A reproducibility study of "User-item fairness tradeoffs in recommendations"

Anonymous authors
Paper under double-blind review

Abstract

Recommendation systems are necessary to filter the abundance of information presented in our everyday lives. A recommendation system could merely recommend items that users prefer the most, potentially resulting in certain items never getting recommended. Conversely, a mere focus on including all items could hurt overall recommendation quality. This gives rise to the challenge of balancing user and item fairness. The paper "User-item fairness tradeoffs in recommendations" by Greenwood et al. (2024) explores these tradeoffs by developing a theoretical framework that optimizes for user-item fairness constraints. Their theoretical framework suggests that the cost of item fairness is low when users have diverse preferences, and may be high for users whose preferences are misestimated. They empirically measured these phenomena by creating their own recommendation system on arXiv preprints, and discovered that the cost of item fairness is indeed low for users with diverse preferences. However, contrary to their theoretical expectations, misestimated users do not encounter a higher cost of item fairness. This study investigates the reproducibility of their research by replicating the empirical study. Additionally, we extend their research in two ways: (i) verifying the generalizability of their findings on a different dataset (Amazon books reviews), and (ii) analyzing the tradeoffs when recommending multiple items to a user instead of a single item. Our results further validate the claims made in the original paper. Moreover, the claims hold true when recommending multiple items, with the cost of item fairness decreasing as more items are recommended.

1 Introduction

Recommendation systems have become increasingly popular due to the rapid expansion of the internet and the vast amount of information it holds (Lü et al. (2012)). They are necessary to filter the abundance of information, ensuring users are just shown a small selection of relevant items rather than being overwhelmed by information overload. Recommendation systems learn the preferences of users and recommend items based on these preferences. One approach is to merely recommend items that the user prefers the most, however, this could result in certain items never getting recommended. To prevent this, algorithmic techniques have been developed to improve item fairness in recommendations (Mehrotra et al. (2018); Wang et al. (2021)). However, too much focus on item fairness may lead to users getting less relevant recommendations, thereby compromising user fairness and, in turn, hurting the overall recommendation quality Wang et al. (2023). Therefore, it is important to consider the balance between user and item fairness (multi-sided fairness).

Previous studies have focused on designing algorithms that strive to balance user fairness, item fairness and overall recommendation quality (Burke et al. (2018); Wang & Joachims (2021)), yet little research exists on the effects and the tradeoffs of multi-sided optimization. Greenwood et al. (2024) address this gap by developing a theoretical framework to analyze these effects and tradeoffs in the context of multi-sided fairness-constrained optimization. They theoretically concluded two phenomena: (i) the cost (i.e. decline in user utility) of item fairness is low when the users have diverse preferences, and (ii) item fairness may have a high cost on users whose preferences are misestimated, such as users who are new to the system (i.e. cold start users). By creating their own recommendation system for arXiv preprints, they empirically measured these phenomena. Similar to their theoretical findings, they found the cost of item fairness to be low for

users with diverse preferences. However, the cost of having misestimated users was already so high such that imposing further item fairness constraints on them did not further increase costs.

Our research focuses on reproducing and extending the experiments of Greenwood et al. (2024). Firstly, we replicate the same experiments using their published code on the same arXiv dataset and Semantic Scholar data. Secondly, we extend the research by examining whether the proposed claims hold for a new dataset consisting of Amazon books reviews. This aims to evaluate the generalizability of the claims regarding multi-sided fairness tradeoffs across different domains, such as book e-commerce. Finally, we further extend the research by analyzing the effects of user-item fairness tradeoffs when recommending multiple items. Greenwood et al. (2024) experiment with recommending only one item, however, recommending multiple items reflects better a real-world scenario as recommendation systems typically suggest multiple items to users rather than a single one. To support this scenario, we adapted their theoretical framework and performed the same experiments.

This paper successfully reproduces the empirical findings of the experiments on the original and newly introduced dataset further validating the claims made by Greenwood et al. (2024). Moreover, we validate the claims as well when recommending multiple items. However, we observed a decreased cost when imposing item fairness constraints as the number of items recommended increases. The GitHub repository containing the code discussed in this paper can be found at the following link: `https://anonymous.4open.science/r/Re_User-Item_Fairness_Tradeoffs_in_Recommendations-7B9A`.

2 Scope of reproducibility

This research investigates the empirical claims concerning the identified phenomena by Greenwood et al. (2024). Their work focuses on the implications of including item fairness constraints alongside user fairness in a recommendation system for arXiv preprints. The fairness definition they adopt follows an individual egalitarian approach, meaning maximizing the utility of the worst-off individual. By employing an in-processing technique for fairness that integrates item fairness directly into the recommendation algorithm, they discovered the following phenomena:

- Discovered phenomenon 1: *"When user preferences are diverse, there is 'free' item and user fairness."* Free fairness implies that item fairness can be imposed with minimal cost to the user while it is beneficial for the items.

- Discovered phenomenon 2: *"Users whose preferences are misestimated can be especially disadvantaged by item fairness constraints."*

Building on these theoretical insights, they attempted to empirically validate these phenomena using an arXiv recommender system, leading to the following claims:

- Empirical claim 1: *"More homogeneous groups of users have steeper user-item fairness tradeoffs – as theoretically predicted, diverse user preferences decrease the price of item fairness."* Notably, the price of fairness becomes substantial only when strict item fairness constraints are imposed.

- Empirical claim 2: *"The 'price of misestimation' is high (users for whom less training data is available receive poor recommendations), but on average item fairness constraints do not increase this cost."*

This study seeks to validate these empirical claims by reproducing the experiments of Greenwood et al. (2024) on the arXiv dataset. To further explore the robustness of their claims, this paper will extend the original work by examining whether the proposed claims hold for the Amazon books reviews dataset. Additionally, this research expands the original study by increasing the number of recommended items, allowing for a deeper analysis of the user-item tradeoff.

3 Methodology

For the reproduction of these empirical findings, we lean on the theoretical framework provided by the paper, and use the GitHub repository of the original paper [1]. We further extend their research by reproducing their claims on a new dataset and by examining the effect of user fairness when the number of recommended items per user is increased.

3.1 Theoretical framework

Greenwood et al. (2024) propose a recommendation system that aims to balance user fairness, item fairness and overall recommendation quality, leading to a multi-sided fairness problem. The principle of egalitarian (social) fairness (i.e. the utility of all agents is given by the utility of the worst-off agent) is used to quantitatively define both user and item fairness, where the balance between the two is parametrized by fairness level γ. This leads to the following optimization constraint:

$$\begin{aligned} U^*(\gamma, \omega) = \max_\rho \min_i U_i(\rho, \omega) \\ \text{subject to} \quad \min_j I_j(\rho, \omega) \geq \gamma \max_\phi \min_j I_j(\phi, \omega) \end{aligned} \tag{1}$$

Here, ρ denotes a set of parameters referred to as *recommendation policy* (ϕ is the set of parameters for the item fairness optimization problem, ρ is the set of parameters for the user fairness optimization problem and – since subjected to the item fairness constraints – to the entire optimization problem). For each user i, $\sum_j \rho_{ij} = 1$, with all $\rho_{ij} \in [0, 1]$. Utility matrix ω is calculated using the cosine similarity between user and item embeddings, where $\omega_{ij} > 0$ denotes the utility of recommending item j to user i. $U_i(\rho, \omega)$ denotes user i's utility from ρ normalized by the utility they would receive from being recommended their best match, and $I_j(\rho, \omega)$ denotes item j's utility from ρ normalized by the utility it receives if it is recommended to every user. Their values are given by:

$$\begin{aligned} U_i(\rho, \omega) &= \frac{\sum_j \rho_{ij}\, \omega_{ij}}{\max_j \omega_{ij}} \\ I_j(\rho, \omega) &= \frac{\sum_i \rho_{ij}\, \omega_{ij}}{\sum_i \omega_{ij}} \end{aligned} \tag{2}$$

In words, Equation 1 seeks to find the parameters ρ for which the minimum normalized user utility is maximized (i.e. most fair according to egalitarian fairness), while the minimum normalized item utility is at least a fraction γ of its optimal value. Note that for $U^*(\gamma = 0, \omega) = 1$ since the optimal recommendation policy ρ would deterministically recommend each user their most preferred item.

