Scientometrics (2023) 128:2703–2736
https://doi.org/10.1007/s11192-023-04636-2

Biases in scholarly recommender systems: impact, prevalence, and mitigation

Michael Färber[1] · Melissa Coutinho[1] · Shuzhou Yuan[1]

Received: 17 December 2021 / Accepted: 17 January 2023 / Published online: 21 March 2023

Abstract

With the remarkable increase in the number of scientific entities such as publications, researchers, and scientific topics, and the associated information overload in science, academic recommender systems have become increasingly important for millions of researchers and science enthusiasts. However, it is often overlooked that these systems are subject to various biases. In this article, we first break down the biases of academic recommender systems and characterize them according to their impact and prevalence. In doing so, we distinguish between biases originally caused by humans and biases induced by the recommender system. Second, we provide an overview of methods that have been used to mitigate these biases in the scholarly domain. Based on this, third, we present a framework that can be used by researchers and developers to mitigate biases in scholarly recommender systems and to evaluate recommender systems fairly. Finally, we discuss open challenges and possible research directions related to scholarly biases.

Keywords Recommender systems · Fairness · Scholarly data

Introduction

Bias refers to the phenomenon of unfairly favoring a group of people or an opinion (O'Neil, 2016; Delgado-Rodriguez & Llorca, 2004; Smetanin & Komarov, 2022), and is highly relevant for recommender systems. For instance, in the case of movie recommendation, users tend to rate popular movies more often, which results in unpopular movies being recommended less frequently (Park & Tuzhilin, 2008; Abdollahpouri et al., 2017). Liu et al. (2016) found that conformity, which implicitly shapes people's behaviors to group norms, strongly influences users' rating behavior in recommender systems. Zehlike et al. (2017) found that people search engines, quite commonly used in job recruiting and to find friends

✉ Michael Färber
michael.faerber@kit.edu

Melissa Coutinho
melcouts97@gmail.com

Shuzhou Yuan
shuzhou.yuan@kit.edu

[1] Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

on social media, are biased based on gender, race, and physical disabilities. Identifying and mitigating bias in AI systems, particularly recommender systems, has therefore been garnering interest from researchers worldwide. For instance, some have proposed new learning-to-rank algorithms to mitigate bias (Zehlike et al., 2017; Abdollahpouri et al., 2019), while others have proposed techniques to better understand and model user preferences for recommendation (Zheng et al., 2020; Liang et al., 2016).

In the scholarly domain, researchers are faced with excessive information overload (e.g., tens of thousands of papers published annually in specific fields; Färber, 2019). Consequently, scholarly recommender systems—which recommend papers (Bogers & Van Den Bosch, 2008), research trends (Zhao et al., 2018), collaborators (Salman et al., 2020), journals, conferences (Klamma et al., 2009), and other entities—are increasingly gaining attention. However, to the best of our knowledge, scholarly bias has not been considered systematically so far. This is surprising, as academic society has an ethical responsibility (Polonioli, 2020). Scholarly bias can affect millions of researchers in various fields. It indirectly influences research output and funding decisions regarding research projects with cumulative multibillion dollar budgets. Bias affects which research ideas get promoted and are finally used in industries, impacting the long-term welfare of industrial nations. With the increase in the number of papers on bias in recommender systems in general, it is important that scholars have a global picture of the various types of bias in recommender systems and bias mitigation methods. In addition, current works on bias are fragmented and use important terms in different ways. For instance, some papers refer to "selection bias" as bias that occurs due to undersampling training data (Wang et al., 2016; Zhang et al., 2019), while others use that term to refer to the bias caused by limited user exposure to recommendation results (Ovaisi et al., 2020), which means only a small number of items are displayed to a user. Also, some papers do not mention the term "bias" in their content, but address at least one type of bias (Liang et al., 2016; Sugiyama & Kan, 2013; Gai & Lei, 2014). Due to these aspects, it might be difficult for researchers interested in this topic to find significant related work. In addition to this, it is difficult to identify and measure bias in scholarly recommender systems despite the pressing need to do so, because these systems face issues related to domain specificity, data sparsity, and data heterogeneity (see "System-caused bias" section). Thus, we believe it is necessary to provide an overview of this area, so as to help researchers understand the current state and future work on this topic.

Biases in *scholarly* recommender systems have been disclosed in several works: For instance, Liang et al. (2016) discussed the recommender systems' "exposure problem", which can result in frequent recommendation of popular scientific articles. Salman et al. (2020) observed gender and racial biases as well as biases based on location in academic expert recommendations used in the hiring process, to find reviewers, or to assemble a conference program committee. In addition, researchers speak of "filter bubbles" ("information bubbles") as the phenomenon that people are isolated from a diversity of viewpoints or content due to online personalization (Nguyen et al., 2014). Polonioli (2020) claimed that recommender systems might isolate users in information bubbles by insulating them from exposure to different academic viewpoints, creating a self-reinforcing bias damaging to scientific progress. Finally, Gupta et al. (2021) found that scholarly recommender systems are biased as they underexpose users to equally relevant items. Figure 1[1] shows the

[1] These statistics were calculated based on the set of papers collected for our literature survey. See "Recommendation bias" section for more information.

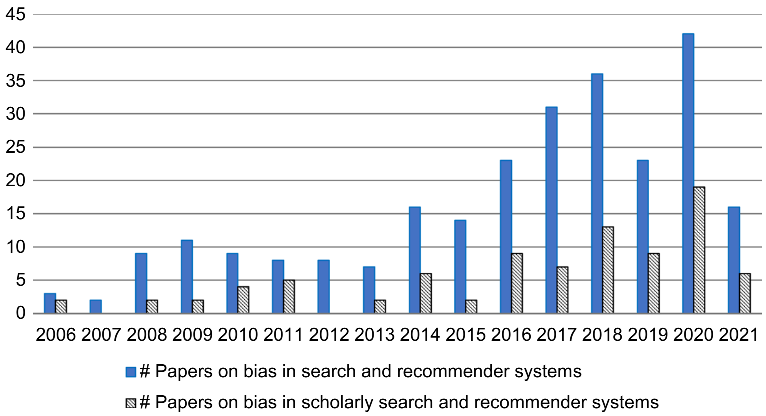


Fig. 1 Number of papers in our corpus by publication year

number of papers on bias in scholarly recommender systems and in recommender systems in general. We observe that bias both in recommender systems in general and in scholarly recommender systems has received a lot of attention. A paper that tackles the popularity bias of recommending scientific articles (Wang & Blei, 2011) won the Test of Time award at the KDD 2021 conference.[2] In addition, the best paper award at SIGIR in 2020[3] was awarded to Morik et al. (2020) for their paper on fairness in learning-to-rank systems. All this indicates that bias in scholarly recommender systems is a timely and important topic.

Few literature surveys in the context of bias in recommender systems *in general* (i.e., independent of the domain) have been published (Chen et al., 2020a; Mehrabi et al., 2021; Piramuthu et al., 2012). For instance, Chen et al. (2020a)'s survey highlighted the different biases a recommender system faces (excluding unfairness), as well as various methods of mitigating them. Our survey differs from these works in the following aspects:

1. We focus on detecting and mitigating biases found in scholarly search and recommender systems. The scholarly domain has not been considered systematically with respect to biases, while it exhibits several peculiarities that make identifying and mitigating biases challenging and not straight forward. To name only a few peculiarities, which cannot be found in other domains, such as news publishing:

   (a) Papers are domain-specific texts, using a specific vocabulary and often following a dedicated structure, such as IMRaD (Wu, 2011). Furthermore, papers are enriched by citations. Both the prose and the citations are prone to bias.
   (b) Papers typically have several authors, often with different background (e.g., gender, institution, country), which can be a ground for bias.
   (c) Research takes place in scientific communities, which are often quite separated from each other, leading to specific social norms.

[2] https://kdd.org/awards/view/2021-sigkdd-test-of-time-award-winners.

[3] https://sigir.org/sigir2020/awards/.

(d) Science thrives on scientific discourse between researchers. In addition, it is increasingly the case that papers are distributed and promoted on social media, which can lead to further effects and bias.

(e) Scholarly systems are known to suffer from cold-start problems (Wang & Blei, 2011). The number of items is much greater than the number of users, leaving many items without any feedback. Vellino (2015) found that the sparsity of the user-item matrix in Mendeley was three times smaller than that of Netflix.

(f) The Matthew effect ("the rich gets richer") was found to be particularly prevalent in academia (Merton, 1968; Zeng & Zuo, 2019; Zhou et al., 2022; Wang & Barabási, 2021).

2. We not only specify a definition of each bias, but also its impact and prevalence in the academic field, showing which biases are particularly relevant and worth to be investigated further.
3. We propose a framework on bias detection and mitigation in *scholarly* recommender systems.

Overall, in this article, we make the following contributions:

1. We define the types of biases in scholarly recommender systems, and describe their characteristics and influence in academia.
2. We summarize existing methods for debiasing scholarly recommender systems, and provide a framework for applying these methods.
3. We identify open challenges and discuss future directions to inspire further research on scholarly bias.

The structure of this article is as follows: In the "Recommendation bias" section, we identify the various biases present in scholarly recommender systems and classify them according to their impact and prevalence. The "Popularity bias" section explains methods of bias mitigation. It also proposes a debiasing framework that can be followed to detect and mitigate biases in scholarly recommender systems. Finally, the "Unfairness" section discusses future research directions and the challenges faced when building or using such systems.

Recommendation bias

In this section, we first introduce definitions of important concepts used throughout the article. We then describe how we obtained our collection of papers concerning scholarly biases, categorize types of scholarly bias according to the lifecycle of (scholarly) recommender systems, and describe the impact and prevalence of the individual biases.

Terminology

In the following, we define important concepts used throughout our article.

- **Bias** refers to the phenomenon of unfairly favoring a group of people or an opinion (O'Neil, 2016). **Scholarly bias** refers to *bias* occurring in the scholarly domain (e.g., when dealing with scientific publications, researchers, affiliations, and venues), such as in scholarly recommender systems.

- **Conformity** is an important factor leading to bias. It is defined as the act of matching attitudes, beliefs, and behaviors to group norms, politics or being like-minded (Cialdini & Goldstein, 2004).
- **Filter bubble** or **information bubble** is a persistent phenomenon existing in search engine and social media (Bruns, 2019). With the personalized recommendation, users are exposed to information targeting their interests and reinforcing their belief instead of seeing balanced information. Consequently, users are isolated in the bubbles, where only the information consistent with their viewpoints can be seen (Pariser, 2011).
- **Echo chamber** is caused by the lack of diverse perspectives and framed by like-minded users, especially on social media platforms (Cinelli et al., 2021). In an *echo chamber*, a group of people having the similar perspective choose to preferentially connect with the people inside the group, and exclude the viewpoints of outsiders (Bruns, 2017).

Paper collection

To collect relevant literature for this survey, we first performed a keyword search following Beel et al. (2016)'s procedure using Google Scholar as a data source.[4] We relied on Google Scholar, as it is not only a commonly used academic search engine (Haddaway et al., 2015), but also covers more citations than other sources (Martín-Martín et al., 2021). We used "bias," "fairness," "recommender" or "recommendation," combined with "paper" or "citation," as keyword queries. Following the snowball sampling technique, we broadened the search by looking through the bibliography of these papers, and downloaded relevant papers that had been cited by the first set of works. To narrow down our results, we searched using keywords such as "academic" or "scholarly" in papers' titles and body text. Overall, 88 documents addressed some kind of bias in scholarly searches or recommender systems, while 171 documents addressed bias in general. The list of all considered publications is available online.[5]

Categories of bias in scholarly recommender systems

According to Chen et al. (2020a), we can differentiate between the following steps of recommender systems:

1. *Training of recommender systems model (Data → Model)*: recommender systems are trained based on observed user-item interactions, user profiles, item attributes, etc.
2. *Recommending items to users (Model → User)*: recommender systems infer user preferences toward items and provide recommendations to the users.
3. *Collecting data from users (User → Data)*: New user actions are integrated into the training data.

We argue that biases can be categorized more intuitively when considering their immediate cause – which is either the human user or the recommender system itself. In the following, we outline the biases of these bias categories.

