Information Processing and Management 57 (2020) 102138

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Toward creating a fairer ranking in search engine results

Ruoyuan Gao*[,a], Chirag Shah[b]

[a] Department of Computer Science, Rutgers University United States
[b] Information School, University of Washington United States

ARTICLE INFO

Keywords:
Information retrieval
Search engine bias
Fairness ranking
Relevance
Diversity
Novelty

2019 MSC:
00-01
99-00

ABSTRACT

With the increasing popularity and social influence of search engines in IR, various studies have raised concerns on the presence of bias in search engines and the social responsibilities of IR systems. As an essential component of search engine, ranking is a crucial mechanism in presenting the search results or recommending items in a fair fashion. In this article, we focus on the top-k diversity fairness ranking in terms of statistical parity fairness and disparate impact fairness. The former fairness definition provides a balanced overview of search results where the number of documents from different groups are equal; The latter enables a realistic overview where the proportion of documents from different groups reflect the overall proportion. Using 100 queries and top 100 results per query from Google as the data, we first demonstrate how topical diversity bias is present in the top web search results. Then, with our proposed entropy-based metrics for measuring the degree of bias, we reveal that the top search results are unbalanced and disproportionate to their overall diversity distribution. We explore several fairness ranking strategies to investigate the relationship between fairness, diversity, novelty and relevance. Our experimental results show that using a variant of *fair ε-greedy* strategy, we could bring more fairness and enhance diversity in search results without a cost of relevance. In fact, we can improve the relevance and diversity by introducing the diversity fairness. Additional experiments with TREC datasets containing 50 queries demonstrate the robustness of our proposed strategies and our findings on the impact of fairness. We present a series of correlation analysis on the amount of fairness and diversity, showing that statistical parity fairness highly correlates with diversity while disparate impact fairness does not. This provides clear and tangible implications for future works where one would want to balance fairness, diversity and relevance in search results.

1. Introduction

Search engines are gateways to the Web, and therefore, it is crucial that they are not only effective at doing what they do, but are also fair. As the effects of search engines continue unabated, various studies have raised concerns on the presence of bias in search engines and the social responsibilities of information retrieval (IR) systems (Fortunato, Flammini, Menczer, & Vespignani, 2006; Introna & Nissenbaum, 2000; Snow, 2018; Tavani, 2016). The crucial component of a search engine that governs what and how people access information on the Web is the ranking algorithm. A fair ranking ensures that the appropriate exposure of protected groups diversifies the topical coverage per information need and improves user awareness of different opinions. But the reality is often far from this ideal situation.

* Corresponding author.
E-mail address: ruoyuan.gao@rutgers.edu (R. Gao).

https://doi.org/10.1016/j.ipm.2019.102138
Received 29 April 2019; Received in revised form 28 September 2019; Accepted 2 October 2019
Available online 16 October 2019
0306-4573/ © 2019 Elsevier Ltd. All rights reserved.