The price of the item fairness constraint can be calculated as the normalized decrease in user fairness when subjected to the item fairness, given by the formula:

$$\pi^F_{U|I}(\gamma', \omega) = \frac{U^*(\gamma = \gamma', \omega) - U^*(\gamma = 1, \omega)}{U^*(\gamma = \gamma', \omega)} \tag{3}$$

Utility matrix ω can be estimated based on the embeddings of users and items to be recommended, however, it should not be ignored that these utilities are mere estimations. This price of misestimation can be computed by:

$$\pi^M_U(\gamma', \omega, \hat{\omega}) = \frac{U^*(\gamma = \gamma', \omega) - \min_i U_i(\hat{\rho}(\gamma'), \omega)}{U^*(\gamma = \gamma', \omega)} \tag{4}$$

Here, ω denotes the true utility matrix, $\hat{\omega}$ denotes the estimated utility matrix, and $\hat{\rho}(\gamma')$ denotes a recommendation policy that solves the recommendation problem (Equation 1) with respect to the misestimated utilities, i.e. $\hat{\rho}(\gamma')$ solves $U^*(\gamma', \hat{\omega})$.

[1] Accessible via https://github.com/vschiniah/ArXiv_Recommendation_Research

For a new user without known preferences (a so-called cold start user), the platform estimates their preference as the average of preferences of existing users. This would, without item fairness constraints, result in the platform recommending generally popular items. However, since an average is taken over a large diversity of items, the preference of this individual new user for a particular item is less strong than users who have more data in their embeddings and prefer a certain item. This encourages the recommendation of generally unpopular items when item fairness constraints are introduced. Therefore, theoretically one would expect user fairness to be worsened more for cold start users as γ increases.

3.2 Datasets

In this paper, we use the arXiv dataset[2] to reproduce and extend the work of Greenwood et al. (2024). Furthermore, we use the Amazon Books Reviews dataset[3] to investigate how their findings translate to a different domain.

3.2.1 Original dataset: arXiv

The arXiv dataset consists of 2,639,142 papers, containing papers published on arXiv from 1991 up until now[4]. Each entry covers a paper, which is defined by fourteen attributes. Just as Greenwood et al. (2024), we only consider papers that possess one or more Computer Science categories and dropped all remaining papers. This leaves us with 707,763 papers. The train and test set are created based on a paper's year of publication: all papers published before 2020 are selected for the train set, and all papers from 2020 are selected for the test set, resulting in 255,138 and 65,948 papers respectively. The category distributions of these sets are visualized in Appendix A.1 Figure 6. The distribution of our dataset slightly differs from the ones presented by Greenwood et al. (2024), shown in Appendix A.2 Figure 7.

Additional information on all test set papers is acquired[5] through API calls to Semantic Scholar[6], using the Semantic Scholar corpus-IDs. This process, including the API calls and the retrieved additional information, is further detailed in Section 3.3.1. All papers where the Semantic Scholar corpus-ID could not be acquired are discarded, leaving 26,254 papers in the test set for further use. To match the size of the test set with the original authors, we sampled 14,307 papers such that the proportions of all subcategories remained unchanged. We kept the train set size unchanged as it is unclear how the dataset sizes by Greenwood et al. (2024) were obtained, and explicit sampling could lead to significant performance cost[7].

3.2.2 New dataset: Amazon books reviews

The Amazon books reviews dataset consists of 3,000,000 reviews where each review is accompanied by attributes of the corresponding book, including title, author, description, and year of publication. The reviews span from 1996 until the end of 2013 and include 212,404 unique books. Books that missed either a description or the year of publication were excluded from our analysis, as the description is essential for creating the user embeddings, and the year of publication is required to split the dataset. After cleaning the data, we retained 78,571 unique books. For the training set, we considered books published before 2007. This training set was used to create embeddings for the reviewers, who represent the users in the recommendation system. It contains 30,506 books, 475,382 reviews, and is contributed to by 237,310 unique reviewers. The test set included books published from 2007 until the end of 2011, resulting in it containing 14,548 books. The test set is the set of books to be recommended. We chose these years to align the test set size with the original paper.

[2]Accessible via https://www.kaggle.com/datasets/Cornell-University/arxiv

[3]Accessible via https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

[4]The dataset is updated weekly. These numbers correspond to the time of writing this, January 11$^{\text{th}}$, 2025.

[5]Since these are the ones on which recommendations will take place. The train set is merely used to create embeddings for authors, as described in Section 3.3.1.

[6]More information about the Semantic Scholar API can be found via: https://www.semanticscholar.org/product/api

[7]E.g.: a possibility exists all papers of certain authors were not yet present in the dataset. Explicit sampling from all papers could therefore unwillingly worsen the performance over all authors.

3.3 Experimental setup

To replicate the original study, the README files in the paper's repository were taken as a leading reference, in coordination with what was described in the paper itself. The code provided largely required minor adjustments, however, some files were missing which we reintroduced to fully run the original pipeline. This pipeline starts with general data preparation where the train and test set are separated and additional information about papers is gathered through API calls. Subsequently, embeddings for papers and authors, and associated similarity scores are calculated. After all preparation, logistic regression is performed to evaluate whether calculated similarity scores are a good measure of successful recommendations. Finally, we run the experiments to validate the main claims of the paper. Below we describe the detailed experimental set-up of reproducing the results on the arXiv dataset (Section 3.3.1) and of our proposed extensions: (i) validating the claims on a different dataset (Section 3.3.2), and (ii) examining the effect of user fairness when multiple items are recommended for each user (Section 3.3.3)

3.3.1 Original setup: arXiv dataset

We maintained the below pipeline of the original paper[8] with minor adjustments.

General data preparation Identically to Greenwood et al. (2024), we retrieved all papers of the arXiv dataset as described in Section 3.2.1. and discarded all papers that do not possess Computer Science as a main category. For this, we added an extra script to produce all mappings from category-ID to main and subcategory, as this file seemed missing. Furthermore, we moved the script to gather additional features of the test set forward; according to the original README, this script is performed later, leading to problems in later scripts, hence we moved it forward. Furthermore, we introduced a script to sample the subcategories proportionally.

Logistic regression data preparation Further additional API calls to the Semantic Scholar API are made to obtain information for logistic regression regarding: all authors that contributed to any paper in the test set, all papers that have a paper from the test set in their citation, and all citations of papers in the test set. Since we could not obtain a Semantic Scholar API key (due to a pause in issuances by high demand), we were restricted to the 1,000 API calls per second shared across all unauthenticated requests to the Semantic Scholar servers[9]. This led to many unsuccessful responses and eventually to the API server blocking our requests, as described in Appendix A.3. This forced us to significantly reduce the test set available for logistic regression down to $1/12^{\text{th}}$ of our initial test set, covering 2,188 papers, approximately $1/6.5^{\text{th}}$ of the number of papers in the test set of Greenwood et al. (2024). To mitigate the first problem specifically, we retried all failed responses four more times before moving on to the next paper.