[4] Two separate searches were conducted in April and October 2021, respectively.

[5] See https://doi.org/10.6084/m9.figshare.17069840.

Human-caused bias

The human-caused bias category includes biases that occur outside of the recommender system's training and deployment steps and covers biases regarding the user behavior. User behavior can be collected explicitly via ratings (Tang & McCalla, 2009) and implicitly through user-item interactions, such as clicks, downloads, and citations. While explicit data tells us explicitly whether a user likes or dislikes an item, it is often difficult and labor-intensive to obtain it (Yang et al., 2009; Naak et al., 2009). Most recommender systems nowadays therefore use implicit feedback data (Pennock et al., 2000; McNee et al., 2002). If user decisions and actions are biased, this results in biased data as well. As recommender system models are typically trained on recorded user behavior (e.g., user interactions with an existing recommender system), the biases in data collected from users leads to biases in model training which could lead to the recommender systems model making biased predictions (i.e., recommendations) to users.

Conformation bias, *position bias*, and *selection bias* are bias types which are mentioned by Chen et al. (2020a) as biases that occur in recommender systems in general and which we consider as being immediately human caused. In the following, we outline what these biases mean in the scholarly domain, as well as their impact and prevalence.

1. Conformation bias

Definition: *Users tend to conform their beliefs with those of their peers, friends, and those in the same communities, even if doing so goes against their own judgment, meaning the rating values do not always signify true user preference* (Chen et al., 2020a).
Impact: In case of the conformation bias, scholars ignore personal beliefs to follow those of their peers. Conformation bias can have a significant impact on the academic society, resulting in ethical and social concerns. It can be understood as non-optimal scientific practice, as researchers should steadily rely on themselves and should take over responsibility for their actions. In particular, social media found its way into science, leading to the effect that scholars may prefer items favored by their social ties or communities.
Prevalence: Researchers are exposed to a variety of social influences, including the urge to fit in with their peers. By evaluating a network epistemology model in which scientists tend to adopt activities that conform to those of their neighbors, Weatherall and O'Connor (2020) investigated the interplay between conformism and the scientific community's epistemic goals. They came to the conclusion that conformism reduces the likelihood of actors taking successful activities (Weatherall & O'Connor, 2020). Studies have also found that the decision of a group can be biased toward the opinion of the group member who started the discussion. Researchers use a variety of social media platforms (Van Dijck et al., 2018), including ResearchGate, LinkedIn,[6] Facebook,[7] Twitter,[8] and Academia,[9] to communicate and disseminate information, such as sharing their own articles with possible readers and looking for collaborators. Scholars who have no paper access due to license restrictions, especially the ones in underdeveloped areas, may obtain paper resources in this way. According to a 2015 survey, 47% of scientists affiliated with the American Association for

[6] https://www.linkedin.com/feed/.

[7] https://www.facebook.com/.

[8] https://twitter.com/.

[9] https://www.academia.edu/.

the Advancement of Research (AAAS) use social media to keep up with new findings and discuss science by sharing their thoughts (Center, 2015). The ability to create social media networks has aided in the communication and collaboration of scientists irrespective of geographical location (Editorial Board, 2018). Scholarly metrics like altmetrics[10] count the number of times a research paper has been downloaded or shared on social media networks like Twitter to determine its qualitative and quantitative impact (Adie & Roe, 2013). Some research also shows that the most cited scientific papers are also the ones with the highest impact according to altmetrics (Torres-Salinas et al., 2013). However, there might be remarkably exceptions in which altmetrics say nothing about the quality of artifacts. For instance, there are three retracted papers within the top 100 articles at altmetric.com as of November 2021. Altmetrics indicators that use data from social networks are ever present in the research landscape. Due to the growing influence of social networks in our daily lives and to simultaneously tackle the data sparsity problem, recent studies tend to integrate social media information, such as altmetrics, into the recommender systems, known as social recommender systems (Zhao et al., 2018; Asabere et al., 2014; Xu et al., 2012b; Lee & Brusilovsky, 2011). Wang et al. (2017) offer a social recommendation model that indicates if individuals have affinities with items favored by their social ties. The conventional explanation is that a user's taste is comparable to and/or affected by her trustworthy social network friends. Scholars may look for information for a quick conclusion, or cease seeking information once a conclusion is reached. In such cases, personalized social recommender systems that use social information to create recommendations could promote conformity in scholarly circles (Polonioli, 2020). Furthermore, Analytis et al. (2020) found that traditional collaborative filtering algorithms for recommender systems, such as the weighted k-nearest neighbors (k-nn) algorithm, produce social influence networks, and that the most influential individuals (e.g., highly cited researchers or famous scientists) benefit the most from the k-nn algorithm. They examined datasets from Faces (DeBruine, 2017) and Jester (Goldberg et al., 2001) and show that as the number of users in a network increases, so does the influence of only a handful of people. Zheng et al. (2020) propose a framework to disentangle user's interest and conformity because a user might click an item, not because she likes it, but because many others have clicked on it. This is true for any platform that uses a social metric, such as likes, downloads, and citations. Therefore, influenced by other's opinions, users' rating values might not signify true preference. Explicit, published indicators or analyses on the prevalence of conformation bias in scholarly datasets were not found by us.

2. Position bias

Definition: *Users tend to interact with items in higher position of the recommendation list regardless of the items' actual relevance* (Collins et al., 2018).

Impact: Position bias makes the evaluation of recommender systems challenging, because it also affects how the users interact with a series of recommendations (Mansoury, 2021). Tracking a user's clicks on a series of recommendations can often be used to infer the relevance of a set of recommendations. This click data can thus be used to assess the effectiveness of a recommender system. However, due to position bias, the likelihood of a user engaging with an item may not reflect its absolute relevance within the set. If position bias

[10] https://www.altmetric.com/.

is not corrected in scholarly search and recommender systems, some recommended items (papers, venues, citations, etc.) will receive less than their “fair” share of clicks, views, downloads, etc. and others will receive more than their “fair” share. Considering the classical *Matthew effect*(Merton, 1968), researchers also tend to use the papers that show up in the top *n* for the purpose of citing. Subsequently, those papers receive more citations and are weighted higher in further search results.

Prevalence: In search and recommender systems (e.g., Google Scholar search), position bias seems to be a widespread phenomenon: According to eye-tracking studies (Joachims et al., 2017a), users are less inclined to look at lower-ranking items in vertical lists, instead focusing on the first few entries. Furthermore, 65 % of users engage with lists in a depth-first manner, clicking on the first item that appears relevant without thoroughly evaluating the full list (Klöckner et al., 2004). When Joachims et al. (2017a) studied the user behavior on Google’s results page (see Fig. 2), they found that users tend to click substantially more often on the first rather than the second ranked item, while they view both items with almost equal frequency. This behavior is similar when using academic search engines like Google Scholar, PubMed, and Web of Science,[11] where the items (e.g., papers) ranked higher receive higher citations, views, clicks, downloads, etc. Wang et al. (2018) also observed a considerable position bias when users search for emails.

3. Selection bias

Definition: *Users sometimes choose to rate items not out of relevance or item quality, but simply because they are influenced by other factors like item popularity, higher citation count of author, paper, etc. which have little or nothing to do with relevance or quality. As a result, observed ratings are not a representative sample of all ratings* (Chen et al., 2020a).

Impact: Selection biases in the data impact search and recommender systems for scholars by making inaccurate predictions. In recommender systems the algorithms used to predict user preferences are designed to have high prediction accuracy on the assumption that the missing implicit ratings are missing at random (MAR), i.e., there is no bias operating over which items are rated and which are not. MAR signals rarely exist in the real world, because it is very unlikely that a recommender system would recommend items completely at random. Implicit data can be missing-not-at-random (MNAR) due to the exposure, position, and popularity bias of the recommender system’s model (Stinson, 2021). This leads to a *skewed observed rating distribution*, because there is a lot of data on certain (popular, relatively old) items, whereas there is very less data on other (state of the art, relatively new) items. This makes learning the preferences of users challenging for a recommender system’s model. Furthermore, recommender system algorithms that depend entirely on observed implicit data or naively based on such missing click data will produce biased recommendations (Hu et al., 2008; Steck, 2010; Marlin & Zemel, 2009). It is also completely up to the user which item she chooses to rate (in the form of clicks, views, downloads, etc.). She might choose to rate an item not because she finds it relevant, but because of item popularity, high citation count, etc. Since her selections are biased, the observed rating distributions are biased as well. For instance, most scholars would choose to cite a paper that has many citations, or one that was presented

[11] https://www.webofknowledge.com/.

at a world-renowned conference, over another paper that has fewer citations, even though the content of the latter was more relevant to them. As another example, papers written by a famous author are more likely to be viewed regardless of relevance or interest, simply because most scholars tend to access them. Wang and Barabási (2021) revealed that in the academic world, famous scientists get credit easier than less famous researchers. This could lead to items being labeled as false positives in the training data and the recommender system making incorrect predictions. Selection bias also occurs in the *sampling process*, when the chosen data sample is biased and not a representative of the whole population (Zadrozny, 2004).

Prevalence: Selection bias can be observed in IR systems of various kind, such as paper search engines (e.g., Google Scholar) and paper/citation/dataset recommender systems. Note that position bias (i.e., bias of considering only top-k results) can also lead to selection bias (i.e., bias due to considered background information) in such systems (Ovaisi et al., 2020). Users rarely view all relevant results, either because the system displays a shortened list of the top-k recommended items or because users do not take the time to review all of the ranked results. Lower ranked, relevant results have a marginal chance of being noticed (and clicked) and may never be boosted in Learning to Rank (LTR) systems, where a ranking model is learned based on training data. Ovaisi et al. (2020) therefore model selection bias in two semi-synthetic datasets from the Yahoo! Learning to Rank Challenge (C14B)[12] by assigning a zero observation probability to documents below a cutoff *k*. In the scholarly domain, Forsati et al. (2017) investigated a selection bias in the CiteULike[13] dataset, containing user ratings for papers. Selection bias is relatively comprehensively diagnosed in non-scholarly datasets: Mena-Maldonado et al. (2021) observed this bias in ratings of the MovieLens 1 M dataset (Maxwell & A, 2015), as the small number of popular items take up most of existing ratings, thereby displaying a MNAR pattern in the data distribution. Selection bias was also found in the Yahoo! base rating dataset (Marlin et al., 2012), where the users have a probability of rating either 1 (very good) or 5 (very bad) more than 50% of the time. Zhang et al. (2019) observed selection bias in a Natural Language Sentence Matching (NLSM) dataset like the QuoraQP dataset during the sampling process. Quora[14] is a social platform where knowledge can be shared by scholars and non-scholars alike. The authors analyzed how the sampling procedure of selecting some pairs of sentences by the providers can bring about an unintended pattern, i.e. selection bias, into the model. We were not able to discover explicit statistics on the prevalence of selection bias in scholarly datasets.

System-caused bias

System-caused bias, which we can also name *model-intrinsic bias*, occurs during the training and deployment of recommender systems. It is caused by unfair data or inducted by the model itself. If the data with bias is used for training a recommender systems model, the bias might also be in the recommender system's results. Due the effects of the feedback loop (i.e., recorded user behavior with the recommender system is used to retrain the recommendation model), the bias in the data might increase even more, leading to the "rich get richer" Matthew effect (Chaney et al., 2015). Moreover, unfair recommended results

[12] https://webscope.sandbox.yahoo.com/catalog.php?datatype=c.

[13] https://www.citeulike.org.

[14] https://www.quora.com/.

