to assess how a particular ranking balances the fairness of items with the utility that the ranking provides to the users. Intuitively, restaurants in different areas of a city, such as business areas of residential neighborhoods, may display different group exposures. Analogously, some groups could receive fair treatment on Yelp but not on Amazon, according to a specific fairness measure, due to the different contents and policies of the two platforms. In this framework, we address the following research issues. **RI 1** Compare the group exposure at the *item level*, that is, by considering in isolation the rankings of reviews for a single item. **RI 2** Examine how groups are treated in different platforms under different fairness measures. **RI 3** Measure the group exposure at the *query level*, that is, for different queries returning a ranking of items, with each item returning a ranking of reviews.

Methodology. As we do not have direct access to the underlying algorithms and datasets used by the platforms, we developed a pipeline including: *data collection, group assignment, and fairness quantification.* Data collection gathers the rankings available for each search query. Group assignment then assigns each review to a group, using available data about reviewers as well as automatic tools that guess each group based on the text of the review and the available information about reviewers. Fairness quantification implements the exposure allocation measures to quantify two types of r-fairness, dubbed *item-level fairness* and *query-level fairness*. For the former, we consider one item at a time. For the latter, we consider different items together, as returned to a user query such as the one in Figure 1.

Summary of findings. Our key observations can be summarized as follows. In real-world data taken from well-known platforms there is high variability of item-level fairness, with some items showing reviews in a fair ranking while others overexposing certain categories of reviewers (**RI1**). Different platforms treat groups very differently, with some more relying on utility than others for producing the ranking (**RI2**). Different queries over the same platform produce rather different exposures for the same groups of individuals, even when the queries return the same set of items (**RI3**). Additionally, we observed that the ranking of reviews is highly sensitive to the activity of reviewers and that the most active reviewers are not necessarily diverse in demographics. This calls for the design of novel fairness policies in the platforms that incorporate reviewer activity and give voice to less active groups. Our results highlight: (i) the importance of injecting methods for evaluating and assessing fairness in Collaborative Rating Platforms to guarantee that the influence of all reviewer groups is balanced and transparent, and (ii) the effectiveness of our approach to address this problem.

Outline. Section 2 discusses related works about fairness in ranking. Section 3 provides preliminary notions and Section 4 introduces the fairness measures used in this paper. Section 5 reports our experimental results. Section 6 presents a discussion of the findings, highlighting their implications and practical applications. Finally, Section 7 presents conclusive remarks.

2 Related work

Fairness is a rapidly expanding topic, from the seminal works of Dwork et al. (Dwork et al., 2012) discussing the fundamental notion of *fairness in classification* to more recent works applying and extending such concept in a wide variety of data management tasks (Efthymiou et al., 2021; Dong et al., 2023). The notion of r-fairness explored in this paper is mostly related to the notion of *fairness in ranking*, which refers to the principle of incorporating fairness requirements into algorithmic rankers.

[1] Code and experiments can be found at the following link: https://github.com/jermathew/R-Fairness

For an extensive discussion we refer the reader to the two-parts survey (Zehlike et al., 2022a,b) and the survey in (Li et al., 2023) on fairness in recommendation. This concept is crucial in applications such as search engines (Singh and Joachims, 2018) and recommendation systems (Wang et al., 2023) where biased ranking can lead to unfair outcomes. A popular way of achieving fairness in ranking is the optimization of *exposure*, formulated in terms of position allocation, that is, positions in the ranking that are occupied by a certain group (Singh and Joachims, 2018). We address the case of collaborative review platforms, where items (e.g., a restaurant, a product or a place) represent multiple rankings of reviews, and items themselves appear ranked (e.g., best-rated restaurants first) upon a user's query. In this scenario, groups that write high-ranked reviews for high-ranked items have the most influence on online customers' decisions.

Multiple rankings. Differently from our case, works (Singh and Joachims, 2019; Yadav et al., 2020; Diaz et al., 2020) addressing exposure over multiple rankings drawn from a stochastic ranking process and the fairness measures represent the expected exposure of items. More specifically, in (Singh and Joachims, 2019) the stochastic process is the training algorithm for learning a ranking policy. In (Yadav et al., 2020) user feedback is also considered. In (Diaz et al., 2020) the stochastic ranking processes represent information retrieval systems. We also mention (Biega et al., 2018b) that proposes a way to amortize fairness over time. None of these works address the case of nested rankings in collaborative rating systems.

Collaborative filtering and query fairness. Fairness in collaborative filtering has been widely studied in (Tang et al., 2023; Gómez et al., 2022; Yao and Huang, 2017; Shao et al., 2022). However, these works primarily focus on ensuring fairness in item recommendations and do not consider the nested ranking structure of collaborative rating systems. In more detail, (Tang et al., 2023) presents a framework for debiasing recommendations to achieve fairness in top-N recommendations, (Gómez et al., 2022) shows that collaborative recommender systems can exhibit geographic imbalances in provider exposure and addresses that with fairness-aware re-ranking approaches, (Yao and Huang, 2017) provides new fairness metrics and objectives to address unfairness in collaborative filtering recommender systems and finally (Shao et al., 2022) introduces a fairness-aware collaborative filtering model. Similarly, works in (Gao and Shah, 2021; Liu et al., 2024) aim to ensure that search results treat different queries fairly by providing balanced exposure across ranked items.