Embeddings and similarity scores Paper embeddings are calculated for both train and test sets by considering each paper's title, abstract and categories, removing stopwords, and letting a TF-IDF vectorizer create embeddings. The embeddings for the users (authors) for which we recommend papers, are created by considering all embeddings of their published papers present in the train set. Utility matrix (ω) is constructed by computing ω_i for each author and paper by the cosine similarity between their embeddings. This results in a similarity matrix for all their previously published papers and each paper j in the test set. ω_{ij} is determined by selecting the maximal similarity score present for paper j. In the original code, author embeddings are created for 1,000 uniquely sampled authors from the test set. Since no embedding can be created for authors that do not appear in the train set, we modified the code to create embeddings for all authors present in both sets, after which we sample 1,000. The results of the original approach are shown in Appendix A.4, and the results of our approach are discussed in Section 4. Furthermore, we adjusted how the calculated similarity score is stored. In the original codebase, the similarity score is stored by overwriting all previously stored similarity scores for each author. We corrected this by writing the similarity score to the data entries which belong to the author it relates to.

Logistic regression To evaluate the recommender system, we performed logistic regression. Since the original repository did not include code for this regression, we implemented it based on the details provided in

[8] As explained in Section 6, our approach is in line with a previous version of the original paper.
[9] As stated here: `https://www.semanticscholar.org/product/api`

the paper. We used the same sample size as the original paper, namely 1128. We added all recommendations per user for our logistic regression, instead of only the top-1 recommendations. The logistic regression was conducted between the similarity score and whether the author cited the recommended paper. To compute the similarity score, we used the maximum TF-IDF score. To ensure robustness, we performed logistic regression three times with random seeds 42, 999, and 123. We calculated the coefficient, standard error, Z-value, and P-value.

Experiments We added a script to generate the data source file for all experiments based on the README file of the original authors, since this script was missing. For each experiment, the utility matrix ω is computed again analogously. The first experiment examines the difference in user-item fairness tradeoff between heterogeneous and homogeneous users. For heterogeneous users, we sample 40 authors and 20 papers out of the entire test set. Then, for 50 values of γ between 0 and 1, U^* is calculated and plotted. In total 10 curves are calculated, after which the mean and (two) standard deviations are plotted. For homogeneous users, all authors are first grouped into 25 clusters. For this clustering, the original code casts the sparse into dense embeddings resulting in significantly more memory usage. We ensured the sparse representations were maintained to significantly save memory without change in result. The process of creating the remainder of the graph is identical to that of the heterogeneous graph, except that for each curve all authors of the same cluster are sampled, which is expected to be 40. The second experiment examines the difference in user-item fairness tradeoff between users of which preference data is present and cold-start users. The latter category is constructed by treating 10% of the sampled users as cold-start users by removing their embedding. Then again, for 50 values of γ between 0 and 1, U^* is calculated and plotted.

3.3.2 Extension: Amazon books reviews dataset

In our recommender system, user embeddings are defined by all book embeddings to which a reviewer made a review prior to 2007. In the original paper, user embeddings were constructed using the title, abstract, and category of each item. To align with this, we used the title, description, and category of each book to create our user embeddings. To evaluate our recommender system, we performed logistic regression following the methodology described in section 3.3.1, adjusted for the Amazon books reviews dataset. The logistic regression was conducted between the similarity score and whether the user actually left a review for the recommended book. The similarity score reflects the similarity between the books reviewed by the user before 2007 (the user embedding), and the possible recommended books from the test set.

For the experiments, we sampled 1,000 users, consistent with the original paper. Moreover, we held all experimental parameters, for conducting the experiments, constant to ensure comparability with the original paper. Specifically, we tested 50 values of γ between 0 and 1, clustered the 1,000 users into 25 clusters using the k-means algorithm for the homogeneous population for the first experiment, and again treated 10% of the population to be misestimated for the second experiment.

3.3.3 Extension: Multiple recommendation per user

We extended the original study on the arXiv dataset by recommending more than one item to each user as it simulates a real scenario better. In real-world recommendation systems like the ones used by streaming services such as Netflix, users are often provided with multiple recommendations instead of a single item. In the paper, we noticed that the optimal policy ρ, which maximizes the minimal user utility (Equation 1), under the constraint $\sum_j \rho_{ij} = 1$, leads to the following: for each user i, the item with index h—which generates the highest utility score $\omega_{i(j=h)}$—is allocated $\rho_{i(j=h)} \to 1$, while all other items have $\rho_{i(j\neq h)} \to 0$. This is because we aim to assign the highest recommendation weights to the items with the greatest utility to achieve the maximum expected utility. When generalizing to recommending multiple items (k) in the optimal policy ρ, the k items with the highest utility should correspond to values in ρ that approach 1, while the remaining items should have values that converge on 0. This can be achieved by the following updated constraint:

$$\sum_j \rho_{ij} = k \tag{5}$$

Thus, for each user i, the allocated recommendation weights must sum up to k, where k controls the number of recommended items. In this experiment, k is set to 3 and 5, to analyze the impact of recommending 3 and 5 items, respectively. As recommending multiple items automatically increases user utility, the normalization of the user utility should be adjusted to maintain an equivalent range for the user utilities across different k values. Therefore, we adjusted the normalization for user i's utility in equation 2 to:

$$U_i(\rho, \omega) = \frac{\sum_j \rho_{ij}\ \omega_{ij}}{\sum_{j \in K}\ \omega_{ij}}, \text{where } K \text{ is the set of the } k \text{ items with the highest utilities for user } i \tag{6}$$

This adjustment ensures that user i's utility is normalized with respect to their best k recommendations. Hence, the only differences from the original setup in section 3.3.1 are setting k to values other than 1 and modifying the normalization.

3.4 Computational requirements

For the experiments described in Section 3.3, we used a node containing nine cores of the Intel Xeon Platinum 8360Y, an NVIDIA A100 GPU, and 60GB of DRAM. In total all computing time took 99 hours. We calculated the CO_2 emission to be approximately 12 kg CO_2[10].

4 Results

In this section, we first briefly review the original results of the paper. Then, we present our findings from the reproducibility study on the arXiv dataset using the original setup. Furthermore, we discuss the results of the two extensions. Since no seed is set when sampling authors and papers for the curves, the curves vary slightly every run due to randomization.

4.1 Original results

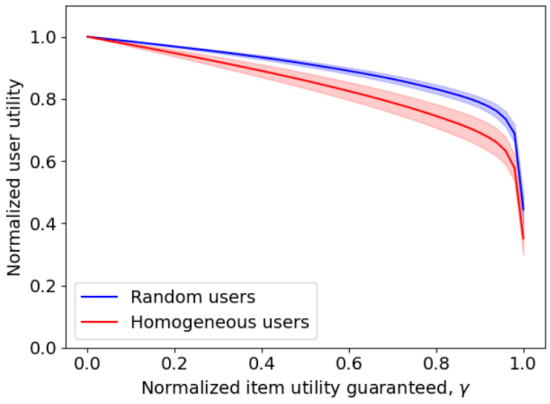
This section briefly reviews the key findings of the original paper by Greenwood et al. (2024). First, the authors validated their arXiv recommender system by performing a logistic regression on the similarity score and whether the paper was actually cited by the author. Table 1 shows a highly significant positive coefficient, confirming a reliable recommender system.