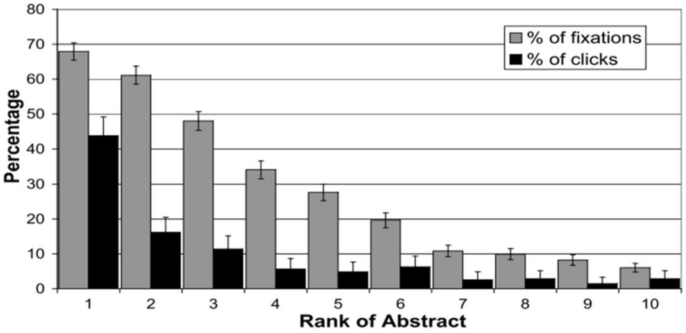


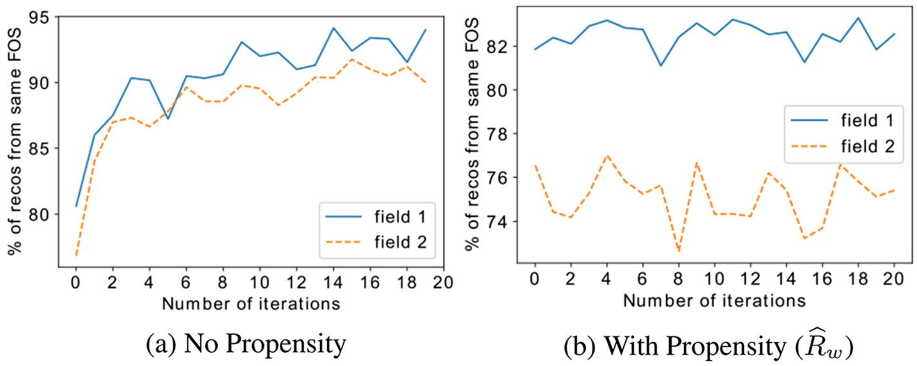
Fig. 2 Figure from Joachims et al. (2017a). Percentage of times an abstract was viewed ("fixations") and clicked depending on the result rank of the result. The abstracts ranked 1 and 2 receive most attention

caused due to biased and imbalanced data might also have a collective, disparate impact on certain groups of people (Courtland, 2018), such as female scholars, scholars from developing countries, and less experienced scholars. This *unfairness* leads to gender and racial biases among many others, which can cause detrimental impact to scholars looking for jobs and promotions, grant funding (with a total budget of more than 2% of the countries' GDP worldwide OECD, 2021), etc.

In the following, we outline the system-caused biases in the context of scholarly recommender systems. Following mainly Chen et al. (2020a), who described bias in recommender systems in general, these are *inductive bias*, *exposure bias*, *popularity bias*, and *unfairness*.

1. Inductive bias (Learning bias)

Definition: *The model makes assumptions to better learn the target function and to generalize the training data* (Chen et al., 2020a).
Impact: Strictly speaking, inductive bias is not part of the traditional definition of bias as unfairly favoring a group of people or an opinion (O'Neil, 2016; Delgado-Rodriguez & Llorca, 2004; Smetanin & Komarov, 2022). Inductive bias is considered to be positive, as it leads to more accurate results. It is added to the model to enable that the solution learned from the observed training data can better adjust to unseen data in the real world and not only to the specific test data. To achieve this, often a regularization term is introduced to avoid overfitting and to achieve better generalization (Mitchell, 1980; McClelland, 1992). Adding inductive bias is supposed to help the system to find desirable solutions without decreasing the performance (Battaglia et al., 2018). It can be especially useful for scholarly recommender systems, because they typically face the data sparsity problem: the user ratings are for very few items known, while the number of items compared to the number of users is very large. Indeed, scholarly recommender systems usually have a much higher data sparsity problem than movie recommender systems, for instance, because a large number of scientific articles and venues keeps getting added the database on a daily basis (Färber, 2019), and it is not possible for scholars to go through all the articles available.
Prevalence: In scholarly search and recommender systems, there is a scarcity of feedback data on user-item interaction. Vellino examined implicit ratings on Mendeley (which covers research papers) and Netflix (which covers movies) and discovered that Netflix had three orders of magnitude less sparsity than Mendeley (Vellino, 2013). Inductive biases are critical to the ability to classify instances that are not identical to the training instances (Mitchell,



(a) No Propensity

(b) With Propensity ($\widehat{R}_w$)

Fig. 3 Figures from Gupta et al. (2021) showing the proportions of recommended papers from the field of study (FOS) 1 and 2. For *no propensity* (*figure* **a**) the proportion of recommended papers from the same FOS increases over time for both FOS 1 and 2. This increases exposure bias and decreases the number of recommended papers from a different FOS over time. In contrast, when the models are trained with *propensity* (*figure* **b**) the proportion of recommended papers from a different FOS remains stable over time

1980), for example, data on newly added papers, female expert recommendations, etc. He et al. (2017) targeted to better generalize the training data using neural networks. To address the data sparsity problem in graph-based recommendation approaches which are mostly used for citation recommender systems, graph embeddings have been proposed in recent works (Tang et al., 2015; Perozzi et al., 2014). The latent feature of the nodes capture neighborhood similarity and community membership. Recommender systems that utilize social information for recommendation are also said to mitigate the data sparsity problem (Yang et al., 2017).

2. Exposure bias

Definition: *Users are exposed to certain relevant items. However, the unobserved items without interaction do not mean that they are irrelevant or represent negative preference* (Chen et al., 2020a).

Impact: Exposure bias is a major issue of implicit feedback datasets. Implicit data does not contain negative feedback: we may know why a user clicked on an item, but we do not know why she did not click on another item. It could be because she did not find the item relevant, but could also be because she was not exposed to the item, and hence did not "see" it. Exposure bias can have the following impact on scholarly recommender systems: (a) Open-acess (OA) documents and papers from a scholar's own field of study (FOS) are recommended more often (Gupta et al., 2021). (b) Exposure bias can exacerbate popularity bias, causing relevant but unpopular items to not be shown (Chen et al., 2020a). (c) Feedback loops (i.e., the recommendation model is trained on interactions of users with an existing recommender system, which might be already biased) also aggravate this bias and diminish the diversity of the recommended results. This could lead to a situation in which users only see a narrow subset of the entire range of recommendations, a phenomenon known as the *filter bubble* (Jiang et al., 2019).

Prevalence: For instance, let us consider a recommender system that recommends relevant citations to users to back-up claims (Färber & Jatowt, 2020), showing certain attributes of the paper, such as title, author information, and abstract. In this case, due to the exposure bias, equally relevant *papers from different or rare fields of study* might be less-cited historically (and therefore less recommended) because users have been preferentially exposed to papers in their own field of study, or less often been exposed to papers from a rare field of study. In an observed citation network, a number of relevant papers are not cited because the user was not exposed to those papers (Gupta et al., 2021). *OA documents* are available to a larger audience, because access is not limited by a "paywall" and can be read by practically anyone that has access to the internet (Harnad et al., 2008; Wang et al., 2015). This increases their probability of being found. Gupta et al. (2021) research exposure bias in citation recommendation. Upon considering two real-world datasets of citation networks, they use the probability that a node is exposed to another node as propensity score, and notice that a recommender system trained directly on the observed data underestimates the probability of being cited given exposure for low propensity nodes relative to high propensity nodes. This creates an exposure bias, which can be seen in the lack of diversity in the recommendation results, because papers from the same field of study are recommended almost 95% of the time, compared to just 82% of the time when propensity was considered (see Fig. 3). Researchers are mainly interested in the publications in their own disciplines, as Ojasoo and Doré (1999) showed. The impact of exposure bias on popularity in the scholarly field can have a similar impact as in other domains. For instance, in studying the MovieLens and Last.fm data sets, Abdollahpouri and Mansoury (2020) found that a user is not at all exposed to a vast majority (more than 80%) of the items. Exposure bias can affect item diversity. Because of the feedback loops associated with recommender systems, the number of highly exposed items that are recommended continuously increases over time, and those items that are less exposed despite being relevant get recommended less often with time (Mansoury et al., 2020). On social media, echo chambers describes the lack of diverse perspectives and the favor of the formation of groups of like-minded users, which leads to framing and reinforcing a shared narrative (Cinelli et al., 2021). Echo chambers can also emerge based on scholarly recommender systems: Jiang et al. (2019) show that feedback loops can generate echo-chambers and filter bubbles. They do this by observing the evolution of a user's interest, that is, when a user's interests change extremely over time. Moreover, Liu et al. (2020) found that if the data from the previous recommender systems are used in the new recommender without correcting inherent biases, then not only will these biases be carried forward in the new recommender systems model, but they might also be amplified due to the effect of the feedback loop (Chaney et al., 2015).

3. Popularity bias

Definition: *The tendency of the recommender system's algorithm to favor a few popular items while under-representing the majority of other items* (Abdollahpouri et al., 2021).

Impact: Algorithms that prefer popular items are often used to assist people in making decisions from a large number of options (e.g., top-ranked papers in search engine results, highly-cited scientific papers). Recent non-scholarly works (Abdollahpouri & Mansoury, 2020; Sun et al., 2019) show that one of the consequences of popularity bias in search and recommender systems is *disfavoring less popular items*, that is, recommendations are exposed to users purely depending on the degree of popularity. This can also create an exposure bias. In the absence of explicit signals or ratings, the systems rely on implicit indications such as popularity and engagement measures. They are simple to use and are frequently used as scalable quality proxies in predictive analytics algorithms. In networks, a bias occurs because all nodes are not equal and the nodes with more links receive more attention. This phenomenon is called "preferential attachment" (Barabási & Albert, 1999; Barabási & Bonabeau, 2003). Sun and Giles (2007) proposed a novel popularity-factor-weighted ranking algorithm that ranks academic papers based on the popularity of the publishing venues. The authors demonstrated that their system outperforms other ranking algorithms by at least 8.5 %. The wisdom of the crowd (Surowiecki, 2005) underpins the utility of such rankings: high-quality options tend to gain early popularity and, as a result, become more likely to be selected since they are more visible. Furthermore, being aware of what is popular can be viewed as a type of social influence; an individual's behavior may be influenced by colleagues or neighbors' choices (Muchnik et al., 2013). These principles indicate that high-quality material will "bubble up" in a system where users have access to popularity or engagement cues (such as ratings, amount of views, and likes), allowing for a cost-effective item ranking (Ciampaglia et al., 2018). Popularity metrics have not only been adopted to popular social media and e-commerce platforms with the general population as users, but they are also used on social media platforms like Papers With Code[15] for researchers to highlight popular and trending items. Many recommender systems disregard unpopular or recently released items with minimal ratings, focusing instead on those with enough ratings to be useful in recommendation algorithms (Park & Tuzhilin, 2008). It is important to tackle popularity bias due to the following reasons:

(a) Popularity bias reduces the extent of *personalization* and harms *serendipity* (Abdollahpouri et al., 2017). In addition, the *diversity* of the recommendation list could be decreased (Abdollahpouri et al., 2019). Different users have distinct preferences. Only recommending popular papers, authors, citations, etc. might not always be beneficial especially for those users that study or wish to research a rare or unexplored subject/field of study. There would also be no surprise element for the user, as she keeps seeing items that she has already seen before. Thus, the recommendation is not based on the actual content, but based on effective rules of thumbs.

(b) *Unfairness* of recommendation results also increases (Chen et al., 2020a). Even if they are good matches, regularly recommending popular items diminishes the exposure of other items, which is unequal with respect to the less popular items.

(c) Popularity bias will boost the exposure opportunities of other popular items (Abdollahpouri & Mansoury, 2020), making them even more popular. As a result, the training data also becomes unbalanced, elevating the so-called *Matthew effect*.

Prevalence: Quite recently in a paper by Zhu et al. (2021), popularity in terms of equal opportunity instead of the conventional popularity bias was considered. According to this paper, given that a user likes both a popular item i as well as a less popular item j, i is observed to rank higher than j even though the user likes both items. Due to position bias exhibited by the user, item i is more likely to be observed and clicked than item j, resulting in bias favoring popular items.

Yang et al. (2018) investigated two types of bias in terms of popularity in the CiteULike dataset, which contains user ratings for scientific publications: (a) **interaction bias** (i.e., the tendency for users to connect with popular items more frequently), and (b) **presentation bias** (i.e., the tendency for recommenders to unfairly present more popular items than long tail ones). Though this differentiation was used to highlight the prevalence of

[15] https://paperswithcode.com/.

popularity bias in scientific articles in the CiteULike dataset, it can also be made for all kinds of scholarly items, such as venues, citations, and experts. *Interaction bias* can be seen clearly in Fig. 4. The n_i^* distribution is highly skewed because the horizontal axis is log scaled: Less than 50 user interactions occur for 99 percent of the goods. However, this kind of bias can be rather well understood from a user perspective as it largely depends on user's behavior and her decision on which items she chooses to interact with. This can also be seen as a result of conformity as users tend to interact with popular items, because these items are most favored by majority of other users, and therefore they choose to *follow the crowd*. For this very reason, we decided to include interaction bias as a type of conformity bias, which we have explained in detail previously (see the "Introduction" section).