Fairness pipelines. Since our focus is building a *fairness assessment* tool rather than a fairness optimizer method, we mention the existence of a number of fairness assessment pipelines. However, they are either not applicable to nested rankings or not expressive enough. In (Heuss et al., 2022) the pipeline relies on the assumption that the exposure of an item in a ranked list depends also on the distribution of groups for previous items (inter-item dependencies). Our pipeline also differs from (Xu et al., 2023), which shows that LLMs can exhibit implicit discrimination in recommendations based on user attributes like usernames.

Other works. We finally mention the notion of multi-sided fairness (Abdollahpouri and Burke, 2019), that considers the impact of the ranking on all the stakeholder, including the users who receive the recommendations and the items being ranked.

3 Preliminaries

As illustrated in Figure 2, a subjective database is typically used to represent individual opinions, expressed through reviews over a set of items and it involves three main components: a set of reviewers $\mathcal{C}$ (for Collaborators), a set of items $\mathcal{I}$ (e.g. restaurants on Yelp or products on Amazon) and a set of reviews $\mathcal{R}$ for those items.

Model. We model $\mathcal{C}$ and $\mathcal{I}$ as tables with their own set of identifiers and attributes (such as restaurant location for items and username for reviewers), and $\mathcal{R}$ with a table in which each tuple includes (i) an item I, (ii) the reviewer who wrote a review for I, and (iii) rating information, which can include: a numerical value representing rate given by the reviewer to the item I, a textual description motivating such rating, and the date of the review. A review may also contain other details, such as the number of other reviewers that liked that review. The only assumption here is that a reviewer can write at most one review per item, as it usually happens in collaborative rating systems. Finally, a review r can be associated with a measure of *utility* $u(r)$. Such a measure can be directly available in the subjective dataset or can be derived from attributes of the review. For instance, in Figure 2 we could define the utility u of a review as the number

$\mathcal{I}$ (Items)

Iid	Name	Rate	City
i1	Ace	4	New York
i2	Kixby	3	New York
i3	Porto	4	Boston

$\mathcal{C}$ (Reviewers)

Uid	Age	Followers
u1	28	12
u2	23	0
u3	$\perp$	115
u4	32	2

$\mathcal{R}$ (Reviews)

Rid	Item	Reviewer	Text	Rate	Date	Likes
r1	i1	u1	"Had lunch here last Sunday and it was..."	4	08/12/24	11
r2	i2	u2	'Mixed feelings. The starters were..."	3	09/12/24	0
r3	i3	u1	"Great food and friendly staff..."'	5	10/12/24	3
r4	i1	u4	"The location is great, but..."	3	12/12/24	1

Figure 2: Example of items, reviewers, and reviews

of "Like" it received.

Queries. Subjective datasets are mainly used in online collaborative rating systems to help users find items of interest (Pitoura et al., 2021). As illustrated in Figure 1, a user can issue a query Q to those systems specifying the user's desiderata, such as "Italian restaurants near me", upon which the system outputs an ordered set of k items $\langle I_1, ..., I_k \rangle$. For each item, I_i, the system shows an ordered list of its n_i reviews $\langle r_{i,1}, ..., r_{i,n_i} \rangle$. Both the rankings on items and reviews are sorted according to a *relevance score*, which is usually application-specific and opaque to the end-user, or other more transparent criteria, such as the cost for an item and the date of issue for a review.

Groups. Reviewers can be part of a variety of groups based on (i) their attributes, such as `Age`, and (ii) any information they expose through the review, such as the `Sentiment` towards reviewed items. We denote with $g \subseteq \mathcal{R}$ a group of reviews written by reviewers that share similarities for one or more attributes. For instance, with respect to the subjective dataset in Figure 2, the group $g = \{r_1, r_3\}$ denotes the set of reviews written by positive reviewers.

Ranking. We can think of rankings of items and reviews as fundamentally a *ranking of reviewers*, where reviewers who write high-ranked reviews for high-ranked items obtain more visibility and have therefore more possibility to influence the choices of users. Several methods have been proposed in the literature to address fairness in ranking (Zehlike et al., 2022a,b; Pitoura et al., 2021). One naive application of such methods is to directly use, for each item, the fairness metrics on $\mathcal{R} \bowtie \mathcal{C}$. Although this could highlight fairness issues on specific items, it would not address the fairness of a query Q, where a high-ranked review for an item i can lose its visibility if i is low-ranked in the result of Q.

4 R-Fairness

In this work, we consider r-fairness, that is, fairness with respect to reviewers. Specifically, we consider two types of fairness, referred to as *item-level fairness* and *query-level fairness*. For the former type, we identify basic methods for fairness of individual items, based on metrics from the literature: dubbed Exposure (Singh and Joachims, 2018), treatment (Singh and Joachims, 2018) and rank equality (Kendall, 1938). Then, we introduce variants of such methods to better capture the properties of our scenario. For the latter type, we introduce a principled aggregation strategy to combine individual items' fairness into query-level fairness.

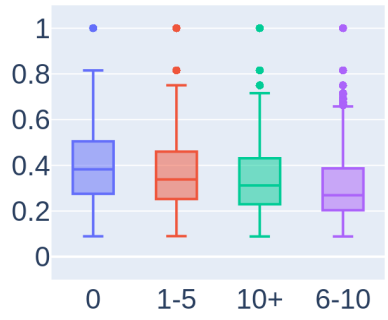
Item-level fairness. We start from the popular fairness measures of exposure defined in (Singh and Joachims, 2018).