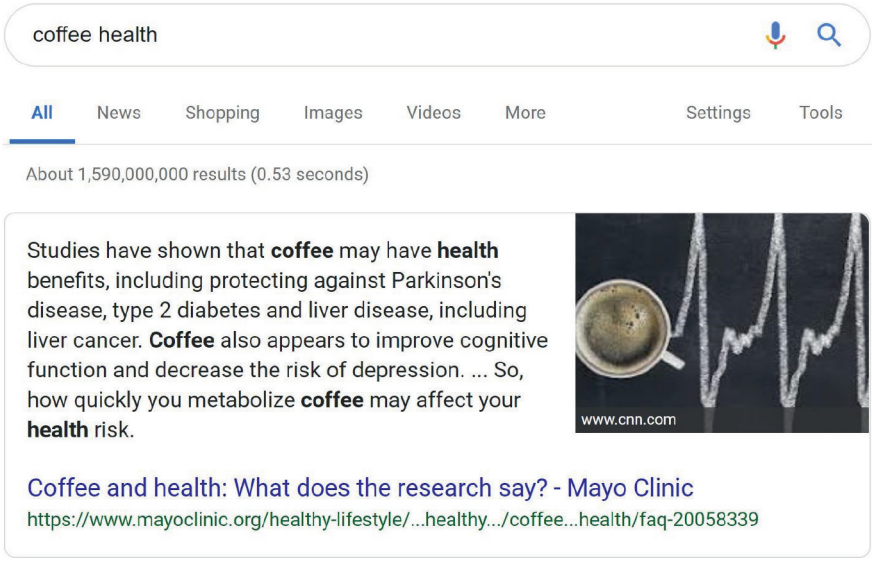
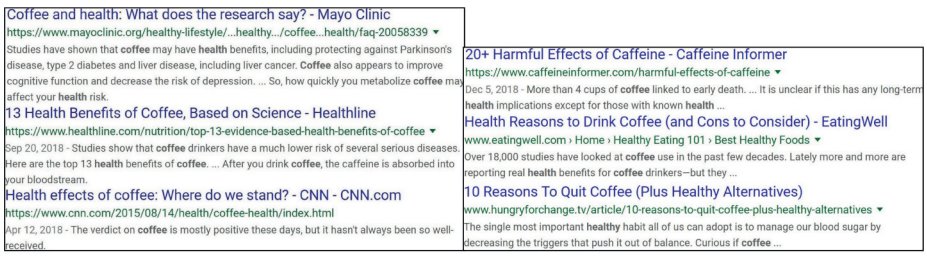


Fig. 1. Featured snippet by Google Web Search for query "coffee health".

Biased presentation of different aspects of the query can lead to problematic cognitive bias. As pointed out in Bar-Ilan, Keenoy, Levene, and Yaari (2009); Haas and Unkel (2017); Kammerer and Gerjets (2012); Keane, O'Brien, and Smyth (2008); Novin and Meyers (2017); Pan et al. (2007); Shokouhi, White, and Yilmaz (2015), the presentation of search engine results affects users' credibility judgment, selection making, and belief and attitude shaping of information. Consider a scenario when a user types in the query "coffee health". Ideally, the search engine should present both the benefits and harms of coffee in a way that the user can have a comprehensive view of coffee's impact on health and weigh the benefits and harms accordingly. Unfortunately, this is not always the case with current search engines. Fig. 1 displays the enhanced snippet by Google web search. This result was particularly processed to be positioned on top of the result page to capture users' attention. Since this snippet is all about benefits of coffee, the user may have an impression that coffee is beneficial to health at the first glance. In fact, all the snippets on the first page are suggesting on coffee's benefits (e.g., Fig. 2(a)). If the user only looks at the first page results, he/she may reach a conclusion that coffee is good for health or the goodness greatly outweighs the badness. However, if the user continues to look at the results on the second page – which happens less often than only looking at the first page, he/she will find that coffee can also be associated with negative impacts on health (Fig. 2(b)).

From the perspective of diversity and novelty, the goodness and badness of coffee are two aspects of the topic, and can be considered as the topical diversity. A diversified set of results should contain as many as possible aspects to present a comprehensive view of the subject. For instance, the enhanced snippet and the first page results are all about health benefits of coffee but ideally they should offer different reasons or perspectives of why coffee is beneficial. Meanwhile, the top results presented should contain as few redundancies as possible so that the user does not have to go through a long list of snippets to find a different variety of information. This is known as improving novelty in search results. Fairness provides a *subjective* moderation on how balanced different subtopics should be presented, hence is essentially different from diversity and novelty. In the example of "coffee health", let us consider the diversity groups of this query's search results to be defined as the benefits and harms of coffee. The top 20 results contain both the goodness and badness of coffee, thus satisfying the diversity requirement. Meanwhile, simply from the snippet offered by Google search, the top 20 results seem to offer different reasons and from different resources, thus each result can be considered to have some novelty compared to the previous results. However, there are 17 results emphasizing the health benefits while only 3 results mentioned the harms. Therefore, according to the statistical parity fairness definition which requires equal exposure of documents from