The results of their experiments are presented in Figure 1. Figure 1a shows that item fairness imposes a higher cost on homogeneous users compared to diverse users, consistent with *empirical claim 1.* This cost remains low except when γ reaches 1, matching strict fairness. Figure 1b demonstrates a substantial cost of misestimation. This cost is so high that item fairness does not have a negative impact on user fairness, corresponding to *empirical claim 2.*

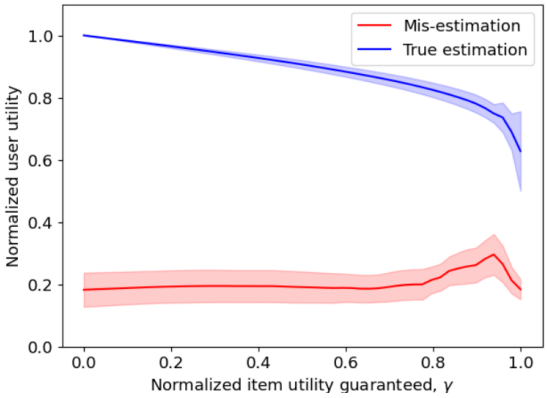
	Coefficient	Standard Error	Z-value	P-value
Similarity score	12.4100	0.058	212.178	0.000

Table 1: Logistic regression results to validate the recommender system with the arXiv dataset (Greenwood et al., 2024).

[10]Taking the average Carbon Efficiency of our country from Moro & Lonza (2018) and the calculation tool of Lacoste et al. (2019)



(a) Homogeneous versus diverse users on the arXiv dataset.

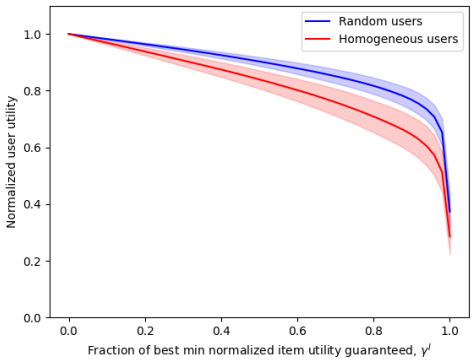


(b) With and without misestimation on the arXiv dataset.

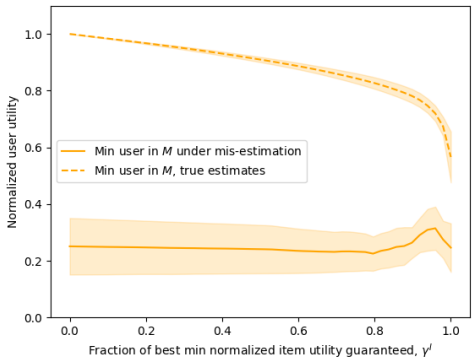
Figure 1: The empirical findings of Greenwood et al. (2024) between user and item fairness, where γ describes the item fairness constraint as explained in Section 3.1.

4.2 Original setup: arXiv dataset

This section presents the results of our reproducibility study on the arXiv subset. Due to API rate limits, we ran the logistic regression for the arXiv dataset on the subset of 2,188 papers instead of 14,307 papers. To ensure that using a smaller subset for the experiments would not impact the results, we conducted the experiments on both sets. Since results on the total set (Appendix A.5 Figure 9)) and on the smaller subset (Figure 2) show a similar trend and no significant differences, we decided to present results on the subset of 2,188 papers for consistency with the dataset on which logistic regression is performed.



(a) Homogeneous versus diverse users on the arXiv dataset.



(b) With and without misestimation on the arXiv dataset.

Figure 2: Empirical findings between the normalized user and item utility. γ describes the item fairness constraint.

Table 2 shows the results of the logistic regression. The coefficient of 6.21 suggests a strong positive relationship between the similarity score and whether or not the author cites the paper in the future. This let us to conclude that recommending items based on similarity score is valid. The coefficient is however lower than what is reported in the original paper (see Table 1), which might be caused by the faulty storage of similarity scores as described in Section 3.3.1.

	Coefficient	Standard Error	Z-value	P-value
Similarity score	6.2092 ± 0.3429	0.1650 ± 0.0093	37.8554 ± 4.0966	0.000

Table 2: Logistic regression results using three random seeds to validate the recommender system with the arXiv dataset.

Figure 2a demonstrates the comparison between heterogeneous and homogeneous users. The resulting curves are similar to those observed in the original paper. For moderate item fairness constraints, so $0 \leq \gamma \leq 0.9$, we observe a slight negative effect on user utility by increasing item fairness. As γ approaches 1, the tradeoff becomes significantly steeper. Homogeneous users exhibit a higher price of fairness, consistent with *empirical claim 1* in the original study.

Figure 2b presents the results of the comparison between users with and without misestimation (i.e. users without and with prior known preferences). The curves of this experiment closely resemble those of the original paper. A similar drop in user fairness is present as for random users in Figure 2a, while user fairness stays consistent, with even a slight increase, for cold start users when increasing γ. This shows, like stated in *empircal claim 2* in the original paper, that the cost of misestimation is already so high that it is not worsened with item fairness constraints as we analyzed in Section 3.1.

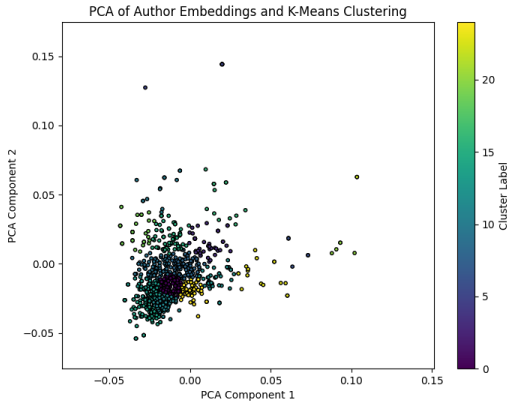
4.3 Extension: Amazon books reviews dataset

This section presents the results of the original experiments conducted on the Amazon books reviews dataset. Table 3 presents the results of the logistic regression. The coefficient of 15.05 suggests a strong positive relationship between the similarity score and whether or not a reviewer leaves a review for a book in our recommendation selection. Thus, we concluded our recommender system to be valid.

	Coefficient	Standard Error	Z-value	P-value
Similarity score	15.05 ± 0.78	0.365 ± 0.012	41.23 ± 1.04	0.000

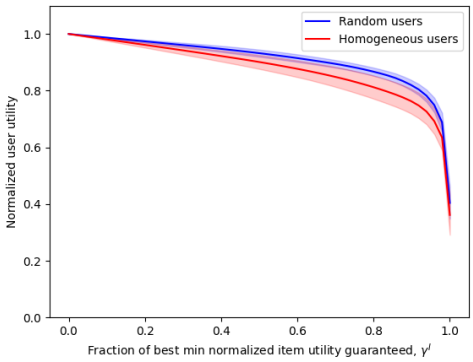
Table 3: Logistic regression results using three random seeds to validate the recommender system with Amazon books reviews dataset.

(a) arXiv dataset.