Definition 4.1 (Exposure) *The exposure of a group g in a ranking $\mathcal{L}$ is defined as*

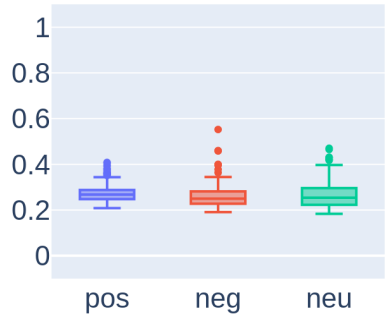
$$Exp_{avg}(g|\mathcal{L}) = \frac{1}{|g|} \sum_{e \in g} Exp(e|\mathcal{L})$$

where $Exp(e|\mathcal{L}) = \frac{1}{log_2(1+pos(e|\mathcal{L}))}$ *and* $pos(e|\mathcal{L})$ *denotes the position of e in $\mathcal{L}$.*

One limitation of the above notion of exposure is that, for large lists, this measure tends to converge to a low value, which becomes indistinguishable for different groups of relevant size. This behavior does not allow the identification of groups that are more visible than others. This is confirmed



(a) Exp_{avg} on Yelp for fans



(b) Exp_{avg} on Amazon for sentiment

Figure 3: Exp_{avg} does not capture variations of exposure in different items because avg flattens for long list.

by the experiments of real-world data reported in Figure 3, in which the exposure is computed for different groups based on the ethnicity (Figure 3(a)) and sentiment (Figure 3(b)). For this reason, we introduce another measure that compares the exposure of the group in a ranking with its best possible outcome. Let $Exp_{best}(g|\mathcal{L})^*$ denote the best exposure a group g can achieve, i.e. the exposure that g would get if members of g were placed in the top $|g|$ positions of the ranked list $\mathcal{L}$:

$$Exp_{best}(g|\mathcal{L})^* = \sum_{i=1}^{|g|} \frac{1}{log_2(1+i)}.$$

We call the new measure *ratio* exposure, defined as follows:

Definition 4.2 (Ratio exposure) *The ratio exposure of a group g in a ranking $\mathcal{L}$ is defined as:*

$$Exp_{ratio}(g|\mathcal{L}) = \frac{\sum_{e \in g} Exp(e|\mathcal{L})}{Exp_{best}(g|\mathcal{L})^*}.$$

Item-level treatment. The next measure takes into consideration the *utility* of (groups of) reviews, under the rationale that it is reasonable for a review in a ranking to have an exposure that is proportional to its utility. For this, let U be an application-dependent function that associates with each review in play its utility, under the hypothesis that U is the same for all queries. We build here on the treatment metric in (Singh and Joachims, 2018) but with Exp_{ratio}. The *utility* of the review according to the opinion of other reviewers. Rating and utility are usually records of values (e.g., for ratings: scores for staff and location plus a descriptive text, in the case of hotels; for utility: positive votes received by other reviewers) and application-dependent. However, the way in which they are implemented does not affect our model.

Definition 4.3 (Ratio treatment) *We define*

$$Treat_{ratio}(g|\mathcal{L}) = \frac{Exp_{ratio}(g|\mathcal{L})}{U(g)}$$

where $U(g)$ is the average utility of members of g.

Item-level rank-equality. Our last item-level fairness metric is a weighted version of the rank equality metric in (Kendall, 1938; Kuhlman et al., 2019). This metric relies on the assumption that there exists a ground truth ranking and measures pairwise disagreements with such ranking using the Kendall-Tau distance. Since a "truly fair" ranking may not be available, we reformulate this metric to distinguish whether the proposed ranking is more or less fair than the utility-based ranking.

Definition 4.4 (Weighted rank equality) *Given two groups g_1, g_2 we define*

$$WRE\ (g_1, g_2|\mathcal{L}) = \frac{1}{N_{g_1,g_2}} \sum_{(e,e') \in g_1 \times g_2} Inv(e, e')$$

where function $Inv(e, e')$ measures the weighted inversion as

$$Inv(e, e') = \begin{cases} W(e, e') & \text{if } W(e, e') < 0 \\ & \text{and } Exp(e) < Exp(e'), \\ -W(e, e') & \text{if } W(e, e') < 0 \\ & \text{and } Exp(e) > Exp(e'), \\ 0 & \text{if } W(e, e') \geq 0 \end{cases}$$

where $W(e, e') = (U(e) - U(e'))(Exp(e|\mathcal{L}) - Exp(e'|\mathcal{L}))$ is the weight of the (possible) inversion of the couple, and $N_{g_1,g_2} = \sum_{(e,e') \in g_1 \times g_2} |W(e, e')|$ is the normalization factor.

Query-level fairness. For query-level, we design a dedicated metric. Assume we are given a query $q(\mathcal{R})$ which is associated to a ordered set of ranked lists $\langle \mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_n \rangle$.