(a) Top 3 ranked results (b) Results ranked at 18-20

Fig. 2. Search results from Google for query "coffee health".

different groups, the diversity fairness is not satisfied. Previous studies on search result diversification such as Jiang et al. (2018); Xu, Xia, Lan, Guo, and Cheng (2017); Yu et al. (2017, 2018) were able to achieve high diversity and relevance by including documents from different subtopic groups and then rank the selected documents. The objective function can be the relevance score, a trade-off between diversity and relevance, or the evaluation metric. Since such objective functions do not consider the balance between different diversity groups, and the fairness definition is subjective and can be different in different application scenarios, the state-of-the-art diversification methods cannot guarantee any fairness, nor do they provide implications on fairness. To sum up, diversity focuses on the coverage of multiple subtopics regarding a search query and novelty reduces the redundant information in each subsequent result. While both diversity and novelty may help improve diversity fairness by surfacing documents from different diversity groups, neither of which guarantees fairness. The diversity fairness should enjoy the benefits of both diversity and novelty, while satisfying the fairness constraints.

In addition, it is worth noting that a fairness ranking algorithm should ideally not sacrifice relevance to improve fairness, but that may not always be possible. Hence, one must consider the trade-off between relevance and fairness. Existing research on fairness ranking have proposed measures to quantify bias from data sources and ranking systems (Kulshrestha et al., 2017), and evaluates fairness of ranked output (Yang & Stoyanovich, 2017). Several studies have developed algorithms and frameworks to implement fairness ranking, subject to different fairness constraints (Celis, Straszak, & Vishnoi, 2018; Geyik, Ambler, & Kenthapadi, 2019; Singh & Joachims, 2018; Wu, Zhang, & Wu, 2018; Zehlike et al., 2017). Yet, no evaluations have been done on search engine results regarding the relevance performance after re-ranking. In addition, previous studies have used labeled data such as job seeker resumes with gender as groups, or TREC (Text REtrieval Conference) news articles with sources as groups, to study the performance of their ranking algorithm. These works do not reveal topical diversity bias present in search engines, and lack discussions on whether diversity fairness can be achieved on the search engine results without sacrificing relevance. In addition, previous works on trade-offs between diversity, novelty and relevance do no consider the need for fairness, thus lack the implications on the trade-offs between these factors when fairness is introduced into the ranking system.

Motivated by these examples and research gaps, we conducted empirical studies to answer the following research questions (RQs).

- **RQ1:** With a general-purpose topical clustering on Web search results, to what extent do search engines show topical bias regarding documents from different clusters? How do we quantitatively measure the degree of bias?
- **RQ2:** How do various fairness ranking methodologies perform on search engine results regarding relevance? What is the impact of the amount of diversity fairness introduced on relevance, diversity and novelty? What are the trade-offs between diversity, novelty, and relevance with respect to different fairness strategies?
- **RQ3:** How robust are different fairness ranking strategies under various relevance measurement and across distinct datasets? Do the trade-offs between diversity, novelty, and relevance always hold for different datasets? Given a dataset (topics, ranking algorithms, quality of relevance), constraints on fairness and relevance, and relevance metric, which strategy should one choose?
- **RQ4:** Is there a correlation between fairness and diversity? What are the relationship and differences between them?

Our experimental results show that for RQ1 the top search results from Google are indeed biased considering the overall document distributions. We propose entropy-based metrics to measure the degree of bias under different fairness constraints and show the degree of bias varies by queries. For RQ2, we demonstrate that a simple fairness ranking strategy can yield good fairness in top search results without loss of relevance. In fact, incorporating diversity fairness results in better relevance and diversity. For RQ3, we compare the performance of different re-ranking strategies on Google search data and on Text REtrieval Conference (TREC) data. We discover that depending on the constraint of fairness, there is not a one-for-all strategy that suits every situation. The performance of a strategy and the trade-offs between different factors may differ in various scenarios. Therefore, we propose appropriate recommendations on which strategies to choose in each scenario. For RQ4, we show that diversity and relevance are highly correlated with the statistical parity fairness, while this is not the case with the disparate impact fairness, although both fairness benefits the diversity and relevance. In other words, the lower the diversity bias, the higher the relevance, diversity and novelty.