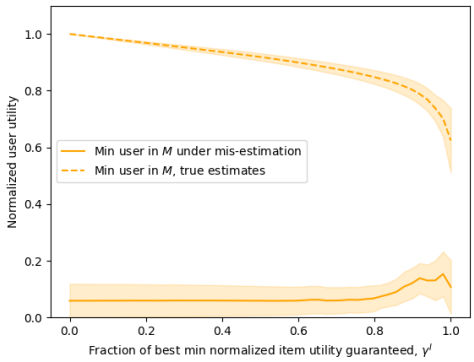


(b) Amazon books reviews dataset. We left out approximately 5 data points that were outliers.

Figure 3: Scatterplots for the test set of both datasets depicting 25 clusters and two PCA components.



(a) Homogeneous versus diverse users on the Amazon books reviews dataset.



(b) With and without misestimation on the Amazon books reviews dataset.

Figure 4: Empirical findings between the normalized user and item utility, where γ is the item fairness constraint, using the Amazon books reviews dataset.

Figure 4a demonstrates the comparison between heterogeneous and homogeneous users. The resulting curves are similar to those observed in the original paper. Homogeneous users exhibit a higher price of fairness for the new dataset, which aligns with *emperical claim 1* reported in the original paper. Notably, the two curves are closer to each other compared to the original study. A possible explanation for this difference is that homogeneous users in the Amazon book reviews dataset may be clustered more heterogeneously. To test this, we computed the silhouette scores for clusters in both datasets, a metric measuring within-cluster similarity relative to other clusters, with lower scores indicating greater heterogeneity. However, we observed that both datasets yield similar scores, suggesting this hypothesis is not supported. Further details can be found in Appendix A.6 Table 4. To further investigate why the two curves are closer to each other, we plotted the authors and the clusters to which they belong. Figure 3 shows that a possible explanation can be sought in the random users. When looking at the scatter plots, all user embeddings are gathered in a small interval of values, as opposed to the arXiv dataset which is more spread. This leads to a higher probability of more similar users being randomly sampled in the Amazon dataset compared to the arXiv dataset. The more homogeneous nature of the random users in the Amazon dataset can lead to more similar user utility between the two curves. This can explain the two curves being closer to each other compared to the original study on the arXiv dataset.

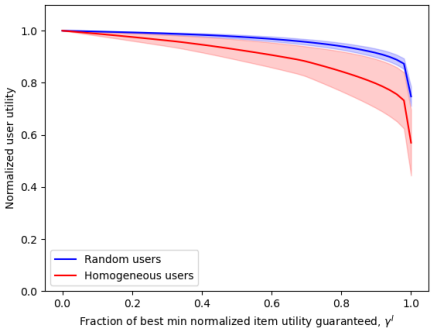
Figure 4b shows the comparison between users with and without misestimation. The curves of this experiment closely resemble those of the original paper, further supporting the robustness of the proposed *empirical claim 2* that item constraints do not increase the cost of misestimation.

4.4 Extension: Multiple recommendation per user

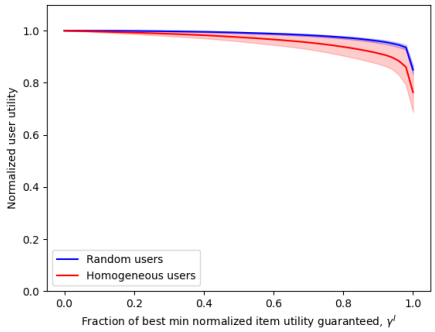
This section displays the experimental results of recommending multiple items, conducted on the arXiv subset of 2,188 papers. Figures 5a and 5b show the differences between homogeneous and heterogeneous users when recommending 3 and 5 items, respectively. The observed curves have a similar trend to those in the original paper, where homogeneous users encounter a higher cost of item fairness than heterogeneous users, and the curves become substantially steeper when γ approaches 1. So, this is consistent with proposed *empirical claim 1*. However, a notable observation is that both curves become less steep as k increases. Intuitively, this makes sense as the negative impact of an individual recommendation is mitigated by considering a larger variety of items. Due to this enlarged variety in recommendations, the gap between recommendation differences for homogeneous and heterogeneous users decreases.

Figures 5c and 5d illustrate the disparity between misestimated and truly estimated users when recommending 3 and 5 items again. As expected, the line of misestimated users shifts upwards for a higher k since misestimations are less problematic when more items are recommended. In this context, when recommend-

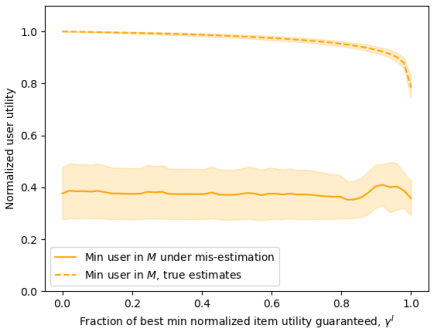
ing more arXiv preprints, the likelihood of the user encountering a relevant paper increases, which leads to a higher utility. Furthermore, we observe that item constraints do not increase the cost of misestimation, which aligns with *empirical claim 2* in the original paper.



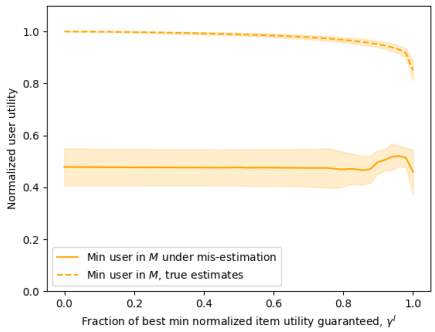
(a) Homogeneous versus diverse users for $k = 3$.



(b) Homogeneous versus diverse users for $k = 5$.



(c) With and without misestimation for $k = 3$.



(d) With and without misestimation for $k = 5$.

Figure 5: Empirical findings between the normalized user and item utility for $k = 3$ and $k = 5$, where γ is the item fairness constraint.

5 Discussion and conclusion

We successfully reproduced the empirical findings described in the original paper by Greenwood et al. (2024). First, we performed a logistic regression on a subset of the dataset, confirming that similarity scores provide a valid estimation for future citations and thus user preferences. Then, we reproduced their findings by conducting experiments on this subset. Furthermore, we extended our experiments to a different dataset, the Amazon Books Reviews dataset, and observed similar findings as in the original paper, further validating their claims. Our results validate *empirical claim 1*: the cost of item fairness is lower for users with diverse preferences and this price is low, except for when γ approaches 1. We similarly validate *empirical claim 2*: the cost of item fairness is not higher for users whose preferences are misestimated. Even when increasing the number of recommended items, homogeneous users experience a higher cost of item fairness. However, both curves become less steep as the number of recommended items increases, suggesting that increasing the number of recommended items seems to lessen the tradeoffs between user and item fairness. Additionally, all experiments indicate that item fairness constraints do not increase the cost of misestimation.

All the results demonstrate a high standard deviation, which could be overcome by increasing the number of items and users. In a newer version of the paper, this adjustment is implemented, but due to time and computational limitations, we were unable to do this ourselves. However, it would be interesting to explore the effect of recommending multiple items when the sample size is bigger. In our experiments, the number

of possible items to be recommended is 20 and the actual number of recommended items is 1, 3, and 5. Recommending 5 out of 20 papers makes up a substantial part of the entire recommendation pool, which could be the reason for the fairness constraints lessening quite a bit and the decreased cost of misestimation. Future work could investigate whether this effect weakens when the recommended items are a smaller part of the total possible items to be recommended.

Overall, in line with the original findings, our results suggest that recommendation system designers should keep in mind that it is beneficial to have a diverse user population and that item fairness constraints should be imposed on the entire population rather than subgroups.