Definition 4.5 (Query-level fairness) *Given a measure $\mathcal{F}$ (e.g., Exp_{cov}) and a group g, we define*

$$\mathcal{F}(g|q(\mathcal{R})) = \frac{1}{count(g, q(\mathcal{R}))} \times \sum_{i=1}^{n} [\mathcal{F}(g|\mathcal{L}_i) \cdot Exp(\mathcal{L}_i|q(\mathcal{R}))]$$

where $count(g, q(\mathcal{R}))$ is the number of items in $q(\mathcal{R})$ in which at least one member of g exists and $Exp(\mathcal{L}_i|q(\mathcal{R}))$ is the individual exposure of the i-th ranked list $\mathcal{L}_i$ in $q(\mathcal{R})$, that is:

$$Exp(\mathcal{L}_i|q(\mathcal{R})) = \frac{1}{log_2(1+i)}.$$

In order to combine the values of $\mathcal{F}(g)$ across different ranked lists $\mathcal{L}_1, ..., \mathcal{L}_n$ we first weight each fairness measure $\mathcal{F}(g|\mathcal{L}_i)$ by the exposure of the corresponding item (e.g. $Exp(\mathcal{L}_i|q(\mathcal{R}))$. This allows us to mitigate the effect of the fairness measure of groups associated with a low-ranked item. Then, we sum up those values and take the average. The latter step is required since each group g may not occur in every item of $q(\mathcal{R})$.

Exp_{avg}, Exp_{ratio}, and $Treat_{ratio}$ have linear complexity in the number of reviews, making them scalable even to larger datasets. Instead, WRE is quadratic in the number of reviews, which requires careful management of the computational load.

Discussion. Our concept of r-fairness differs from the traditional notion of item fairness. While item fairness typically assesses whether certain categories of items, e.g., Italian restaurants or Chinese restaurants in the context of a restaurant recommender system, are overexposed compared to others in the overall ranking, our approach to r-fairness focuses on the reviews within each restaurant and considers them in the context of the restaurant rankings. Analogously, r-fairness is fundamentally different from the position-weighted fairness within a single ranking list. These differences are illustrated in detail with a practical example in Appendix A.

5 Experiments

We run extensive experiments to examine how the various measures of r-fairness vary.

Datasets. We use two real-world datasets, dubbed Yelp and Amazon, the latter collected during these studies and available on demand. `Yelp` consists of business reviews from the `yelp.com`[2] website. `Yelp` collects ≈ 7M reviews from ≈ 2M reviewers on 150K businesses across 11 metropolitan areas in the USA, Canada, and Europe. For each business, we collect name, position, average rating, and category. For each review, we collect the rating, and, if any, the feedback of each review given by others, who can assign a vote among useful, funny, and cool. `Amazon` consists of product reviews from the `Amazon.com` website. We selected 213 items, and collected ≈ 800K reviews from ≈ 608K reviewers, with an average of ≈ 3K reviews per item. Items for `Amazon` are selected from different categories. In both cases, the items returned by the query are ranked by the aggregate rating (e.g., stars) assigned to the item. For each review, we collect text, rating, publication date, and user name. We also collected the feedback of each review given by other reviewers. The reviewers' activity in both Yelp and Amazon platforms is reported in Figure 4. For the Yelp dataset we choose as groups the number of reviews, fans, years since subscription and attitude, where the latter is determined by the average number of stars $avg_\star$ given by the user: 'supporter' if $avg_\star > 4$, 'hater' if $avg_\star < 2$, 'normal' otherwise. For `Amazon` we inferred the sentiment and the gender.

Inference. Reviewers' gender on `Amazon` is predicted from their username using `gender-guesser` [3]. In case of uncertainty, we set a null value. The sentiment is inferred from the texts of the reviews using OpenAI's `gpt-3.5-turbo`.

Methodology. Computing Exp_{avg}, Exp_{ratio} and $Treat_{ratio}$ of group g requires $O(|g|)$ operations. Computing the weighted rank equality requires instead $O(|\mathcal{L}|^2)$ operations in the worst case, corresponding to all possible couples of reviews of g and reviews of other groups. For this reason, in our experiments, we limit the computation of weighted rank equality to the first 100 reviews per item.

5.1 Item-level Exposure

Experiments are reported in Figure 5 for Yelp and Figure 6 for Amazon. The barplots related to a group refer to the items in which the group is present. So the barplots for the attitude 'normal' refer to all the items, being this attitude present in every ranking, while for the attitude 'hater' the measures refer to about 60% of the items.

[2] https://business.yelp.com/data/resources/open-dataset/

[3] https://pypi.org/project/gender-guesser/

#Reviews		Attitude		Years		Fans	
Group	**Activity**	**Group**	**Activity**	**Group**	**Activity**	**Group**	**Activity**
6-20	26.0	Normal	53.7	6-10	54.3	0	52.5
1-5	20.5	Supporter	40.5	10+	30.2	1-5	25.8
100-500	18.9	Hater	5.6	0-5	15.3	6-10	15.6
21-50	17.6					10+	5.9
51-100	11.2						
500+	5.5						

(a)

Gender		Sentiment	
Group	**Activity**	**Group**	**Activity**
Unknown	46.6	Positive	49.5
Female	28.1	Negative	30.8
Male	25.2	Neutral	19.6
Andy	0.03		

(b)

Figure 4: Summary of reviewers' activity on Yelp (a) and Amazon (b). The activity is in percentage.