The rest of the article is organized as follows: Section 2 discusses related work. Section 3 defines fairness and the top-k fairness ranking problem, followed by our entropy-based degree of bias metric. Section 4 describes the dataset and bias analysis in Google search results. Section 5 explains the fairness ranking strategies we explore. The experimental results and analysis are presented in Section 6. We conclude our work in Section 7.

2. Related work

There has been a wide range of discussions on the presence of bias in IR and its social impact. Many fairness definitions have since been proposed and emphasized for different application scenarios. Several studies have investigated the fairness ranking problem under different fairness definitions and application scenarios, and motivate our work on the search engine setting. In this section, we present a brief review on previous related works.

2.1. Notions of fairness

Fairness and bias are often considered to be two sides of the same coin. Whether it is fairness or bias, their notions differ in areas of study and focus on social impact. In computer systems, (Friedman & Nissenbaum, 1996) defines bias as “systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others”. In supervised machine learning, fairness

is considered as equal false positives and true positives rates for different groups (Hardt, Price, & Srebro, 2016). In social impact, fairness is often concerned in terms of *individual fairness* (Biega, Gummadi, & Weikum, 2018; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Kamiran, Calders, & Pechenizkiy, 2010; Kusner, Loftus, Russell, & Silva, 2017; Lahoti, Weikum, & Gummadi, 2018) and *group fairness* (Chouldechova, 2017; Kamiran & Calders, 2009; Pedreshi, Ruggieri, & Turini, 2008; Singh & Joachims, 2018). Dwork et al. (2012) discusses the connection between individual fairness and group fairness. Individual fairness treats similar individuals similarly; as such, statistical parity treats the demographics of those receiving positive (or negative) classifications the same as the demographics of the population as a whole. See Narayanan (2018) for a more comprehensive explorations of fairness definitions.

From the aspect of ranking, Singh and Joachims (2018) proposes to view group fairness from a perspective of exposure. In Singh and Joachims (2018), demographic parity is considered as equal total exposure of each group, disparate treatment as equal average exposure of members from each group, and disparate impact as exposure proportional to average utility of a group. These definitions are similar to the concepts of diversified overviews in opinion mining (Demartini & Siersdorfer, 2010). Corresponding to demographic parity, *balanced overview* of search results are defined as having both positive and negative documents. A *neutral overview* shows only objective documents, which can be considered as documents that cover all groups of interest. A *realistic overview* resembles the disparate impact in which the presented documents should reflect the original opinion population. Our view of fairness ranking is motivated by Demartini and Siersdorfer (2010) and Singh and Joachims (2018). Specifically, we investigate fair presentation of search results in terms of *statistical parity* (balanced overview) and *disparate impact* (realistic overview). Instead of classifying documents based on ethnic groups or opinion polarities via sentiment analysis, our study aims to present a perspective of document diversity with general purpose clusters. In fact, considering the possibilities of all kinds of queries and the large size of Web data, it is reasonable to assume the absence of labeled documents or trained classifier given any group definition. Our approach, therefore, provides a more scalable and practical solution to bringing fairness to search results.

2.2. Bias in search engines

Bias in IR may arise from the source data, algorithmic or system bias, and cognitive bias. Algorithms that learn from and mirror real world statistics may unavoidably carry social bias from the original data to the IR system (Barocas & Selbst, 2016). For example, it has been shown that direct usage of click-through data in search engines may introduce bias (Joachims, 2002; Wang, Bendersky, Metzler, & Najork, 2016). Natural language processing and machine learning techniques that learn semantic embedding of human language text encodes human-like semantic biases (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Caliskan, Bryson, & Narayanan, 2017). Kulshrestha et al. (2017) develops a framework that quantifies the amount of bias that arise from the data source and from the ranking system. Because of the bias in data and algorithms, various bias can surface in search engine results. For instance, gender stereotypes have been discovered in image search results for various occupations (Kay, Matuszek, & Munson, 2015), character traits (Otterbacher, Bates, & Clough, 2017), and in resume search engines (Chen, Ma, Hannák, & Wilson, 2018). Weber, Garimella, and Borra (2012) shows bias of political leanings in search queries and results.