Yang et al. (2018) calculated the average number of times that an item with the observed popularity $n^* \in [1, max(n_i^*)]$ was recommended, denoted by f(n^*), in order to account for the presentation bias. A biased recommender may provide a $f(n^*)$ that is linearly or exponentially rising, whereas an unbiased system should anticipate one that is basically flat with a small slope (see Fig. 5). In the scholarly field, popular items are not only limited to highly-cited research papers, but can also include famous authors, conference venues, academic events, etc.

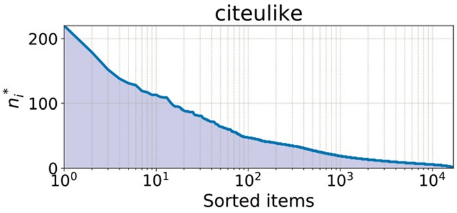


Fig. 4 Figure from (Yang et al., 2018). The distribution of n_i^* (the observed number of interactions with item i) in the citeulike dataset. The items are presented in descending order of n_i^*. The horizontal axis is log scaled for better visualization. The n_i^* distribution is skewed and the user interactions are significantly biased

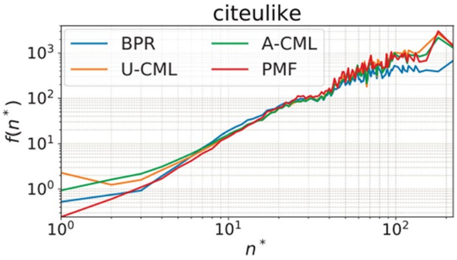


Fig. 5 Figure from (Yang et al., 2018). Empirically estimated f(n^*) on the CiteULike dataset and four recommendation algorithms. $f(n^*)$ denotes the average number of times that an item with observed popularity n^* was recommended. Both axes are log scaled. Therefore, exponential growth is linear in the figure. This shows significant presentation bias

4. Unfairness

Definition: *The system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others* (Chen et al., 2020a).

Impact: Unfairness is a serious and ongoing issue in academia where certain groups are discriminated based on gender (Holman et al., 2018), race (Fang et al., 2000), socio-economic status (Petersen et al., 2014), and university prestige (Way et al., 2019) among many others. Such inequalities have been found in grant funding (Lee & Ellemers, 2015), credit for collaborative work (Sarsons, 2017), hiring and promotions (Nielsen, 2016), peer-review (Shah, 2021), authorship (West et al., 2013) (Huang et al., 2020), and citations (Caplar et al., 2017). The difficulty of fairness has prevented recommender systems from becoming more pervasive in our culture. Particularly, distinct user groups are typically underrepresented in data depending on characteristics like gender, affiliation, or degree of education. The recommender system models are quite likely to learn these overrepresented or underrepresented groups when training on such unbalanced data, reinforce them in the ranked results, and possibly lead to systematic discrimination and decreased visibility for disadvantaged groups. Moreover, due to the effects of using historical data in supervised learning settings (see also the feedback loop for recommender systems, i.e., using the information of user behavior with recommender systems to create recommender system models), these results have the potential to (unknowingly) influence user thinking and behavior, and also make users part of the *filter bubble*, again re-enforcing these beliefs. These biases amplify over time due to the *Matthew effect*.

Prevalence: Andersen and Nielsen (2018) observed an indirect **gender bias** in the Web of Science's selection criteria for its citation resources. Mohammad (2020) observed gender gaps in Natural Language Processing (NLP) both in authorship as well as citations. Among all the NLP papers published from 1965 to 2000 in the ACL Anthology,[16] they found that only about 30% were written by women. They also used citation counts extracted from Google Scholar to show that, on average, male-first authors are cited markedly more than female-first authors, even when the experience and area of work are considered. Recommendations in educational and career choices are another important motivating application for fair recommender systems. Students' academic choices can have significant impacts on their future careers and lives. Polyzou (2020) found a course recommender system to give biased results to a student because of pre-existing biases and socio-technical issues present in input data. For instance, a system might rarely advise female computer science students to take intensive coding programs. If patterns from the last 50 years continue, parity between the number of male and female authors will not be attained in this century, according to Wang et al.'s detailed examination of gender trends in the computer science literature. They came to the conclusion that there is a persistent gender gap in the literature on computer science that might not be closed without deliberate action (Wang et al., 2021a). Google Scholar is of the most popular academic search engines which uses citation counts as the highest weighted factor for ranking publications. Therefore, highly cited articles are found significantly more often in higher positions than articles that have been cited less often (Beel & Gipp, 2009). According to Parker et al., a small group of scholars receive a disproportionately significant amount of citations in the scientific community. They looked at the social characteristics of these highly cited scientists and discovered that

[16] https://aclanthology.org/.

the majority of them are male, nearly entirely based in North America and Western Europe (Parker et al., 2010). Most search and recommender systems for scholars therefore show unfairness against less-cited authors, gender, and race, among many others in the recommended results.

Bias in **expert recommendation** within academia has been researched relatively fairly by many scholars. Expert recommender systems have been used for the following (see Salman et al., 2020): (1) to hire and recruit professors and researchers in academic positions, (2) to recommend experts to evaluate patents, (3) to identify reviewers for scientific conferences, and (4) to assemble a conference program committee. One study by the Nature Magazine (Lerback & Hanson, 2017) shows that only 20% of the peer-reviewers are women and are thus largely under-represented. Another study (Yin et al., 2011) shows that women and authors from developing countries were under-represented as editors and as peer-reviewers. The National Science Foundation (NSF) developed an automated reviewer selection system that considers different demographic features while selecting peer reviewers (Hettich & Pazzani, 2006) to tackle this problem. Bias in a peer review process can also be seen in the geo-location of the reviewer. For example, a study in (Publons, 2018) shows that the US dominated the peer review process by 32.9% while its publications represent only 25.4% of all publications.

Unfairness in **language** of scientific publication can also be observed. Journals not published in English tend to have a lower impact factor (Vanclay, 2009). Matías-Guiu and García-Ramos reason that this may not be due to the language itself, but to the fact that they are not included in authorship networks, since most English speaking authors do not read them, and hence, do not cite them either (Matías-Guiu & García-Ramos, 2011). May (1997) also found serious English language bias in the Institute for Scientific Information (ISI) database, as most non-English journals are not covered by the Science Citation Index (SCI). Cross-lingual citations make up less than two percent of all citations (Saier et al., 2021). In the field of recommender systems, Torres et al. (2004) observed that a recommender in a user's native language was greatly preferred to one in an alternative language, even if the items themselves recommended were in the alternative language (e.g., a Portuguese-based recommender recommending scientific papers in English).

Another important instance of unfairness in the scholarly field is the phenomenon of **citation bias**. Citations are key elements in the evolution of scientific knowledge and of the scientific discourse. They enable particular research findings to survive over time and to develop into academic consensus. Given the large body of scientific literature, it is often unfeasible to cite all published articles on a specific topic, particularly in case of page limitations given by journals and conference proceedings, although the San Francisco Declaration on Research Assessment (DORA; http://www.ascb.org/dora/) encourages publishers to not insist in such constraints). As a result, some selection of citations needs to take place. If this selection is influenced by the actual results of the research presented in the article, then citation bias occurs (Song et al., 2010). Duyx et al. (2017) conducted a systematic review and meta-analysis on the citation behavior of about 52 studies from the Web of Science Core Collection and Medline.[17] They found that positive articles, that is, articles that support an author's belief/claim, are cited about twice as often as negative ones, i.e., articles

[17] https://www.medline.com/.

that are critical to an author's belief/claim. Greenberg (2009) conducted a similar analysis and found that positive articles received 94% of the total citations.

Summary Overall, we can highlight the following aspects concerning biases in scholarly recommender systems:

1. All biases concerning recommender systems in general are also relevant to scholarly recommender systems.
2. In particular, *unfairness* in the form of gender bias has been addressed due to unbalanced gender distribution in academia (Sapiezynski et al., 2017; Islam et al., 2021).
3. While scholarly recommender systems aim to provide diversity in terms of content, they might also be responsible for "echo chambers" in science (West & Bergstrom, 2021) due to the *exposure bias* in underlying data. This is frequently observed in citation networks (Gupta et al., 2021).
4. The impacts of the *Matthew effect* cannot be ignored: most search and recommender systems display their results in an order influenced by citation counts and related criteria, and frequently cited papers are cited disproportionately more often as they increase in popularity.

Bias mitigation in scholarly recommender systems

Scholarly bias mitigation methods

According to D'Alessandro et al. (2017), bias mitigation methods can be grouped into the following three categories. They align with the steps of recommender systems' lifecycles, mitigating bias in the data collection, training, and recommendation steps (see the "Categories of bias in scholarly recommender systems" section):

1. **Mitigation during preprocessing**: This type of mitigation approach is deployed before the creation of a recommender system's model. Using these mitigation techniques, the dataset is transformed with the aim of removing the underlying bias from the data before modeling. However, these methods are usually hard to implement and may be ineffective, as they cannot remove the bias present in the machine learning model itself. These techniques are mostly used to decrease the level of discrimination and unfairness in the data (Calmon et al., 2017).
2. **Mitigation during processing**: This mitigation approach employs techniques that can be considered modifications of the traditional learning algorithms to address bias during the model training phase (e.g., utilizing bias-indicating metrics, such as propensity scores).
3. **Mitigation during postprocessing**: This mitigation approach uses techniques in which bias processing is performed after model training. These mitigation methods are noteworthy because the user's perception of and interaction with the results are essential.

Figure 6 and Tables 1 and 2 provide an overview of the mitigation methods mentioned in our paper collection. Overall, we see that concrete bias mitigation methods have been published with respect to all biases types outlined in the "Recommendation bias" section.

As dealing with bias during model training requires particular methods (e.g., neural network architectures), a large portion of publications is dedicated to system-caused bias.

Debiasing framework

Notably, most bias mitigation approaches in recommender systems only focus on one bias dimension. Only one paper (Polonioli, 2020) addresses three types of bias while focusing on the ethics of recommender systems. However, in reality, not just one but many types of bias affect recommender systems' performance simultaneously. Thus, in this paper, to mitigate the mixture of biases found during the steps of recommender systems' lifecycles, we provide a universal debiasing framework for scholarly recommendation.

Our debiasing framework is designed for the *scholarly domain*. As outlined in the "Introduction" section, in contrast to non-scholarly recommender systems (e.g., movie recommender systems), scholarly recommender systems have various unique characteristics. For instance, the content (full-text) of papers and the topology of networks (e.g., citation networks) are subject to scholarly biases. In addition, researchers, institutions, and papers are embedded into scientific communities and societal behavior. Technically, scholarly recommender systems typically need to deal with noise and the lack of negative feedback. Overall, our debiasing framework focuses on identifying and finding solutions pertaining to the mentioned issues that are frequently observed in scholarly AI systems.

Our debiasing framework consists of three steps:

1. **Detect biases:** To debias a recommender system, it is first important to detect the kinds of biases that occur (and that can be measured). This can vary depending on the use case. To determine what biases a system is dealing with, it is important to know the number of false negatives and false positives from the feedback data. These numbers can be determined via a simple confusion matrix or standard evaluation metrics, such as precision, recall, etc. User click-data can also be helpful. We can list the following characteristics of each bias type as rules of thumb to determine the type of bias in the data:

   (i) Exposure bias: if the number of false negatives is too high;
   (ii) Position bias: if the click data are imbalanced;
   (iii) Selection bias: if the number of false positives is too high;
   (iv) Conformation bias: if click data are similar within a community.