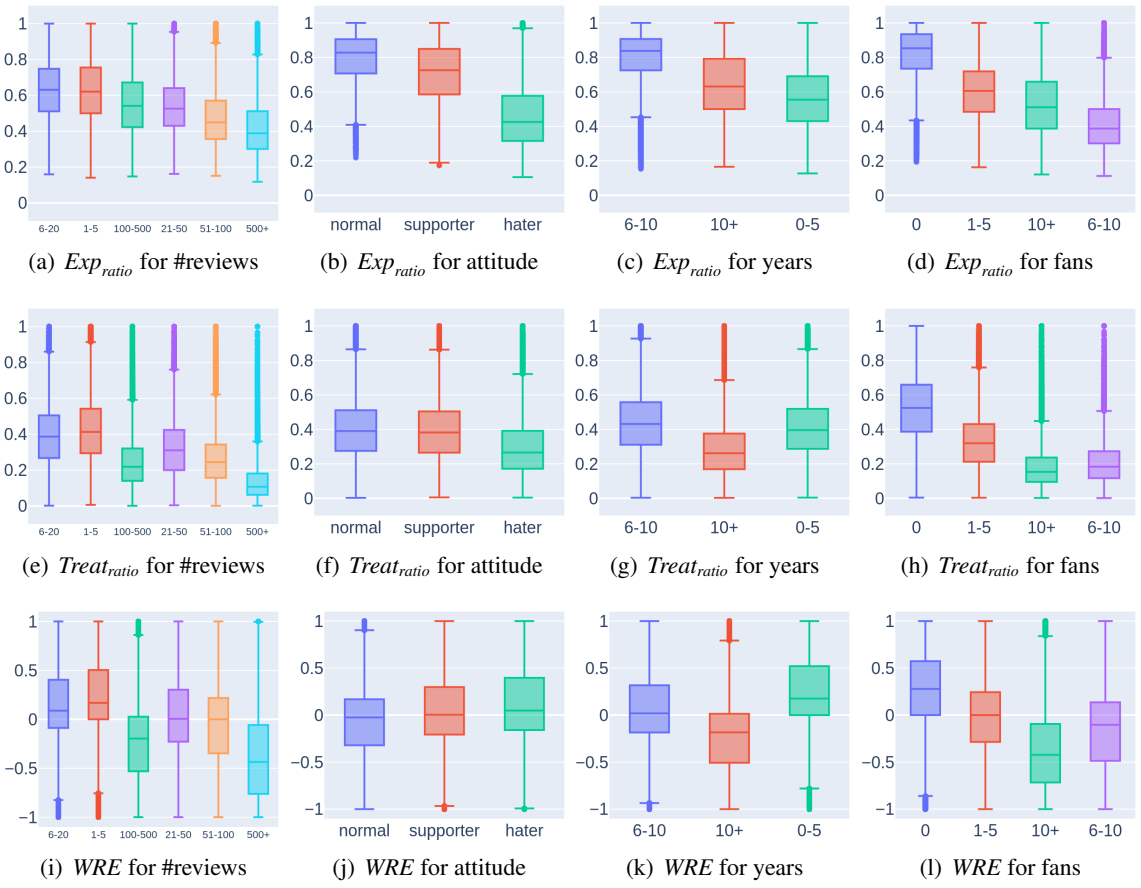
RI1. The barplots highlight that the introduced Exp_{ratio} measure is more discriminating than the Exp_{avg} one. Indeed, with long rankings, the Exp_{avg} tends to flatten out, and all groups have very similar Exp_{avg} values. This is due to the decreasing values of the exposure function. The same happens even if we consider a few reviews per item. On the counterparts, the Exp_{ratio} discriminates among different groups, by giving more relevance to the top positions. Although groups with larger cardinality tend to have higher values of Exp_{ratio} (the larger the cardinality of a group, the higher the probability that a review of the group is in the first positions) there are notable exceptions: the Exp_{ratio} of the group with attitude 'normal' is about two times larger than that of 'hater', although the cardinality of 'normal' is about ten times larger. Another sign of the flattening of Exp_{avg} is that all the groups have almost the same range, while in Exp_{ratio}, $Treat_{ratio}$ and *WRE* different groups have different ranges.

RI2. Regarding **RI 2**, we observe that there are no evident differences between Yelp's and Amazon's datasets when the fairness measure is Exp_{avg} or $Treat_{ratio}$, and the general trend is that the higher the activity, the greater the exposure/treatment. Contrarily, there is a clear discrepancy with the rank equality measure. The weighted rank equality in Amazon is approximately zero for each group. This happens because the rankings are close to being sorted by utility, and all the inversions concerning this sorting are among reviews with close utility values. Instead, on Yelp dataset, the weighted rank equality varies between -1 and 1 for each group. This means that for each group there exist items in which it is over-represented with respect to its utility, and other items in which it is under-represented. Moreover, this implies that Yelp's reviews are not ranked by utility.

5.2 Query-level Exposure

This section addresses **RI 3**, which aims to measure the query-level exposure of groups for different queries by using the formula stated in Definition 4.5. Different queries may return different sets of items or the same set of items with different rankings. We focus on the latter case, indeed, the fairness of groups calculated on different items is expected to be different.

Query generation. In these experiments, we focus on Yelp's dataset because it is more versatile for generating real-world queries. First, we chose a set of 71 Italian restaurants in the same city (Santa Barbara), and then we ordered them according to different queries. The real-world queries reported in Figure 7 are generated by the proximity of the center of the city (*distance asc*), overall evaluation (*stars desc*), number of reviews (*#reviews*) and price (*price asc* and *price desc*). The goal is to show that the query-level fairness of a group changes as the queries change, even if the set of the restaurants is the same. We stress that a direct comparison with the plots in Subsection 5.1 is not possible because we now consider only a subset of restaurants. To make the experiments more realistic, we consider only the first fifteen reviews per restaurant. Indeed, when someone consults rating platforms, he/she usually considers only the first reviews per item (e.g., the first page), and the items are read by following the order given by the platform. It is unrealistic for someone to read all the reviews of an item before reading the reviews of the following.