From a cognitive perspective, biased search results can lead to unfair distribution of opportunities and resources (Biega et al., 2018; Kearns, Roth, & Wu, 2017). Tavani (2016) discusses the ethical responsibility of search engines. Purcell, Rainie, and Brenner (2012) finds that almost two-thirds of the surveyed adult users believe that search engines are fair and unbiased. Users often rely on the presentation, especially ranking, of search results for credibility judgment, selection making, and belief and attitude shaping of information (e.g., Allam, Schulz, & Nakamoto, 2014; Bar-Ilan et al., 2009; Bråten, Strømsø, & Salmerón, 2011; Haas & Unkel, 2017; Kammerer & Gerjets, 2012; Keane et al., 2008; Van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017; Ludolph, Allam, & Schulz, 2016; Novin & Meyers, 2017; Pan et al., 2007), as well as preference and decision making (Epstein & Robertson, 2015). Novin and Meyers (2017) identifies four cognitive biases that may arise when students use search engines (Priming, Anchoring, Framing, and the Availability Heuristic). Shokouhi et al. (2015) studies the anchoring effect in IR and implies that IR systems should take into consideration of the anchoring effect. Browne Graves (1999) shows that portrayals often reinforce negative stereotypes. Kay et al. (2015) shows that gender bias in image search results exaggerates users' perception about gender stereotypes. Bråten et al. (2011) shows that readers of low topical knowledge may trust less trustworthy sources, and fail to differentiate relevance. Mehrotra et al. (2017) demonstrates the existence of different satisfaction levels across demographics on using search engines.

In this article, we explore lack of diversity in the top results of Google search as a lens to bias in search engines. We discuss how such bias can be present in search results and how it can be addressed without sacrificing much relevance.

2.3. Diversity, novelty and fairness ranking

The utility of a search result or recommendation item not only depends on relevance, but also diversity and novelty. While diversity addresses the ambiguity of search queries, novelty focuses on reducing redundancy in search results (Clarke et al., 2008). Both diversity and novelty are closely connected to fairness. Diversity promotes topical fairness by enriching the overall topic aspect coverage. Novelty boosts fairness ranking by surfacing different documents to avoid positioning bias. However, the optimization goals of diversity and novelty are fundamentally different from fairness ranking. For diversity, the goal is to cover as many aspects of the topic as possible and does not consider the proportion each aspect should be covered. For novelty, the goal is to increase the information gain. To some extend, increasing the coverage of aspects increases the novelty and fairness, but there is no guarantee on the variety of subtopics covered or the distribution of subtopics in the results. Fairness ranking can be viewed as a bartender who

mixes the right amount of diversity and novelty to serve a certain kind of fairness concoction. Fairness aims to distribute the different aspects according to a subjective requirement. To sum up, diversity, novelty, and fairness are different notions.

There have been a wide range of works on diversification which can be classified into explicit and implicit search result diversification (SRD), depending on whether the subtopic groups underlying a query are explicitly known. For example, Jiang et al. (2018) employs recurrent neural networks and max pooling in the learning framework to achieve explicit SRD. Xu et al. (2017) develops a learning to rank framework that directly optimizes the diversity evaluation metrics for implicit SRD. Yu et al. (2017) and Yu et al. (2018) formulate the implicit SRD as a process of selecting and ranking exemplar documents through integer linear programming. The exemplar documents are considered to maximize the relevance while being representative with respect to the non-selected documents. This method ensures the exposure of documents from the majority subtopic group. Hence it will not produce a fair ranking according to statistical parity fairness if the collection of documents are highly biased towards one subtopic group. For implicit SRD, cluster-based methods are often used where similar documents are grouped together and the clusters are used to represent the diversity groups (e.g. Carpineto, D'Amico, & Romano, 2012; Celis, Kapoor, Salehi, & Vishnoi, 2019; Drosou & Pitoura, 2010; Santos, Macdonald, & Ounis, 2015; Wang & Koopman, 2017; Yu et al., 2017; 2018). Previous works such as Carbonell and Goldstein (1998); Clarke et al. (2008) have investigated the trade-offs between diversity, novelty and relevance. Yet no work has investigated the trade-offs when fairness is introduced. At the same time, Mehrotra, McInerney, Bouchard, Lalmas, and Diaz (2018); Singh and Joachims (2018) show that relevance decreases as fairness increases. But no comprehensive analysis is provided whether this is always the case. Our work makes the first practical implications on the trade-offs between diversity, novelty, fairness, and relevance in the fairness ranking setting.