   This list is only meant to serve as a brief orientation. To detect and mitigate biases in recommendation results, popularity and fairness metrics such as statistical parity (or group fairness; Biega et al., 2018), precision, and recall (Wang & Blei, 2011) need to be determined as well. For a detailed explanation on how to figure out the types of bias a recommender system faces, please refer to our descriptions in the "Recommendation bias" section.
2. **Selection of mitigation method(s):** Once the bias(es) has been found, one can refer to Tables 1 and 2, as well as to a detailed description in the "Bias mitigation in scholarly recommender systems" section, to select suitable mitigation methods. Developers can choose between methods that are performed before, during, or after data processing.

Table 1 Mitigation methods for human-caused biases

Bias	Stage in rec. sys.	Mitigation method	References
Conformation bias (Unreliable feedback data)	In	Causal embedding	Zheng et al. (2020)
		Other methods	Wang et al. (2017), Chaney et al. (2015), Tang et al. (2012), Ma et al. (2009)
Position bias (unreliable click data)	In	Propensity score	Agarwal et al. (2019a), Fang et al. (2019), Agarwal et al. (2019b), Ai et al. (2018), Joachims et al. (2017b), Hu et al. (2017), Swaminathan et al. (2015), Raman and Joachims (2013), Chapelle et al. (2012)
	Other	Click models	Vardasbi et al. (2020), Borisov et al. (2016), Shen et al. (2012), Xu et al. (2012a), Chapelle and Zhang (2009), Dupret and Piwowarski (2008), Craswell et al. (2008)
Selection bias (unreliable positive feedback data)	In	Propensity score	Saito (2020), Schnabel et al. (2016)
	Post	ATOP	Lim et al. (2015), Steck (2010)

Table 2 Mitigation methods for system-caused biases

Bias	Stage in rec. sys.	Mitigation method	References
Exposure bias (unreliable negative feedback data)	In	Confidence weights & other heuristics	Saito (2020), Chen et al. (2018), Sidana et al. (2018), Lian et al. (2017), He et al. (2016), Li et al. (2010), Pan and Scholz (2009), Hu et al. (2008), Pan et al. (2008)
		Sampling	Liu et al. (2019), Zhang et al. (2018), Karvelis et al. (2018), Caselles-Dupré et al. (2018), Rendle et al. (2012)
		Exposure-based model	Pradhan and Pal (2020), Liang et al. (2016)
	Post	Propensity score	Gupta et al. (2021), Yang et al. (2018)
Popularity bias (popular items get recommended more often)	In	Regularization	Abdollahpouri et al. (2021), Chen et al. (2020b), Abdollahpouri et al. (2017)
	In	Adversarial learning	Krishnan et al. (2018)
	In	Propensity score	Yang et al. (2018)
	Post	Personalized long-tail promotion	Abdollahpouri et al. (2019)
Unfairness (unfairness to certain groups)	Pre	Re-balancing/re-sampling	Cem Geyik et al. (2019), Asudeh et al. (2019), Biega et al. (2018), Pedreshi et al. (2008)
	In	Regularization	Morik et al. (2020), Beutel et al. (2019), Burke et al. (2018), Kamishima and Akaho (2017), Yao and Huang (2017), Abdollahpouri et al. (2017), Kamishima et al. (2016), Zemel et al. (2013), Kamishima et al. (2013)
	In	Adversarial learning	Bose and Hamilton (2019), Edwards and Storkey (2015)
	In	Causal modeling	Zhang and Bareinboim (2018), Nabi and Shpitser (2018), Wu et al. (2018), Kusner et al. (2017)
	In	Multi-level citation networks	Son and Kim (2018)
	Post	Fairness in ranking	Biega et al. (2020, 2018), Yang and Stoyanovich (2017), Zehlike et al. (2017)

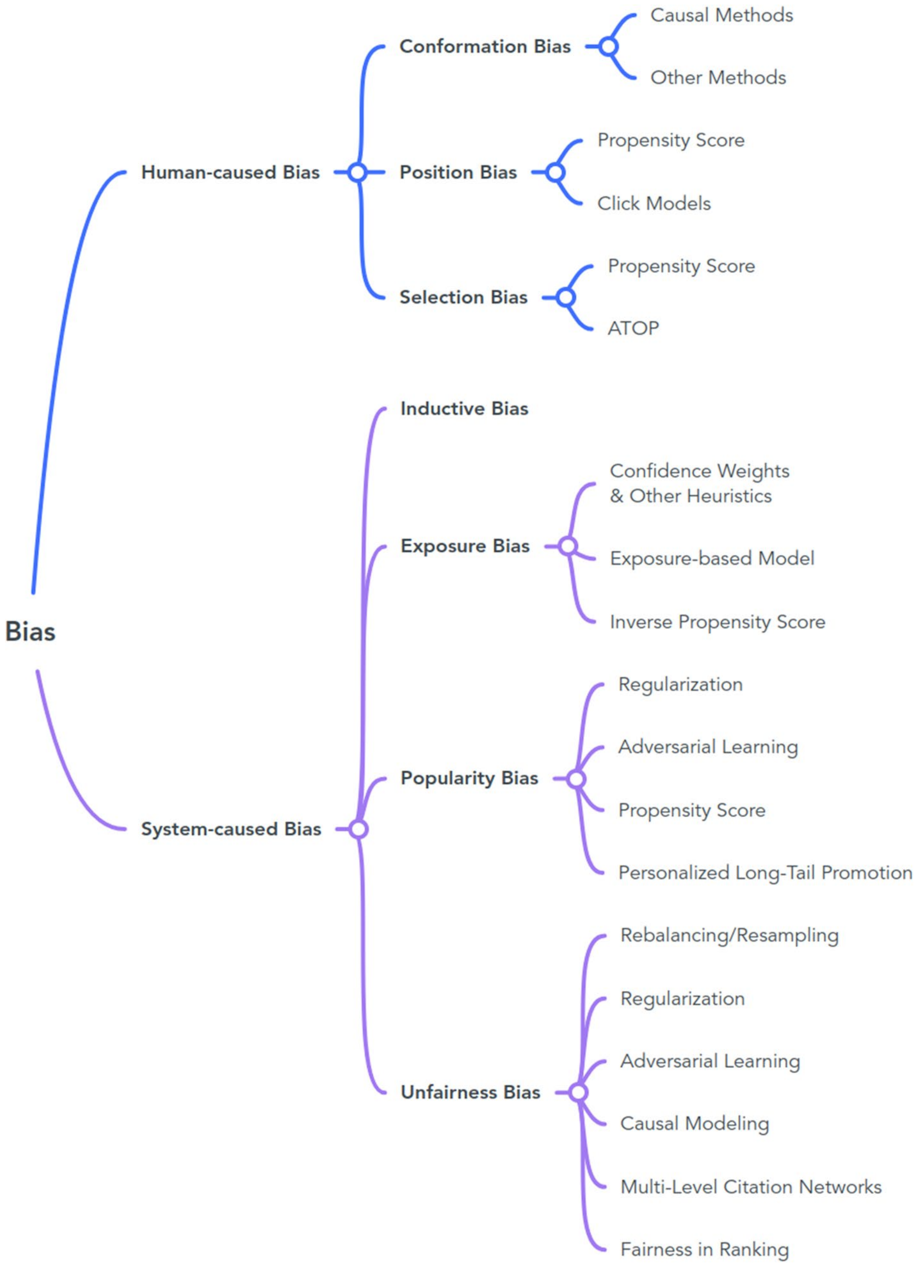


Fig. 6 Overview of identified bias mitigation methods

This helps mitigate bias in the data collection, training, and recommendation steps (see the " Categories of bias in scholarly recommender systems" section).

3. **Model assessment and evaluation:** After applying the appropriate mitigation method, a recommender systems model can be evaluated against baselines to evaluate whether its recommendation results are debiased and its performance changed.

Example Use Case 1: To illustrate the need for and usage of a debiasing framework for scholarly recommender systems, let us consider the following scenario. Data scientists at the digital library ACM[18] might notice a decline in the click-through rate (CTR) of their recommender system. They perform an offline evaluation and find that their model fails to beat simple baselines, concluding that their recommendation algorithm is still biased despite using general techniques like regularization during model training. They realize that they need a comprehensive framework to debias their recommender system with respect to several biases. By analyzing the click data, they find that the click data is highly imbalanced, but that it forms clusters. Thus, both *position bias* and *conformation bias* need to be addressed. Our framework (see Tables 1 and 2) then reveals that *position bias* can be reduced by introducing a propensity score (e.g., Agarwal et al. 2019a; Fang et al. 2019; Agarwal et al. 2019b) or by using click models (e.g., Vardasbi et al. 2020), while *conformation bias* can be reduced in various ways, such as by means of causal embeddings (Zheng et al., 2020) as one of the most recent approaches.
Example Use Case 2: Finding suitable researchers to review papers is often not an easy task for editors. Several academic publishers (e.g., Elsevier) offer online recommendation systems to find suitable reviewers for a given paper. However, since the editor's selection of reviewers can be slightly biased (e.g., selecting people who already have a high h-index or have already reviewed many papers in the journal, i.e., popular people), it makes sense to determine suitable debiasing methods. The present case indicates that the *popularity bias* should be reduced. In the event that the publisher does not develop the recommendation system itself, but only has a license and obtains the recommendations via an API, a mitigation method during post-processing should be looked into. In this case, our framework (Table 2, Popularity Bias, Stage "post") suggests that the personalized long-tail promotion method of Abdollahpouri et al. (2019) can be used by the publisher to re-rank the results before displaying them to the users.

Future directions and open challenges

In this section, we discuss the open challenges that scholars might face while building bias-aware recommender systems for academia, and hope that this will stimulate future research concerning scholarly biases and their mitigation in scholarly recommender systems. We do this by transferring the challenges mentioned by Chen et al. (2020a) to the scholarly domain and by introducing additional aspects, such as using scholarly knowledge graphs to provide explainable scholarly recommender systems.

Nontrivial Statistical Techniques: Inverse propensity scoring (IPS) is a technique widely used to provide an unbiased estimate of the ranking metric of interest and to re-weight clicked items (Wang et al., 2021b). IPS-based methods are the most frequently used bias mitigation method in recommender systems in general (Chen et al., 2020a). However,

[18] https://dl.acm.org/.

these methods are only effective if the propensities are correctly estimated. An important question here is how propensity scores can be estimated accurately. Although some proposed methods provide accurate estimation given a simple bias, such as position bias (Swaminathan et al., 2015; Chapelle et al., 2012), it can be difficult to estimate propensity in the presence of complicated biases, such as exposure and selection biases (Lee et al., 2021). Unlike position bias, these do not simply rely on the position of an item, but on many other factors such as item popularity, availability (open access vs non-open access items), and feedback loops. Most publications on bias in our paper collection do not address this issue. Therefore, we support Chen et al. (2020a)'s argument that further research needs to be performed in this area, particularly in the scholarly field with its typical sparse data (e.g., limited amount of click data).

Addressing Biases Simultaneously: We have noticed that the various methods proposed to mitigate bias in recommender systems usually address only one or at most two biases, while in reality multiple types of bias might occur simultaneously. For instance, to decide whether to download a paper, a researcher might be influenced by the people in her scientific community due to conformation and popularity biases. We found few papers that address multiple biases simultaneously. Chen et al. (2021), for instance, used meta-learning techniques to learn the debiasing parameters from training data. Jin et al. (2020) provided a transfer learning approach that mitigates one or multiple biases in downstream classifiers by transfer learning from an upstream model. To tackle the issue of multiple bias types systematically, two solutions seem to be viable: (1) a general debiasing framework (which we provide in this article) to identify all necessary bias mitigation methods that need to be applied one by one; or (2) a mitigation method that can tackle multiple biases simultaneously—and ideally, model and recognize dependencies between biases.

Addressing Bias in Practice: While debiasing methods have been proposed in research, methodologies and guidelines for practitioners to address bias in real use cases (e.g., software developers of recommender systems and data scientists) is largely missing. While we provide a debiasing framework in this article, we see a need to provide more software libraries to identify and mitigate biases.