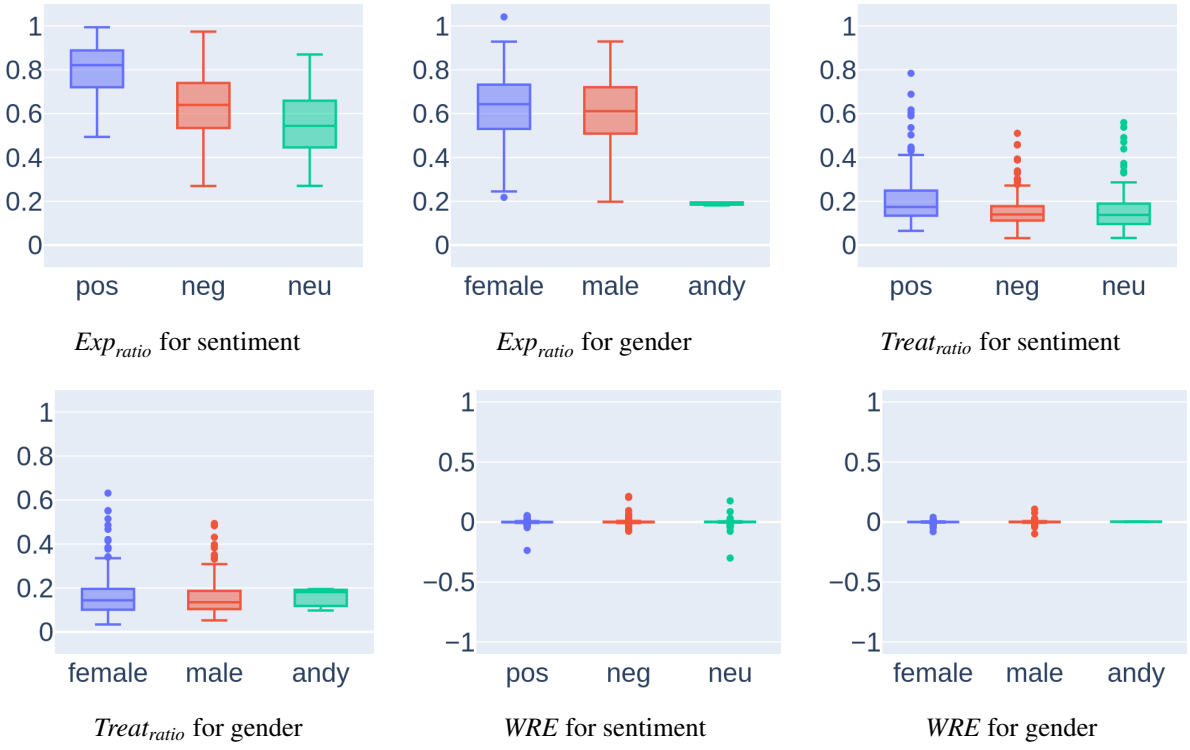


(a) Exp_{ratio} for #reviews (b) Exp_{ratio} for attitude (c) Exp_{ratio} for years (d) Exp_{ratio} for fans

(e) $Treat_{ratio}$ for #reviews (f) $Treat_{ratio}$ for attitude (g) $Treat_{ratio}$ for years (h) $Treat_{ratio}$ for fans

(i) *WRE* for #reviews (j) *WRE* for attitude (k) *WRE* for years (l) *WRE* for fans

Figure 5: Box-plots showing how each group is treated in different items under different fairness measures on Yelp. Different group families on columns and different measures on rows. Each data point is an item.



Exp_{ratio} for sentiment Exp_{ratio} for gender $Treat_{ratio}$ for sentiment

$Treat_{ratio}$ for gender *WRE* for sentiment *WRE* for gender

Figure 6: Box-plots showing how each group is treated in different items under different fairness measures on Amazon. Each data point is an item.

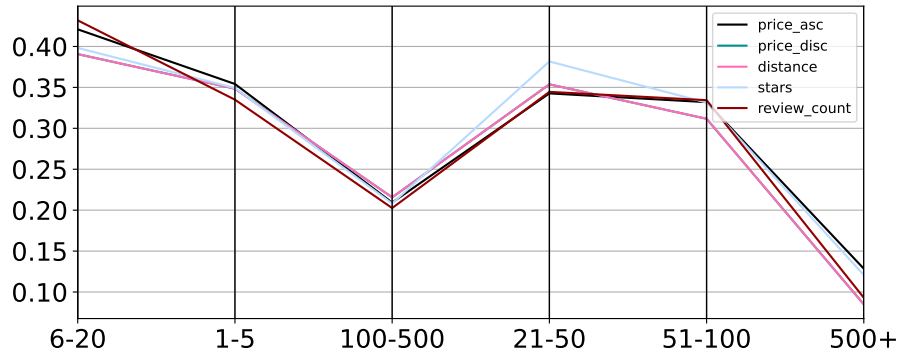


Figure 7: Query-level fairness for $Treat_{ratio}$ on Yelp for #review.