There have been several studies focusing on the fairness ranking problem. Yang and Stoyanovich (2017) proposes fairness measures for ranked outputs based on statistical parity. They compare different group distributions on different prefixes of the ranking list and average the differences in a discounted manner. Zehlike et al. (2017) defines a fair top-k problem to ensure that the proportion of protected groups in the top-k ranking remains above a given threshold. Celis et al. (2018) defines a constrained ranking maximization problem and proposes an efficient approximation algorithm for maximizing the quality metric under a set of constraints of each item. Singh and Joachims (2018) proposes a conceptual and computational framework for fairness ranking which maximizes the utility for the user while satisfying some fairness constraints. Geyik et al. (2019) develops a fairness-aware ranking framework that improves the fairness for individuals without affecting the business metrics. Wu et al. (2018) employs the causal graph to detect and remove both direct and indirect rank bias, and shows that that casual graph approach performs better than statistical parity based approaches in terms of identification and mitigation of rank discrimination. The closest works to our article are by Celis et al. (2019) and Mehrotra et al. (2018). Celis et al. (2019) frames the problem of controlling polarization in personalization as a multi-armed bandit problem and proposes constrained and unconstrained ε-greedy algorithms to optimize for relevance. Our proposed algorithm differs in that we do not consider the loss of relevance as regrets, since the system estimated relevance may be very different from the user judged relevance according to information need. Therefore, instead of optimizing for the minimal relevance regret while subject to fairness constraints, we directly optimize for fairness and system estimated relevance at the same time, and evaluate on the user judged relevance. Mehrotra et al. (2018) addresses the supplier fairness in the two-sided marketplace platform and proposes heuristic based strategies to jointly optimize fairness and relevance. However, their goal is to select a set of items instead of providing a rank of the recommended items, hence is substantially different from our work.

Each of these works develops their ranking approaches under strict mathematical definitions of the fairness they measure, yet they lack analysis of the robustness of their approaches across distinct datasets and various evaluation metrics. Meanwhile, when there is less emphasis on the fairness constraint and more concern on the relevance, there is a lack of an effective way to evaluate such algorithms in terms of bias. In addition, the previous works on fairness ranking assumes all data is labeled with group information, which is not the case for any large-scale search engine. Hence, it is interesting to explore the degree of bias, the relationship between relevance and fairness, and the advantages and disadvantages of various ranking strategies on search engines with a general cluster of documents.

3. Problem definition

In this work, we focus on the top-k fairness ranking problem in search engines. We consider relevance as the utility of the search results. Table 1 is a list of notations used in this article. Formally, for a user query and a set of diversity groups $G = \{G_1, G_2, \cdots\}$, let $D = \{d_1, d_2, \cdots\}$ denote the collection of documents retrieved by the search engine with respect to the user query, $Rs(D)$ be the ranking provided by the search engine. Our goal is to return a ranked set of k documents $D(k) \subseteq D$ with ranking $Rf(D(k))$, s.t. $Rf(D(k))$ is fair regarding G while minimizing the loss of relevance of the original top-k documents.