Addressing Bias in Evaluation: Evaluating bias is a tricky endeavor, as detecting and mitigating specific biases depends greatly on the available data. Most evaluation methods either require accurate propensity scores or unbiased click data. Both can be difficult to obtain in implicit feedback data. Uniform data is sampled with equally-sized sets from each class. Although uniform data provide unbiased information, their small scale and price are insufficient and expensive to evaluate recommender systems, mainly due to high variance found in the data. Moreover, due to the increasing presence of popularity bias and unfairness in the data (see the "Recommendation bias" section), many works propose different kinds of evaluation methods, which can be inconsistent across different kinds of results. Therefore, more research into formulating new metrics for evaluating recommender systems is needed, with or without biased click data.

Deeper Understanding and Modeling of User Behavior: Recommender system biases are caused by human factors, as humans use the recommender systems and typically provide historical data for model training. Thus, we argue that it is worthwhile to model and understand the users themselves (e.g., researchers grouped by experience and scientific discipline) in their daily working environments. For instance, we can assume that paper and citation recommendation systems will be well received in the future. However, these systems are assumed to be prone to biases (Färber & Jatowt, 2020).

Specifically, citation bias has been analyzed in two main contexts in the literature: to explain the scholars' self-citation behavior, and to show that scholars cite papers but disproportionally criticize papers or specific claimsless often. While the first phenomenon has been addressed in a few journal reports (Van Noorden and Singh Chawla, 2019; Aksnes, 2003), little literature can be found to tackle the latter issue, which deals with the context in which a scholar cites. To solve this problem, it is important to understand scholars' citation behavior.

Explainable Recommender Systems: Making recommender systems more explainable has been considered key in recent years. This is particularly important in the scholarly field, one which requires a high level of trust. With respect to biases, explanations of recommendations can help users make informed decisions—facilitating human agency—and can reduce certain biases via human-computer interaction. Existing frameworks and ethics guidelines for trustworthy AI (Jobin et al., 2019), such as the one provided by the European Union (Commission, 2020), can be implemented here.

In addition, scholarly knowledge graphs are worth mention as a way of modeling academic knowledge explicitly, paving the way for scholarly knowledge graph-based recommender systems (Ayala-Gómez et al., 2018). For instance, the Microsoft Academic Knowledge Graph (MAKG)[19] (Färber, 2019) contains eight billion triples about publications and associated entities, such as authors, venues, and affiliations. Wikidata (https://wikidata.org), OpenCitations (Peroni & Shotton, 2020), the Open Research Knowledge Graph (ORKG) (Jaradeh et al., 2019), and AIDA (Angioni et al., 2021), among others, are further noteworthy knowledge graphs.

Conclusion

In this paper, we provided readers with an overview of biases present in scholarly recommender systems. In particular, we outlined the impact and prevalence of each type of bias. We then investigated which of the published bias mitigation methods are particularly relevant in the context of scholarly recommender systems. We noticed that scholarly biases have been rather underexplored so far, particularly in the context of scholarly recommender systems. It became clear that we easily run the risk of leaving biases undetected in academia, potentially disadvantaging millions of researchers. Therefore, we proposed a simple-kept framework that readers can follow to help them choose the right mitigation method for their scholarly recommender systems. Last but not least, we made suggestions for future research in this area.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

[19] Microsoft Academic website and underlying APIs was retired on Dec. 31, 2021. https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/.

References

Abdollahpouri, H., & Mansoury, M. (2020). Multi-sided Exposure Bias in Recommendation. *CoRR*, arxiv:2006.15772

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *RecSys 2017-Proceedings of the 11th ACM Conference on Recommender Systems*, ACM, pp. 42–46, https://doi.org/10.1145/3109859.3109912

Abdollahpouri, H., Burke, R., & Mobasher, B. (2019). Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the 32nd International Flairs Conference*

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., & Malthouse, E. (2021). User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP),* ACM, p. 12, https://doi.org/10.1145/3450613.3456821

Adie, E., & Roe, W. (2013). Altmetric: Enriching scholarly content with article-leveldiscussion and metrics. *Learned Publishing, 26*(1), 11–17. https://doi.org/10.1087/20130103

Agarwal, A., Zaitsev, I., Takatsu, K., & Joachims, T. (2019a). A general framework for counterfactual learning-to-rank. In *SIGIR 2019-Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Vol. 10, pp. 5–14, https://doi.org/10.1145/3331184.3331202

Agarwal, A., Zaitsev, I., Wang, X., Li, C., Najork, M., & Joachims, T. (2019b). Estimating position bias without intrusive inter-ventions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, ACM, https://doi.org/10.1145/3289600.3291017

Ai, Q., Bi, K., Luo, C., Guo, J., & Croft, W. B. (2018). Unbiased learning to rank with unbiased propensity estimation. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2018. pp. 385–394, https://doi.org/10.1145/3209978.3209986

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics, 56*(2), 235–246. https://doi.org/10.1023/A:1021919228368

Analytis, P. P., Barkoczi, D., Lorenz-Spreen, P., & Herzog, S. M. (2020). The structure of social influence in recommender networks. In *The Web Conference 2020-Proceedings of the World Wide Web Conference, WWW 2020*, ACM, https://doi.org/10.1145/3366423.3380020

Andersen, J. P., & Nielsen, M. W. (2018). Google scholar and web of science: Examining gender differences in citation coverage across five scientific disciplines. *Journal of Informetrics, 12*, 950–959. https://doi.org/10.1016/j.joi.2018.07.010

Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., & Motta, E. (2021). Aida: A knowledge graph about research dynamics in academia and industry. *Quantitative Science Studies, 2*(4), 1356–1398.

Asabere, N. Y., Xia, F., Meng, Q., Li, F., & Liu, H. (2014). Scholarly paper recommendation based on social awareness and folksonomy. *International Journal of Parallel, Emergent and Distributed System, 11*, 211–232. https://doi.org/10.1080/17445760.2014.904859

Asudeh, A., Jagadish, H. V., Stoyanovich, J., & Das, G. (2019). Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, ACM, https://doi.org/10.1145/3299869.3300079

Ayala-Gómez, F., Daróczy, B., Benczúr, A., Mathioudakis, M., & Gionis, A. (2018). Global citation recommendation using knowledge graphs. *Journal of Intelligent and Fuzzy Systems, 34*(5), 3089–3100. https://doi.org/10.3233/JIFS-169493

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American, 288*(5), 60–69.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., & Faulkner, R. et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261

Beel, J., & Gipp, B. (2009). Academic search engines, google scholar, ranking algorithm. In *12th International Conference on Scientometrics and Informetrics (ISSI'09)* (pp. 230–241). International Society for Scientometrics and Informetrics

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries, 17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Berger, S., Feldhaus, C., & Ockenfels, A. (2018). A shared identity promotes herding in an information cascade game. *Journal of the Economic Science Association, 4*(1), 63–72. https://doi.org/10.1007/s40881-018-0050-9

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., & Goodrow, C. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, https://doi.org/10.1145/3292500.3330745

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, ACM, pp. 405–414, https://doi.org/10.1145/3209978.3210063

Biega, A. J., Diaz, F., Ekstrand, M. D., & Kohlmeier, S. (2020). Overview of the TREC 2019 fair ranking track. *CoRR*arxiv:2003.11650v1

Bogers, T., & Van Den Bosch, A. (2008). Recommending scientific articles using citeULike. In *RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems*, ACM Press, pp. 287–290, https://doi.org/10.1145/1454008.1454053

Borisov, A., Markov, I., De Rijke, M., & Serdyukov, P. (2016). A neural click model for web search. In *25th International world wide web conference, WWW 2016, International World Wide Web Conferences Steering Committee*, pp. 531–541, https://doi.org/10.1145/2872427.2883033

Bose, A., & Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. In *36th International Conference on Machine Learning, PMLR*, pp. 715–724, https://proceedings.mlr.press/v97/bose19a.html

Bruns, A. (2017). Echo chamber? What echo chamber? Reviewing the evidence. In *6th Biennial Future of Journalism Conference* (FOJ17)

Bruns, A. (2019). Filter bubble. *Internet Policy Review*. https://doi.org/10.14763/2019.4.1426

Burke, R., Sonboli, N., & Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of Machine Learning Research, PMLR*, Vol. 81, pp. 202–214, https://proceedings.mlr.press/v81/burke18a.html

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, NIPS'17

Caplar, N., Tacchella, S., & Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy, 1*(6), 1–5. https://doi.org/10.1038/s41550-017-0141

Caselles-Dupré, H., Lesaint, F., Royo-Letelier, J., & Royo-Letelier, J. (2018). Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3240323.3240377

Cem Geyik, S., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, https://doi.org/10.1145/3292500.3330691

Center, PR. (2015). How scientists engage the public. https://www.pewresearch.org/science/2015/02/15/how-scientists-engage-public/

Chaney, AJ., Blei, DM., & Eliassi-Rad, T. (2015). A probabilistic model for using social networks in personalized item recommendation. In *RecSys 2015-Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, pp. 43–50, https://doi.org/10.1145/2792838.2800193

Chapelle, O., & Zhang, Y. (2009). A dynamic Bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web-WWW '09*, ACM Press, https://doi.org/10.1145/1526709

Chapelle, O., Joachims, T., Radlinski, F., & Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS), 30*(1), 45. https://doi.org/10.1145/2094072.2094078

Chen, C., Zhang, M., Liu, Y., & Ma, S. (2018). Missing data modeling with user activity and item popularity in recommendation. In *Proceedings of the 14th Asia Information Retrieval Societies Conference*, Springer, AIRS'18, pp. 113–125, https://doi.org/10.1007/978-3-030-03520-4_11

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020a). Bias and Debias in recommender system: A survey and future directions. pp. 1–20, arxiv:2010.03240

Chen, J., Dong, H., Qiu, Y., He, X., Xin, X., Chen, L., Lin, G., & Yang, K. (2021). AutoDebias: Learning to Debias for Recommendation. In *SIGIR 2021-Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Vol. 10, pp. 21–30, https://doi.org/10.1145/3404835.3462919

Chen, Z., Xiao, R., Li, C., Ye, G., Sun, H., & Deng, H. (2020b). ESAM: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM. p. 10, https://doi.org/10.1145/3397271.3401043

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591–621.

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports, 8*(1), 1–7. https://doi.org/10.1038/s41598-018-34203-2

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. In *Proceedings of the National Academy of Sciences*. Vol. 118(9)

Collins, A., Tkaczyk, D., Aizawa, A., & Beel, J. (2018). A study of position bias in digital library recommender systems. arxiv:1802.06565

Commission, E. (2020). Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair news-feature. *Nature, 558*(7710), 357–360. https://doi.org/10.1038/d41586-018-05469-3

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining-WSDM '08*https://doi.org/10.1145/1341531

D'Alessandro, B., O'Neil, C., & Lagatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data, 5*(2), 120–134. https://doi.org/10.1089/big.2016.0048

DeBruine, B., & Lisa, J. (2017). Face research lab London set. https://doi.org/10.6084/m9.figshare.5047666.v3

Delgado-Rodriguez, M., & Llorca, J. (2004). Journal of Epidemiology & Community Health. *Bias, 58*(8), 635–641.