Findings. First of all, we note that the exposure of a group varies depending on the chosen query. Clearly, the range of each group is much narrower compared to the plots in Figure 5, as the total number of reviews considered is considerably smaller (hundreds of thousands compared to about a hundred). Another difference concerns the trend of activity: while in the plots in Figure 5, higher activity almost always corresponded to higher exposure or treatment, this does not apply with query-level fairness in Figure 7: the groups '21-50' and '51-100' have the values of $Treat_{ratio}$ close to or higher than that of the group '1-5', even if their activity are lower. Finally, we observe that, given two groups, we cannot establish *a priori* which one has the higher exposure or treatment, as there are some inversions for different queries. Similar results are obtained with Amazon's dataset and with Yelp's dataset for the other groups and fairness measures. The plots are omitted due to space limitations.

6 Discussion

The primary application of the framework we have proposed it to assess r-fairness, helping different stakeholders, such as platform developers and regulators, to uncover ranking biases. Our analysis has uncovered different cases of unfairness, which opens opportunities for developing novel methods to address these issues and enforce fairness.

Enforcing fairness. R-fairness can be used in principle both for post-hoc optimization, by rearranging both reviews and items for a given query, and for guiding fairness-aware ranking algorithms. Instead of making post-hoc adjustments, such algorithms would dynamically decide the ranking of both items and reviews. One approach for the design of those algorithms would be extending traditional fairness-aware ranking methods such as (Singh and Joachims, 2018) to the nested setting of r-fairness. However, this would require handling non-trivial cases where, for instance, reviewers that are fairly ranked at the item-level remain underexposed at the query-level due to low-ranked items. The design of algorithms that integrate query-level fairness directly into the ranking process represents a valuable direction for future work.

Benefits. Assessing r-fairness can enhance transparency in ranking policies, helping users and stakeholders better understand potential biases. Enforcing r-fairness can prevent systematic underexposure of reviewer groups across queries, even when fairness is maintained within each item.

7 Conclusions

In this paper, we have introduced the notion of r-fairness (fairness with respect to reviewers) in collaborative rating platforms, focusing on how the ranking of subjective data impacts fairness among different reviewers' groups. We have then proposed a comprehensive assessment framework for r-fairnessthat offers insights into the exposure of various reviewer groups. Our experiments with real-world data have shown that the item-ranking process impacts fairness outcomes, often amplifying the visibility of certain reviewer groups over others. The insights gained from our study pave the way for a more transparent and equitable design of collaborative rating systems.

Future work aims at exploring additional platforms and developing more sophisticated fairness metrics that better capture subjective data to support a fairer digital environment where the influence of all user groups is balanced and transparent.

Limitations

While our study introduces a novel pipeline to assess r-fairness in collaborative rating platforms, a few limitations should be acknowledged. First, we relied on automatically inferred demographic gender attributes using the tool `gender-guesser`. While this approach was necessary due to the unavailability of real demographic data, it does introduce uncertainties. The accuracy of these predictions is not guaranteed, which could lead to biases in our fairness assessments. Access to actual demographic data in future research would allow for more accurate and reliable findings. Additionally, our experiments were conducted using data from two platforms: Yelp and Amazon. Although these platforms are representative of collaborative rating environments, the extent to which our findings apply to other datasets and platforms is uncertain. To address this, future work should consider analyzing a wider range of datasets and platforms to strengthen the generalizability of our conclusions. Lastly, while we have included the code used to run our experiments, it does not encompass a fully runnable version of the entire pipeline. Moreover, due to privacy concerns, we have not released the crawled Yelp and Amazon data publicly, as it might inadvertently expose personal data. Although these measures are necessary to protect sensitive information, they do restrict the ability of others to fully reproduce and validate our findings. We plan to explore methods for anonymizing data to improve the reproducibility of our research.

Ethics Considerations

Ethical considerations play a crucial role concerning the use of inferred demographic data like gender. These attributes were predicted using automated tools based on user names, rather than relying on verified information. This can lead to inaccuracies and may inadvertently reinforce harmful stereotypes or biases. Additionally, using predicted demographic data without explicit consent raises ethical questions about making assumptions regarding reviewers' identities. Given that our analysis involves real reviewers data from existing platforms, privacy and data protection are also critical issues. While we strive to protect personal information, there's always a risk that some data could still be identifiable. This risk highlights the necessity of prioritizing reviewers privacy throughout our research, which is why we have chosen not to release the data publicly and to perform only analysis on large aggregations of individuals.

Acknowledgments

We thank the anonymous reviewers for their feedback on this work. We thank Eleonora Bartolomucci and Dario Di Nardo for their contributions to an earlier version of this work. This work was partially supported by the HORIZON Research and Innovation Action 101135576 INTEND "Intent-based data operation in the computing continuum", the SEED PNR Project "Frontiers in Linking records: knOWledge graphs, Explainability and tempoRal data", the Sapienza Research Project B83C22007180001 "Trustworthy Technologies for Augmenting Knowledge Graphs", the PNRR-MUR project PE0000013-FAIR "ENDURANCE: Engineering Data Augmentation for Data Centric Artificial Intelligence", and the PRIN-MUR project 2022XERWK9 "S-PIC4CHU: Semantics-based Provenance, Integrity, and Curation for Consistent, High-quality, and Unbiased data science". Jerin George Mathew was financed by the Italian National PhD Program in AI. The work of Sihem Amer-Yahia is partially supported by DataGEMS, funded by the European Union's Horizon Europe Research and Innovation programme, under grant agreement No 101188416.

References

Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158*.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018a. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018b. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 405–414, New York, NY, USA. Association for Computing Machinery.

L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*.

Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 275–284, New York, NY, USA. Association for Computing Machinery.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602.

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*, pages 214–226.