3.1. Fairness definition

Our notions of fairness are motivated by Demartini and Siersdorfer (2010) and Singh and Joachims (2018), in which two types of fairness are defined. The first type is the *statistical parity fairness* – a ranking is considered to be fair if the exposure of documents from different groups is equal; The second type is the *disparate impact fairness* – a ranking is fair if the exposure of documents from different groups has the same distribution as in the entire collection of documents. Since our goal is to investigate bias in search engines and study the relationship between fairness and relevance, we do not define exposure and fairness in terms of relevance as in Singh and Joachims (2018). Instead, we interpret exposure from the perspective of number of documents from each group, and its impact on the entropy of the exposed documents. Formally, let $g = |G|$ be the number of groups, $g_i = |G_i|$ be the number of documents in group

Table 1
List of key notations .

Notation	Meaning
d_i	a document in the search results
$D = \{d_1, d_2, \cdots\}$	collection of documents/search results
$n = \|D\|$	number of documents in D
G_i	a diversity (subtopic) group
$g_i = \|G_i\|$	number of documents in group G_i
$G = \{G_1, G_2, \cdots\}$	a set of disjoint diversity groups
$g = \|G\|$	number of groups in G
$Rs(D)$	ranking on D provided by the search engine
$Rf(D)$	ranking on D provided by the fair ranking algorithm
k	number of documents in the top results
$D(k) \subseteq D$	a set of k documents
p_i	probability of documents from group G_i in $D(k)$
$G_i(k)$	documents from group G_i in $D(k)$
$f_i = \|G_i(k)\|$	number of documents in $G_i(k)$
$F = \{f_i \| i = 1, 2, \cdots, g\}$	fairness constraint
$E(G)$	entropy with respect to G
r	rank position
DB	degree of bias

$G_i \in G$, $f_i = |G_i(k)|$ be the number of documents that are in the top-k ranked results $D(k)$ and in group G_i. Define fairness constraint $F = \{f_i | i = 1, 2, \cdots, g\}$. Let p_i be the probability of documents from group G_i in $D(k)$, thus $p_i = f_i/k$. The entropy of $D(k)$ with respect to G is

$$E(G) = -\sum_{i}^{g} p_i \log_2 p_i. \tag{1}$$

Such a formulation of entropy provides a way to measure fairness or the degree of bias. The higher the entropy, the better balanced mix of different topical groups and higher the statistical parity fairness. We show in Section 3.2 how entropy can be employed to measure the degree of bias for both types of fairness.

3.1.1. Statistical parity constraint

In many cases, fair exposure simply means equal exposure of different groups. Accordingly, a fair ranking requires the number of documents from each group to be equal. In this work, we define the *statistical parity fairness* as follows. For a set of k documents $D(k)$,

$$f_i = \frac{k}{g}, \forall\, G_i \in G, \tag{2}$$

thus $p_i = 1/g$, and the entropy

$$E(G) = -\sum_{i}^{g} \frac{1}{g} \log_2 \frac{1}{g} = \log_2 g. \tag{3}$$

A top-k ranking of documents $D(k)$ is fair under the statistical parity definition if it satisfies the constraint $F = \{f_i | i = 1, 2, \cdots, g\}$ where f_i is defined in Eq. 2. Note that entropy is the maximum when the distribution is uniform, thus statistical parity corresponds to achieving the maximum entropy of the selected documents.

3.1.2. Disparate impact constraint

As described in Demartini and Siersdorfer (2010) and Kay et al. (2015), a fair presentation of search results should reflect truthfully about the real world distribution of different groups. This is particularly desirable when a user wants to perceive the majority opinion on an unfamiliar subject. For example, a user would like to know whether coffee is good for one's health. Not only does he/she desire a comprehensive summary of both the goodness and badness about coffee, but he/she wants to know what do most people say about coffee. In this case, if 70% of the discussions are about the benefits of coffee, then 70% of the top search results should present opinions on the goodness of coffee. The disparate impact fairness is defined as the proportion of documents from each group in the top-k results being the same as in the entire document collection. Formally, we say a top-k ranking of documents $D(k)$ is fair under the *disparate impact fairness* if it satisfies the constraint $F = \{f_i | i = 1, 2, \cdots, g\}$ where f_i is defined in Eq. 4,

$$f_i = k \cdot \frac{g_i}{n}, \forall\, G_i \in G, \tag{4}$$

thus $p_i = g_i/g$, and the entropy

$$E(G) = -\sum_{i}^{g} \frac{g_i}{n} \log_2 \frac{g_i}{n}. \tag{5}$$

A fair ranking with respect to disparate impact should achieve the entropy derived above. However, note that achieving this entropy does not guarantee fairness since it cannot infer the proportion of documents from a particular group.