Dupret, G., & Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 331–338, https://doi.org/10.1145/1390334.1390392

Duyx, B., Urlings, M. J., Swaen, G. M., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology, 88*, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Board, Editorial. (2018). Social media for scientists. *Nature Cell Biology, 20*(12), 1329. https://doi.org/10.1038/s41556-018-0253-6

Edwards, H., & Storkey, A. (2015). Censoring representations with an adversary. In *4th international conference on learning representations, ICLR 2016 Conference Track Proceedings, International Conference on Learning Representations, ICLR*, arxiv:1511.05897v3

Fang, D., Moy, E., Colburn, L., & Hurley, J. (2000). Racial and ethnic disparities in faculty promotion in academic medicine. *JAMA, 284*(9), 1085–1092. https://doi.org/10.1001/JAMA.284.9.1085

Fang, Z., Agarwal, A., & Joachims, T. (2019). Intervention harvesting for context-dependent examination-bias estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, SIGIR 2019, pp. 825–834, https://doi.org/10.1145/3331184.3331238

Färber, M. (2019). The Microsoft Academic knowledge graph: a linked data source with 8 billion triples of scholarly data. In *Proceedings of the 18th International Semantic Web Conference*, ISWC'19, pp. 113–129, https://doi.org/10.1007/978-3-030-30796-7_8

Färber, M., & Jatowt, A. (2020). Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries, 21*(4), 375–405. https://doi.org/10.1007/s00799-020-00288-2

Forsati, R., Barjasteh, I., & Esfahanian, A. H. (2017). Semi-supervised collaborative ranking with push at the top. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2017, ACM, pp. 401–408, https://doi.org/10.1145/3110025.3110144

Gai, L., & Lei, L. (2014). Dual collaborative topic modeling from implicit feedbacks. In *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics*, SPAC 2014, IEEE, pp. 395–404, https://doi.org/10.1109/SPAC.2014.6982723

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4*(2), 133–151. https://doi.org/10.1023/A:1011419012209

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ, 339*(7714), 210–213. https://doi.org/10.1136/bmj.b2680

Gupta, S., Wang, H., Lipton, Z. C., & Wang, Y. (2021). Correcting exposure bias for link recommendation. In *Proceedings of the 38th International Conference on Machine Learning*, ICML'21

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE, 10*(9), e0138237.

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., & Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review, 34*(1), 36–40. https://doi.org/10.1080/00987913.2008.10765150

He, X., Zhang, H., Kan, M. Y., & Chua, T. S. (2016). Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR 2016-Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 549–558, https://doi.org/10.1145/2911451.2911489

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In *26th International World Wide Web Conference, WWW 2017, International World Wide Web Conferences Steering Committee*, pp. 173–182, https://doi.org/10.1145/3038912.3052569

Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: Lessons learned at the national science foundation. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 862–871, https://doi.org/10.1145/1150402.1150521

Holman, L., Stuart-Fox, D., & Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented? *PLoS Biology, 16*(4), 8–65. https://doi.org/10.1371/JOURNAL.PBIO.2004956

Hu, Y., Volinsky, C., & Koren, Y. (2008). Collaborative filtering for implicit feedback datasets. In *IEEE*, ICDM'08, pp. 263–272, https://doi.org/10.1109/ICDM.2008.22

Hu, Z., Wang, Y., Peng, Q., & Li, H. (2017). A novel algorithm for unbiased learning to rank. *CoRR*

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences, 117*(9), 4609–4616. https://doi.org/10.1073/PNAS.1914221117

Islam, R., Keya, KN., Zeng, Z., Pan, S., & Foulds, J. (2021). Debiasing career recommendations with neural fair collaborative filtering. In *Proceedings of the World Wide Web Conference*, pp. 3779–3790

Jaradeh, MY., Oelen, A., Farfar, KE., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., & Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP'19, pp. 243–246, https://doi.org/10.1145/3360901.3364435

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, AIES'19, pp. 383–390, https://doi.org/10.1145/3306618.3314288

Jin, X., Barbieri, F., Davani, A. M., Kennedy, B., Neves, L., & Ren, X. (2020). Efficiently mitigating classification bias via transfer learning. *CoRR*, arxiv:2010.12864

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum, 51*(1), 4–11. https://doi.org/10.1145/3130332.3130334

Joachims, T., Swaminathan, A., & Schnabel, T. (2017b). Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, ACM, WSDM'17, pp. 781–789, https://doi.org/10.1145/3018661.3018699

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kamishima, T., & Akaho, S. (2017) Considerations on recommendation independence for a find-good-items task. In *Workshop on Responsible Recommendation*, Vol. 6, https://doi.org/10.18122/B2871W

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2013). Efficiency improvement of neutrality-enhanced recommendation. Tech. rep.

Kamishima, T., Akaho, S., Asoh, H., & Sato, I. (2016). Model-based approaches for independence-enhanced recommendation. In *Proceedings of the 16th International Conference on Data Mining Workshops, IEEE*, https://doi.org/10.1109/ICDMW.2016.23

Karvelis, P., Gavrilis, D., Georgoulas, G., & Stylios, C. (2018). Topic recommendation using Doc2Vec. In *Proceedings of the International Joint Conference on Neural Networks*. 2018, https://doi.org/10.1109/IJCNN.2018.8489513

Klamma, R., Pham, M. C., & Cao, Y. (2009). You never walk alone: Recommending academic events based on social network analysis. In *Proceedings of the First International Conference on Complex Sciences*, Springer, Complex'09, pp. 657–670, https://doi.org/10.1007/978-3-642-02466-5_64

Klöckner, K., Wirschum, N., & Jameson, A. (2004). Depth- and breadth-first processing of search result lists. Tech. rep.

Krishnan, A., Sharma, A., Sankar, A., & Sundaram, H. (2018). An adversarial approach to improve long-tail performance in neural collaborative filtering-1.5pt. In *Proceedings of the 27th ACM International*

Conference on Information and Knowledge Management, ACM https://doi.org/10.1145/3269206.3269264

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st Conference on Neural Information Processing Systems, Neural information Processing Systems Foundation*, pp. 4067–4077, arxiv:1703.06856v3

Lee, D. H., & Brusilovsky, P. (2011). Improving recommendations using watching networks in a social tagging system

Lee, J. W., Park, S., & Lee, J. (2021). Dual unbiased recommender learning for implicit feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, SIGIR'21, pp. 1647–1651, https://doi.org/10.1145/3404835.3463118

Lee, R. V. D., & Ellemers, N. (2015). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences, 112*(40), 12349–12353. https://doi.org/10.1073/PNAS.1510159112

Lerback, J., & Hanson, B. (2017). Journals invite too few women to referee. *Nature, 541*(7638), 455–457. https://doi.org/10.1038/541455a

Li, Y., Hu, J., Zhai, C., & Chen, Y. (2010). Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management-CIKM '10* https://doi.org/10.1145/1871437

Lian, D., Ye, Y., Zhu, W., Liu, Q., Xie, X., & Xiong, H. (2017). Mutual reinforcement of academic performance prediction and library book recommendation. In *Proceedings-IEEE International Conference on Data Mining*, ICDM, pp. 1023–1028, https://doi.org/10.1109/ICDM.2016.105

Liang, D., Charlin, L., McInerney, & J., Blei, D. M. (2016). Modeling user exposure in recommendation. In: *Proceedings of the 25th International World Wide Web Conference, WWW'16*, pp. 951–961, https://doi.org/10.1145/2872427.2883090

Lim, D., Mcauley, J., & Lanckriet. G. (2015). Top-N recommendation with missing implicit feedback. In *Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, https://doi.org/10.1145/2792838.2799671

Liu, D., Cheng, P., Dong, Z., He, X., Pan, W., & Ming, Z. (2020). A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, SIGIR 2020, pp. 831–840, https://doi.org/10.1145/3397271.3401083

Liu, Y., Cao, X., & Yu, Y. (2016). Are you influenced by others when rating? Improve rating prediction by conformity modeling. In *Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, https://doi.org/10.1145/2959100.2959141

Liu, Y., Tian, Z., Sun, J., Jiang, Y., & Zhang, X. (2019). Distributed representation learning via node2vec for implicit feedback recommendation. *Neural Computing and Applications, 32*(9), 4335–4345. https://doi.org/10.1007/S00521-018-03964-2

Ma, H., King, I., & Lyu, MR. (2009). Learning to recommend with social trust ensemble. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, SIGIR'09, https://doi.org/10.1145/1571941.1571978

Mansoury, M. (2021). Understanding and mitigating multi-sided exposure bias in recommender systems. arXiv:2111.05564

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. In *Proceedings of the International Conference on Information and Knowledge Management*, ACM, pp. 2145–2148, https://doi.org/10.1145/3340531.3412152

Marlin, B., Zemel, RS., Roweis, S., & Slaney, M. (2012). Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*. pp. 267–275, arxiv:1206.5267

Marlin, B. M., & Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems-RecSys '09*, ACM Press

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google scholar, microsoft academic, scopus, dimensions, web of science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics, 126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Matías-Guiu, J., & García-Ramos, R. (2011). Editorial bias in scientific publications. *Neurología (English Edition), 26*(1), 1–5. https://doi.org/10.1016/s2173-5808(11)70001-3

Maxwell, H. (2015). The MovieLens datasets. *ACM Transactions on Interactive Intelligent Systems (TiiS), 5*(4), 58. https://doi.org/10.1145/2827872

May, R. M. (1997). The scientific wealth of nations published by : American association for the advancement of science the scientific wealth of nations. *Science, 275*(5301), 793–796.

McClelland, J. (1992). The interaction of nature and nurture in development: A parallel distributed processing perspective (parallel distributed processing and cognitive neuroscience pdp. cns. 92.6)

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., & Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM, pp. 116–125, https://doi.org/10.1145/587078.587096

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mena-Maldonado, E., Cañamares, R., Castells, P., Ren, Y., & Sanderson, M. (2021). Popularity bias in false-positive metrics for recommender systems evaluation. *ACM Transactions on Information Systems (TOIS), 39*(3), 1–43. https://doi.org/10.1145/3452740

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–62. https://doi.org/10.1126/science.159.3810.56

Mitchell, T. M. (1980). The need for biases in learning generalizations. Tech. rep.

Mohammad, S. M. (2020). Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, ACL 2020, pp. 7860–7870, https://doi.org/10.18653/v1/2020.acl-main.702

Morik, M., Singh, A., Hong, J., & Joachims, T. (2020). Controlling fairness and bias in dynamic learning-to-rank. In *SIGIR 2020-Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Vol. 10, pp. 429–438, https://doi.org/10.1145/3397271.3401100

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science, 341*(6146), 647–651. https://doi.org/10.1126/science.1240466

Naak, A., Hage, H., & Aïmeur, E. (2009). A multi-criteria collaborative filtering approach for research paper recommendation in papyres. In *Proceedings of the 4th International Conference on E-Technologies: Innovation in an Open World*, Springer, MCETECH'09, Vol. 26, pp. 25–39, https://doi.org/10.1007/978-3-642-01187-0_3

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, https://ojs.aaai.org/index.php/AAAI/article/view/11553

Nguyen, T. T., Hui, P. M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. In *WWW 2014-Proceedings of the 23rd International Conference on World Wide Web*, ACM, pp. 677–686, https://doi.org/10.1145/2566486.2568012, https://doi.org/10.1145/2566486.2568012

Nielsen, M. W. (2016). Limits to meritocracy? Gender in academic recruitment and promotion processes. *Science and Public Policy, 43*(3), 386–399. https://doi.org/10.1093/SCIPOL/SCV052

OECD. (2021). Main Science and Technology Indicators, Vol. 2021. OECD Publishing

Ojasoo, T., & Doré, J. (1999). Citation bias in medical journals. *Scientometrics, 45*(1), 81–94.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown

Ovaisi, Z., Ahsan, R., Zhang, Y., Vasilaky, K., & Zheleva, E. (2020). Correcting for selection bias in learning-to-rank systems. In *Proceedings of the World Wide Web Conference 2020*, ACM, WWW'20, pp. 1863–1873, https://doi.org/10.1145/3366423.3380255

Pan, R., & Scholz, M. (2009). Mind the gaps: Weighting the unknown in large-scale one-class collaborative filtering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 667–675, https://doi.org/10.1145/1557019.1557094

Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., & Yang, Q. (2008). One-class collaborative filtering. In *Proceedings of the IEEE International Conference on Data Mining*. pp. 502–511, https://doi.org/10.1109/ICDM.2008.16

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you

Park, Y. J., & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems-RecSys '08*, ACM Press

Parker, J. N., Lortie, C., & Allesina, S. (2010). Characterizing a scientific elite: The social characteristics of the most highly cited scientists in environmental science and ecology. *Scientometrics, 85*(1), 129–143. https://doi.org/10.1007/s11192-010-0234-4

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568, https://doi.org/10.1145/1401890.1401959

Pennock, D. M., Horvitz, E., Lawrence, S., & Giles, C. L. (2000). Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Morgan Kaufmann, UAI '00*, pp. 473–480

Peroni, S., & Shotton, D. M. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies, 1*(1), 428–444. https://doi.org/10.1162/qss_a_00023

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 701–710, https://doi.org/10.1145/2623330.2623732

Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2014). Reputation and Impact in Academic Careers. *Proceedings of the National Academy of Sciences, 111*(43), 15316–15321. https://doi.org/10.1073/PNAS.1323111111

Piramuthu, S., Kapoor, G., Zhou, W., & Mauw, S. (2012). Input online review data and related bias in recommender systems. *Decision Support Systems, 53*(3), 418–424. https://doi.org/10.1016/j.dss.2012.02.006

Polonioli, A. (2020). The ethics of scientific recommender systems. *Scientometrics, 126*(2), 1841–1848. https://doi.org/10.1007/S11192-020-03766-1

Polyzou, A. (2020). Models and algorithms for performance prediction and course recommendation in higher education. PhD thesis

Pradhan, T., & Pal, S. (2020). CNAVER: A content and network-based academic venue recommender system. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2019.105092

Publons, (2018). Global state of peer review. https://publons.com/community/gspr

Raman, K., & Joachims, T. (2013). Learning socially optimal information systems from egoistic users. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, ECML-PKDD'13, Vol. 8189, pp. 128–144, https://doi.org/10.1007/978-3-642-40991-2_9

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pp. 452–461, arxiv:1205.2618

Saier, T., Färber, M., & Tsereteli, T. (2022). Cross-lingual citations in English papers: A large-scale analysis of prevalence, usage, and impact. *International Journal on Digital Libraries, 23*(2), 179–195.