Vasilis Efthymiou, Kostas Stefanidis, Evaggelia Pitoura, and Vassilis Christophides. 2021. Fairer: entity resolution with fairness constraints. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3004–3008.

Ruoyuan Gao and Chirag Shah. 2021. Addressing bias and fairness in search systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2643–2646.

Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2022. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management*, 59(1):102719.

Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. 2022. Fairness of exposure in light of incomplete exposure estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 759–769.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Keith Kirkpatrick. 2016. Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Commun. ACM*, 59:16–17.

Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The world wide web conference*, pages 2936–2942.

Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. 2019. Subjective databases. *Proceedings of the VLDB Endowment*, 12(11):1330–1343.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48.

Hao Liu, Raymond Chi-Wing Wong, Zheng Zhang, Min Xie, and Bo Tang. 2024. Fair top-k query on alpha-fairness. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 2338–2350. IEEE.

Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28.

Pengyang Shao, Le Wu, Lei Chen, Kun Zhang, and Meng Wang. 2022. Faircf: Fairness-aware collaborative filtering. *Science China Information Sciences*, 65(12):222102.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228.

Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *Advances in neural information processing systems*, 32.

Wang-Chiew Tan. 2020. Unleashing the power of subjective data: Managing experiences as first-class citizens. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, page 3610.

Jiakai Tang, Shiqi Shen, Zhipeng Wang, Zhi Gong, Jingsen Zhang, and Xu Chen. 2023. When fairness meets bias: a debiased framework for fairness aware top-n recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 200–210.

Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2015. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR*, abs/1510.02377.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2023. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *arXiv preprint arXiv:2311.07054*.

Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2020. Fair learning-to-rank from implicit feedback. In *SIGIR*.

Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *SSDM*, page 22.

Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *CIKM*, pages 1569–1578.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022a. Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys (CSUR)*.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022b. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Computing Surveys (CSUR)*.

A Difference between r-fairness and position-weighted fairness

We emphasize that r-fairness cannot be regarded as the position-weighted fairness within a single ranking list, even if we consider, among the various measures we have introduced, the one based on the work of Singh and Joachims (Singh and Joachims, 2018).

We show this point with a toy example. Let us assume that the user issues a query Q and gets as a result a ranked list of two business, B_1 and B_2, which we report below, along with their corresponding set of reviews. For each review we also report the reviewer ID and the corresponding group.

Reviewer	Group
r_1	X
r_2	Y
r_3	X
r_4	Y

Table 1: Business B_1 (ranked in position 1)

Reviewer	Group
r_5	Y
r_6	Y
r_7	X
r_8	X

Table 2: Business B_2 (ranked in position 2)

Let's say we want to compute the r-fairness for group Y with respect to the query Q using Exp_{avg}, which builds on the metric in (Singh and Joachims, 2018). We first need to calculate Exp_{avg} for group Y for each business:

$$\begin{aligned} Exp_{avg}(Y, B_1) &= \frac{1}{2}\left(\frac{1}{log_2(1+2)} + \frac{1}{log_2(1+4)}\right) \\ &= 0.53 \\ Exp_{avg}(Y, B_2) &= \frac{1}{2}\left(\frac{1}{log_2(1+1)} + \frac{1}{log_2(1+2)}\right) \\ &= 0.82 \end{aligned}$$

We then compute a weighted average of the above exposure scores, weighted by exposure of the items (i.e. business) in the ranking:

$$\begin{aligned} F(Y, Q) = &\frac{1}{2} Exp_{avg}(Y, B_1) \frac{1}{log_2(1+1)} \\ &+ \frac{1}{2} Exp_{avg}(Y, B_2) \frac{1}{log_2(1+2)} \\ &= 0.52 \end{aligned}$$

Now, if we instead combine all reviews into a single ranked list and apply the same formula, where we replace Q with a single list of reviews composed by concatenating the reviews from B_1 and B_2 (denoted as $B_1 + B_2$) we would get:

$$\begin{aligned} F(Y, B_1 + B_2) = &\frac{1}{4}\left(\frac{1}{log_2(1+2)} + \frac{1}{log_2(1+4)}\right)\frac{1}{log_2(1+1)} \\ &+ \frac{1}{4}\left(\frac{1}{log_2(1+5)} + \frac{1}{log_2(1+6)}\right)\frac{1}{log_2(1+2)} \\ &= 0.38 \end{aligned}$$

The result differs because, in our metric, the score assigned to each reviewer does not decrease monotonically. For example, Reviewer 5 (r_5) in Business 2 (B_2) has a higher weight than Reviewer 4 (r_4) in Business 1 (B_1). In fact:

$$\begin{aligned} Exp_{avg}(r_5, B_2) &= \frac{1}{2} \cdot \frac{1}{log_2(1+1)} \cdot \frac{1}{log_2(1+2)} \\ &= 0.32 \\ Exp_{avg}(r_4, B_1) &= \frac{1}{2} \cdot \frac{1}{log_2(1+4)} \cdot \frac{1}{log_2(1+1)} \\ &= 0.22 \end{aligned}$$

This reflects the realistic behavior of a user who read Reviewer 5's review before Reviewer 4's review, even though it belongs to a lower-ranked business because such a review appears earlier when looking at the reviews across multiple businesses. In fact, users typically skip around and read only the top reviews from one business before moving to the next. This scenario cannot be captured by combining all reviews into a single list, with weights strictly decreasing with each business' position in the ranking, which describes the unrealistic behavior of a user that always read all reviews from one business before moving to the next.