6 Final remarks

How we deviated from the original authors. Due to time constraints, limited computing resources and the computational complexity of optimizing the convex problem, we were confined to reproduce the older original NeurIPS version of the paper, accessible via `https://neurips.cc/virtual/2024/poster/94638`. This version deviates from the most recent version in three aspects. Firstly, it draws smaller samples of papers (20 instead of 200) and authors (40 instead of 500) for each curve. Secondly, it first samples 1,000 papers to sample each curve from. Both the old and new version of the paper share very similar graphs and identical conclusions. Thirdly, for creating homogeneous authors it creates more clusters (25 instead of 10). These choices resulted in significantly less memory usage and faster computing times, with very similar plots and an identical conclusion section. Lastly, as described in Section 3.3.1, we ensured the clustering for creating homogeneous groups was performed on sparse embeddings.

What was easy. The original code was easy to find and publicly available. Together with a well-organized repository, it allowed us to reproduce their results. Furthermore, the original research already discussed several useful extensions in their appendix, answering quite some of our questions and therefore enabling efficient progress. The foundation of enabling multiple-item recommendations was already present in the original code, enabling easier integration of this extension.

What was difficult. Loading the original dataset from Kaggle resulted in more data than described by the original authors, which made it unclear how they ended up with the reported number of papers in the train and test set. Also, some adjustments were necessary to obtain a working pipeline. This required us to more thoroughly examine their code, which was challenging since only parts of the code contained comments. Moreover, to determine whether a recommendation was a good one, we needed data from the Semantic Scholar API. Lacking an API key, we were unable to make all the requests needed to get the data for every paper. Besides, the original paper did not provide code or a clear explanation to perform the logistic regression, hence we had to assume it was conducted on all the similarity scores between authors and items for 1,128 authors. Furthermore, extending their theoretical framework for multiple-item recommendation was challenging, because their framework did not explicitly incorporate that only one item is to be recommended. It is implicitly included in both their framework and code. Due to limited comments, fully integrating this was challenging.

Communication with original authors. During this replication study, we contacted the authors for the following two points:

- The linked GitHub repository does not provide a file for the logistic regression. As a result, we were unsure how it was performed: on all similarity scores between all authors and items or just on the highest similarity score.
- After downloading the data, filtering for the Computer Science category, and performing other preprocessing steps, we obtained a training set of 225,138 papers and a test set of 65,948 papers. However, the original paper reports working with a training set of 139,308 papers and a test set of 14,307 papers.

Unfortunately, due to time constraints, we did not receive a response in time before the submission deadline.

References

R. Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *FAT*, 2018. URL https://api.semanticscholar.org/CorpusID:46816230.

Sophie Greenwood, Sudalakshmee Chiniah, and Nikhil Garg. User-item fairness tradeoffs in recommendations. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://neurips.cc/virtual/2024/poster/94638.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1–49, 2012. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2012.02.006. URL https://www.sciencedirect.com/science/article/pii/S0370157312000828. Recommender Systems.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 2243–2251, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272027. URL https://doi.org/10.1145/3269206.3272027.

Alberto Moro and Laura Lonza. Electricity carbon intensity in european member states: Impacts on ghg emissions of electric vehicles. *Transportation Research Part D: Transport and Environment*, 64:5–14, 2018. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2017.07.012. URL https://www.sciencedirect.com/science/article/pii/S1361920916307933. The contribution of electric vehicles to environmental challenges in transport. WCTRS conference in summer.

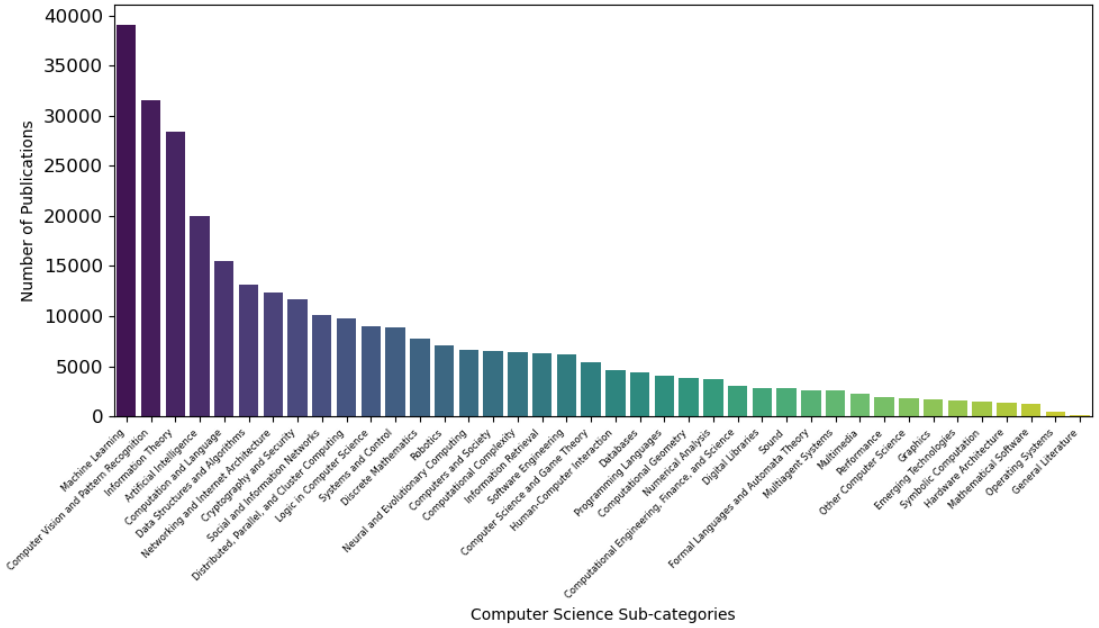
Lequn Wang and Thorsten Joachims. User fairness, item fairness, and diversity for rankings in two-sided markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pp. 23–41, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386111. doi: 10.1145/3471158.3472260. URL https://doi.org/10.1145/3471158.3472260.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:232110379.

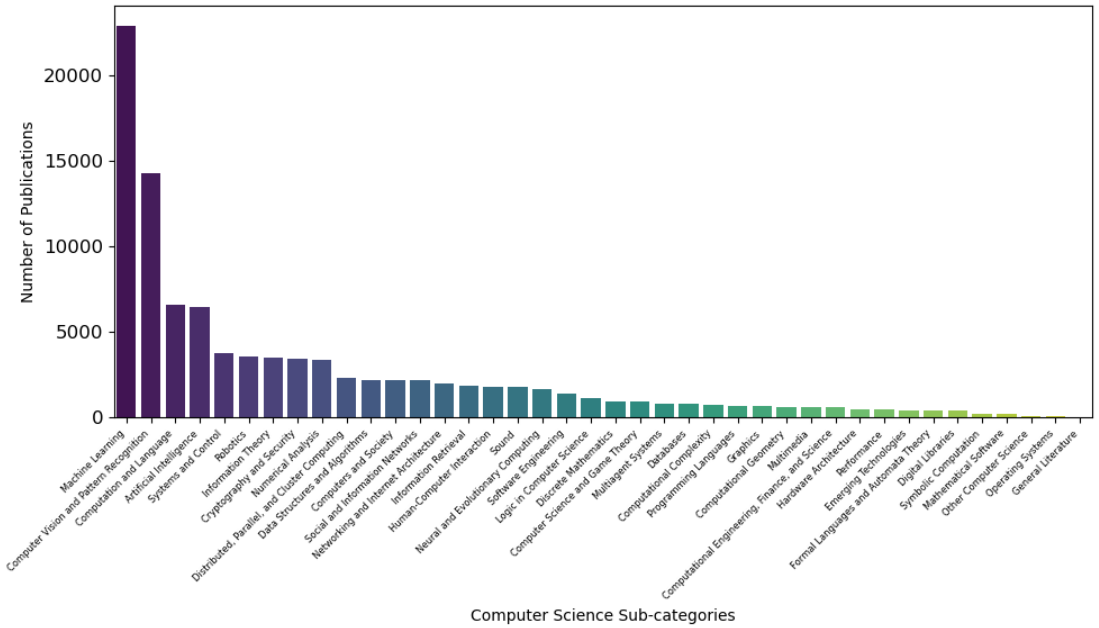
Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3), February 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL https://doi.org/10.1145/3547333.