Saito, Y. (2020). Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, SIGIR'20, pp. 309–318, https://doi.org/10.1145/3397271.3401114

Salman, O., Gauch, S., Alqahtani, M., Salah Ibrahim, M., Alqahatani, M., Ibrahim, M., & Alsaffar, R. (2020). Incorporating diversity in academic expert recommendation. In *Proceedings of the 12th International Conference on Information, Process, and Knowledge Management*, eKNOW'20

Sapiezynski, P., Kassarnig, V., Wilson, C., Lehmann, S., & Mislove, A. (2017). Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the FATREC Workshop on Responsible Recommendation, FATREC'17*

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review, 107*(5), 141–45. https://doi.org/10.1257/AER.P20171126

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. Tech. rep.

Shah, N. B. (2022). An overview of challenges, experiments, and computational solutions in peer review. *Communications of the ACM, 65*(6), 76–87.

Shen, S., Hu, B., Chen, W., & Yang, Q. (2012). Personalized click model through collaborative filtering. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining-WSDM '12*, ACM Press https://doi.org/10.1145/2124295

Sidana, S., Laclau, C., & Amini, M. R. (2018). Learning to recommend diverse items over implicit feedback on PANDOR. In *Proceedings of the 12th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3240323.3240400

Smetanin, S., & Komarov, M. (2022). Misclassification bias in computational social science: A simulation approach for assessing the impact of classification errors on social indicators research. *IEEE Access, 10*, 18886–18898. https://doi.org/10.1109/ACCESS.2022.3149897

Son, J., & Kim, S. B. (2018). Academic paper recommender system using multilevel simultaneous citation networks. *Decision Support Systems, 105*, 24–33. https://doi.org/10.1016/j.dss.2017.10.011

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1–193.

Steck, H. (2010). Training and testing of recommender systems on data missing not at random. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, pp. 713–722, https://doi.org/10.1145/1835804.1835895

Stinson, C. (2021). Algorithms are not neutral: Bias in collaborative filtering. arxiv:2105.01031

Sugiyama, K., & Kan, M. Y. (2013). Exploiting potential citation papers in scholarly paper recommendation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, ACM Press, pp. 153–162, https://doi.org/10.1145/2467696.2467701

Sun, W., Khenissi, S., Nasraoui, O., & Shafto, P. (2019). Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion of The 2019 World Wide Web Conference*, ACM, WWW'19, pp. 645–651, https://doi.org/10.1145/3308560.3317303

Sun, Y., & Giles, C. L. (2007). Popularity weighted ranking for academic digital libraries. In *Proceedings of the 29th European Conference on IR*, Springer, ECIR 2007, pp. 605–612, https://doi.org/10.1007/978-3-540-71496-5_57

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Swaminathan, A., Joachims, T., Gammerman, A., & Vovk, V. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research, 16*, 1731–1755.

Tang, J., Gao, H., & Liu, H. (2012). mTrust: Discerning multi-faceted trust in a connected world. In *Proceedings of the fifth ACM international conference on Web search and data mining-WSDM '12*, ACM Press http://www.epinions.com/user-nancy35c

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, ACM, WWW 2015, pp. 1067–1077, https://doi.org/10.1145/2736277.2741093

Tang, T. Y., & McCalla, G. (2009). A multidimensional paper recommender: Experiments and evaluations. *IEEE Internet Computing, 13*(4), 34–41. https://doi.org/10.1109/MIC.2009.73

Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens. In *Proceedings of the ACM IEEE International Conference on Digital Libraries, JCDL 2004*, ACM, pp. 228–236, https://doi.org/10.1145/996350.996402

Torres-Salinas, D., Cabezas-Clavijo, A., & Jimenez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in web 2.0. *Comunicar,* 53–60. https://doi.org/10.3916/c41-2013-05

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Van Noorden, R., & Singh Chawla, D. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature*. https://doi.org/10.1038/d41586-019-02479-7

Vanclay, J. (2009). Bias in the journal impact factor. *Scientometrics, 78*, 3–12. https://doi.org/10.1007/s11192-008-1778-4

Vardasbi, A., De Rijke, M., & Markov, I. (2020). Cascade model-based propensity estimation for counterfactual learning to rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, https://doi.org/10.1145/3397271.3401299

Vellino, A. (2013). Usage-based vs. citation-based methods for recommending scholarly research articles. arxiv:1303.7149

Vellino, A. (2015). Recommending research articles using citation data. *Library Hi Tech, 33*(4), 597–609. https://doi.org/10.1108/LHT-06-2015-0063

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 448–456, https://doi.org/10.1145/2020408.2020480

Wang, D., & Barabási, A. L. (2021). *The science of science*. Cambridge University Press.

Wang, L. L., Stanovsky, G., Weihs, L., & Etzioni, O. (2021). Gender trends in computer science authorship. *Communications of the ACM, 64*(3), 78–84. https://doi.org/10.1145/3430803

Wang, N., Qin, Z., Wang, X., & Wang, H. (2021b). Non-clicks mean irrelevant? Propensity ratio scoring as a correction. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining*, ACM, WSDM'21, pp. 481–489, https://doi.org/10.1145/3437963.3441798

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics, 103*(2), 555–564. https://doi.org/10.1007/S11192-015-1547-0

Wang, X., Bendersky, M., Metzler, D., & Najork, M. (2016). Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, SIGIR'16, pp. 115–124, https://doi.org/10.1145/2911451.2911537

Wang, X., Hoi, S. C., Ester, M., Bu, J., & Chen, C. (2017). Learning personalized preference of strong and weak ties for social recommendation. In *26th International World Wide Web Conference, WWW 2017, International World Wide Web Conferences Steering Committee*, pp. 1601–1610, https://doi.org/10.1145/3038912.3052556

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., & Najork, M. (2018). Position bias estimation for unbiased learning to rank in personal search. In ACM. https://doi.org/10.1145/3159652.3159732

Way, S. F., Morgan, A. C., Larremore, D. B., & Clauset, A. (2019). Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences, 116*(22), 10729–10733. https://doi.org/10.1073/PNAS.1817431116

Weatherall, J. O., & O'Connor, C. (2021). Conformity in scientific networks. *Synthese, 198*, 1–22. https://doi.org/10.1007/s11229-019-02520-2

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences, 118*(15), e1912444117. https://doi.org/10.1073/pnas.1912444117

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS ONE, 8*(7), e66212. https://doi.org/10.1371/journal.pone.0066212

Wu, J. (2011). Improving the writing of research papers: Imrad and beyond. *Landscape Ecology, 26*(10), 1345–1349.

Wu, Y., Zhang, L., & Wu, X. (2018). On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 10*, 2536–2544. https://doi.org/10.1145/3219819.3220087

Xu, D., Liu, Y., Zhang, M., Ma, S., & Ru, L. (2012a). Incorporating revisiting behaviors into click models. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM 2012, Vol. 12, pp. 303–311, https://doi.org/10.1145/2124295.2124334

Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y., & Xu, W. (2012). Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems, 54*(1), 564–573. https://doi.org/10.1016/J.DSS.2012.08.003

Yang, B., Lei, Y., Liu, J., & Li, W. (2017). Social collaborative filtering by trust. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(8), 1633–1647. https://doi.org/10.1109/TPAMI.2016.2605085

Yang, C., Wei, B., Wu, J., Zhang, Y., & Zhang, L. (2009). CARES: A ranking-oriented CADAL recommender system. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, ACM Press, pp. 203–211, https://doi.org/10.1145/1555400.1555432

Yang, K., & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In: *ACM International Conference Proceeding Series*, ACM, p 6, https://doi.org/10.1145/3085504.3085526

Yang, L., Wang, C., Cui, Y., Belongie, S., Xuan, Y., & Estrin, D. (2018). Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*, ACM, RecSys 2018, pp. 279–287, https://doi.org/10.1145/3240323.3240355

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *31st Conference on Neural Information Processing Systems, Neural information processing systems foundation*, pp. 2922–2931, arxiv:1705.08804v2

Yin, H., Cui, B., & Huang, Y. (2011). Finding a wise group of experts in social networks. In *Proceedings of the 7th International Conference on Advanced Data Mining and Applications*, Springer, ADMA'11, pp. 381–394, https://doi.org/10.1007/978-3-642-25853-4_29

Zadrozny, B. (2004) Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21th International Conference on Machine Learning* pp. 903–910, https://doi.org/10.1145/1015330.1015425

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A fair top-k ranking algorithm. In *Proceedings of the International Conference on Information and Knowledge Management, CIKM'17*, https://doi.org/10.1145/3132847.3132938

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, PMLR, pp. 325–333

Zeng, Y., & Zuo, S. (2019). The matthew effect in computation contests: High difficulty may lead to 51% dominance? In *The World Wide Web Conference, Association for Computing Machinery*, WWW '19, pp. 2281–2289, https://doi.org/10.1145/3308558.3313593

Zhang, C., Yu, L., Zhang, X., & Chawla, N. V. (2018). Task-guided and semantic-aware ranking for academic author-paper correlation inference. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence*, Vol. 2018, pp. 3641–3647, https://doi.org/10.24963/IJCAI.2018/506

Zhang, G., Bai, B., Liang, J., Bai, K., Chang, S., Yu, M., Zhu, C., & Zhao, T. (2019). Selection bias explorations and debias methods for natural language sentence matching datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL)*, ACL 2019, pp. 4418–4429, https://doi.org/10.18653/v1/p19-1435

Zhang, J., & Bareinboim, E. (2018). Fairness in decision-making-the causal explanation formula. In: Thirty-second AAAI conference on artificial intelligence, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949

Zhao, P., Ma, J., Hua, Z., & Fang, S. (2018). Academic social network-based recommendation approach for knowledge sharing. *ACM SIGMIS Database, 49*(4), 78–91. https://doi.org/10.1145/3290768.3290775

Zheng, Y., Gao, C., Li, X., He, X., Jin, D., & Li, Y. (2020). Disentangling user interest and conformity for recommendation with causal embedding. In *Web Conference 2021 (WWW'21)*, ACM, Ljubljana, Vol. 12

Zhou, Y., Zhu, L., Wu, C., Huang, S., & Wang, Q. (2022). Do the rich grow richer? An empirical analysis of the Matthew effect in an online healthcare community. *Electronic Commerce Research and Application, 52*, 101125. https://doi.org/10.1016/j.elerap.2022.101125

Zhu, Z., He, Y., Zhao, X., Zhang, Y., Wang, J., & Caverlee, J. (2021) Popularity-Opportunity Bias in Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ACM, WSDM 2021, pp. 85–93, https://doi.org/10.1145/3437963.3441820