A Appendix

A.1 arXiv dataset distribution of this study



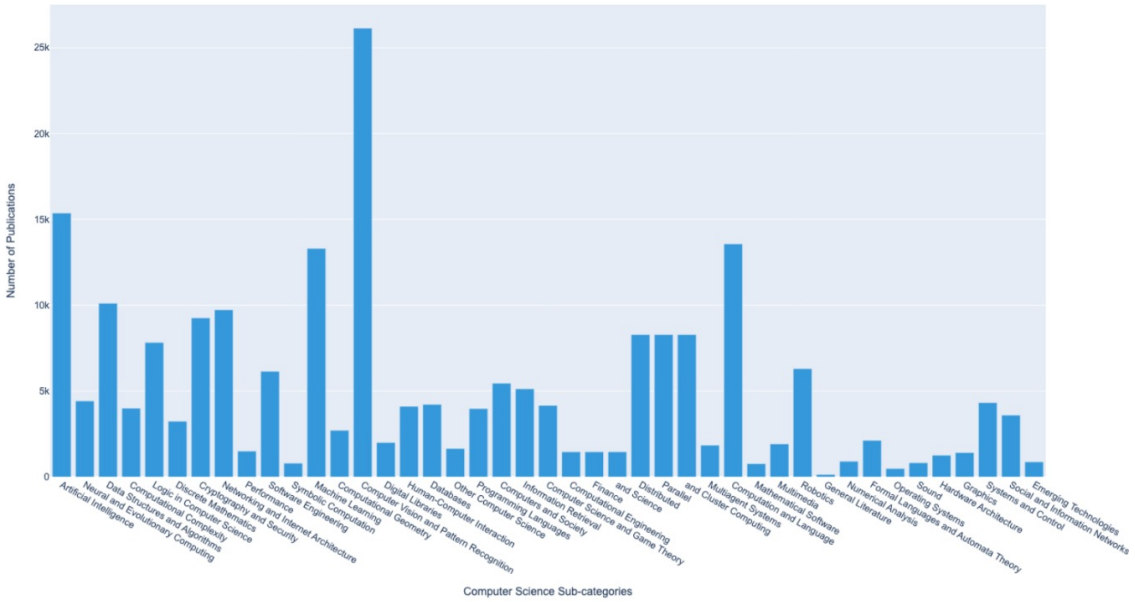
(a) Distribution of papers by subcategory in the train set (i.e. papers published before 2020).



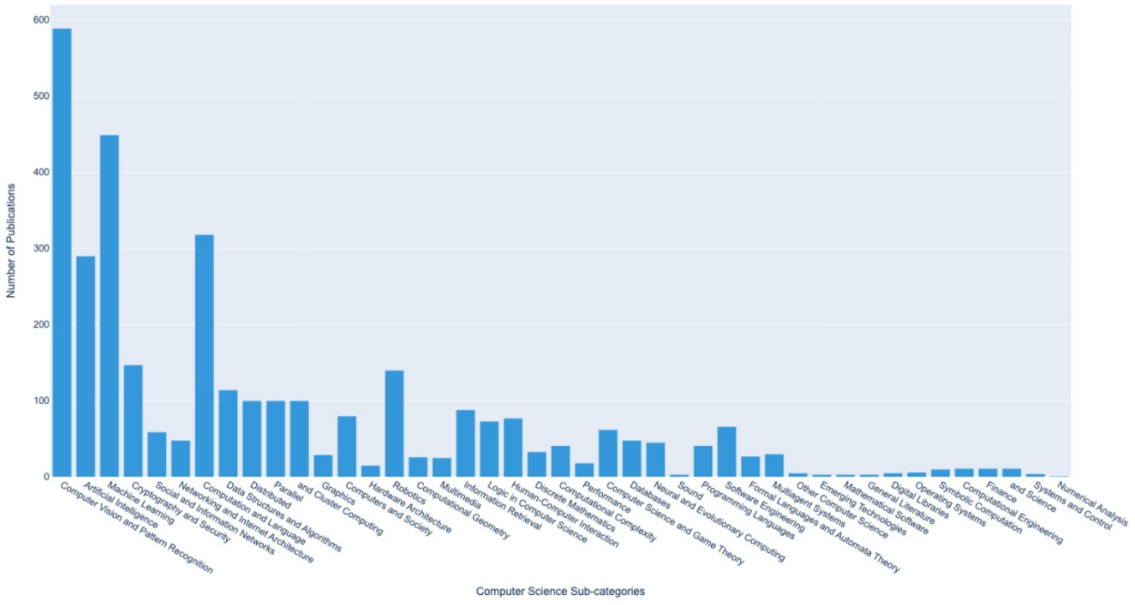
(b) Distribution of papers by subcategory in the test set (i.e. papers published in 2020).

Figure 6: Distribution of papers by subcategory per dataset. A paper possessing multiple subcategories is counted multiple times.

A.2 arXiv dataset distribution of original paper



(a) Distribution of papers by subcategory in the train set (i.e. papers published before 2020).



(b) Distribution of papers by subcategory in the test set (i.e. papers published in 2020).

Figure 7: Distribution of papers by subcategory per dataset (Greenwood et al., 2024).

```
429: Too Many Requests. Please wait and try again or apply for a key for a higher rate
limits.
```

A.3 API rate limit error

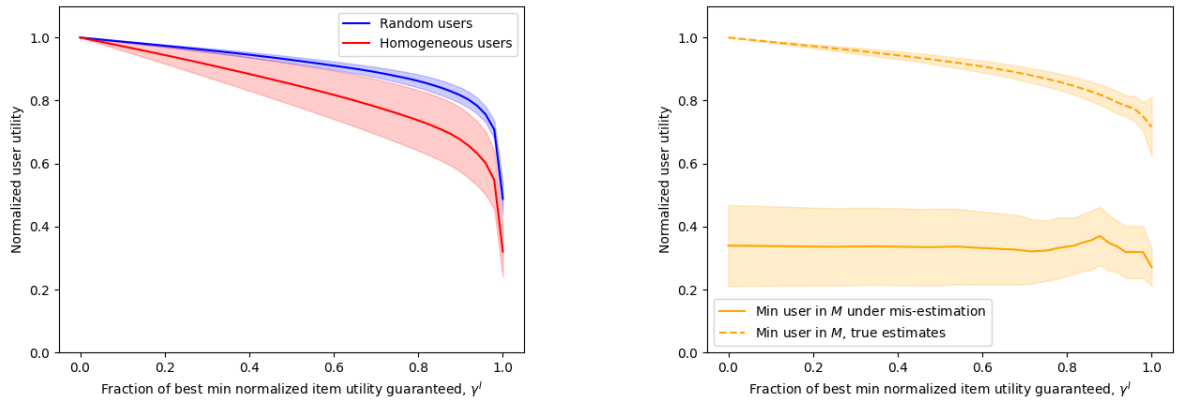
When making an API request when more than 1,000 other unauthenticated requests were performed, the following error was given:

After a large number of requests (sometimes in the hundreds, sometimes in the few thousands), occasionally the server kicked all requests for dozens of minutes and the Python script was terminated. This is likely caused by the server blocking our IP as a common rate-limiting strategy.

```
requests.exceptions.ConnectionError: HTTPSConnectionPool(host='api.semanticscholar.org'
, port=443): Max retries exceeded with url: *URL* (Caused by NameResolutionError
("<urllib3.connection.HTTPSConnection object at 0x7fd8db3a7f20>: Failed to resolve
 'api.semanticscholar.org' ([Errno -3] Temporary failure in name resolution)"))
```

A.4 Test set sampling method in provided code

Figure 8 shows plots produced by the sampling method in the originally provided code. This method differed from what was described in the paper by how authors were sampled that could be chosen to generate the curves. Here, first, 1,000 authors were randomly sampled from the test set, after which embeddings were made for authors that also appeared in the training set, resulting in a significantly smaller pool of authors (659 authors) from which was sampled to create the plots.

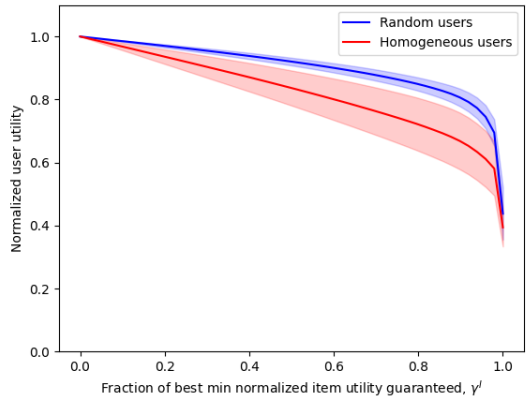


(a) Homogeneous versus diverse users on the original dataset.

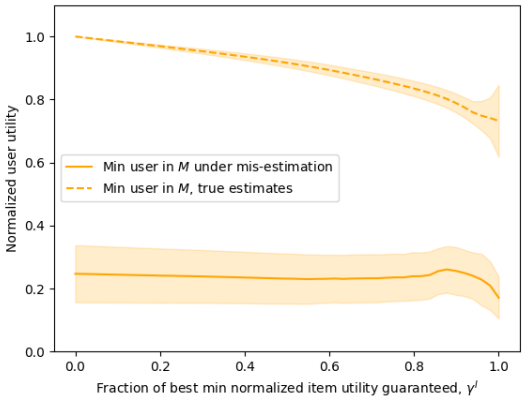
(b) With and without misestimation on original dataset.

Figure 8: Empirical findings between the normalized user and item utility by sampling from 659 authors from a subset of 2,188 papers. γ describes the item fairness constraint as explained in Section 3.1.

A.5 Test set sample of 14,307



(a) Homogeneous versus diverse users on the arXiv dataset.



(b) With and without misestimation on the arXiv dataset.

Figure 9: Empirical findings between the normalized user and item utility by sampling from 1,000 authors from a subset of 14,307 papers. γ describes the item fairness constraint as explained in Section 3.1.

A.6 Amazon books reviews dataset

This section provides results of our explorative data analysis of the Amazon books reviews dataset.

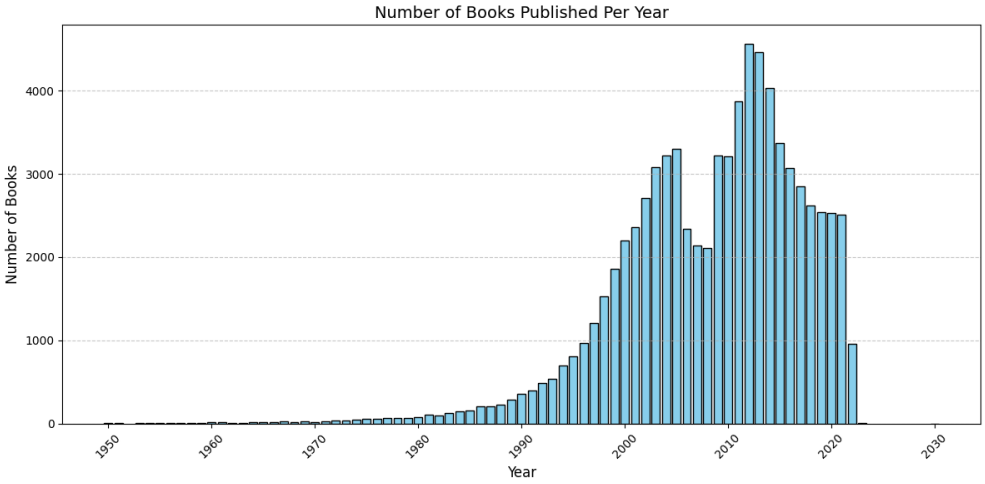


Figure 10: Number of books published per year in the Amazon books reviews dataset (48 books between 1802 and 1950 were dropped for visualization).

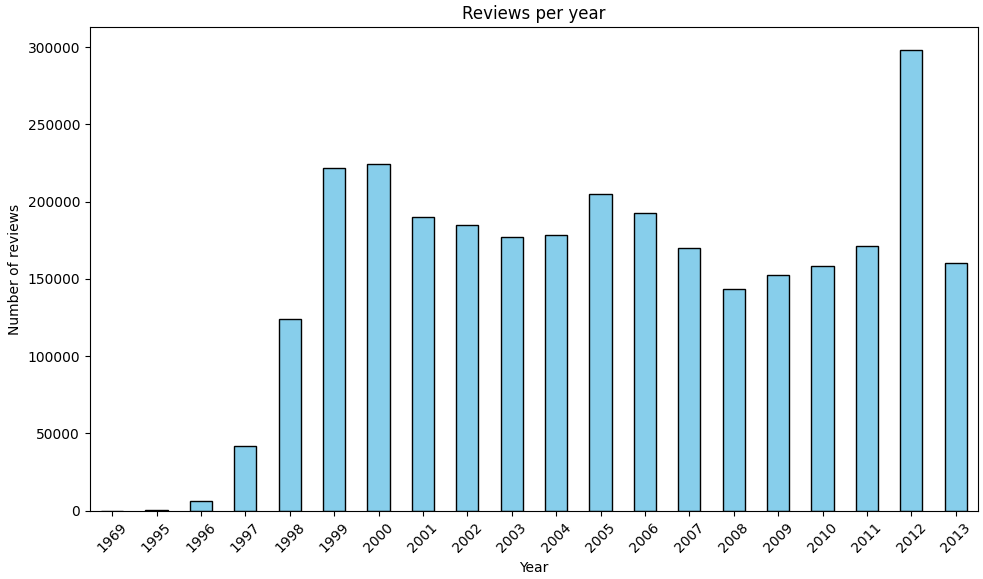


Figure 11: Number of reviews per year in the Amazon books reviews dataset.

	Silhouette score
arXiv preprints dataset	0.3401 ± 0.0038
Amazon books reviews dataset	0.3447 ± 0.0096

Table 4: Silhouette scores for three random seeds (42, 999, 123).