3.2. Degree of bias

We show that entropy based metrics can be used to effectively measure the degree of bias in ranking results. For example, for statistical parity constraint, we can compare how balanced the items are from each group using the entropy of the collection of documents. Let E^* be the "ideal" entropy of k documents when the statistical parity fairness constraint is satisfied, E_a be the entropy of the top-k ranking results provided by a ranking algorithm. The degree of bias under the statistical parity fairness constraint can be defined as the relative distance between the entropy E_a and E^* as

$$DB = \frac{|E^* - E_a|}{E^*}. \tag{6}$$

For disparate impact constraint, we are comparing how the distribution in the top-k results differs from the true distribution in the entire set of search results. Relative entropy (also known as Kullback-Leibler divergence) can be used to measure the degree of bias under this fairness constraint.

$$DB = D_{\mathrm{KL}}(P||Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right). \tag{7}$$

Here, P is the true distribution of groups and Q is the distribution in the top-k search results. Note that the distribution P cannot contain zeros by definition – there must be at least one item in each group since an empty group means the group does not exist. While in the top-k results, the distribution Q may contain zeros for some groups since it is only a partial set of the entire collection. In this case, we cannot use Kullback-Leibler divergence directly since Q containing zeros where P does not are not well defined. We applied Laplace smoothing to the distribution of Q to avoid this issue.

The above are just a few examples of using entropy as a measure of degree of bias. Other entropy based measures such as mutual information gain can also be applied to compare the differences between the distribution observed and the target fair distribution, which provides a measure of the degree of bias. In other words, the degree of bias can be quantified by measuring how far the entropy of a ranking result is from the ideal entropy defined by the fairness constraint. For a set of queries, we can measure the overall degree of bias at rank r by taking the average degree of bias for all queries. As a result, we have a uniform metric to compare the effectiveness of different fairness ranking strategies.

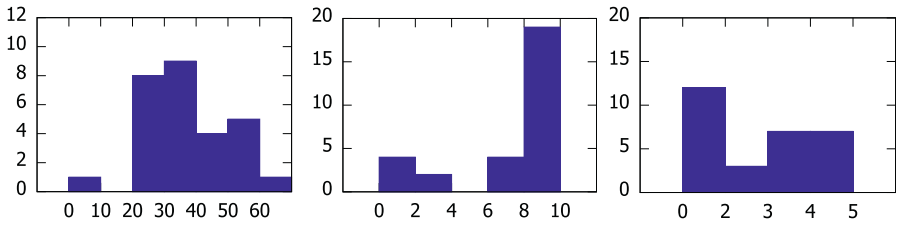
4. Bias in search engines

In this section, we describe our dataset and analysis pertaining to identifying topical diversity bias in search engine results. To begin our investigations, we used Google search as a case study to understand bias and fairness ranking in search engines. We took 30 queries from Google Trends[1] on March 3, 2019 for analyzing the presence of bias in search engine results. We then took another 70 queries from Google Trends spanning the week of June 23, through June 29, 2019 with 10 queries per day. This resulted in a diverse collection of 100 popular queries from different days and months. For examining different diversity fairness strategies and investigating the relationship between relevance, diversity, novelty and fairness, we used the entire 100 queries (see Section 6). We crawled the first 100 results per query, eliminating advertisements, widgets, etc. We excluded widgets to avoid duplicate documents since the widget content was from the search result documents. For each result, we collected the *snippet* (title, url, excerpt from Webpage) along with the actual Webpage converted to plain text. For some Webpages we were not able to crawl the content due to privacy policy, thus we excluded such results from our collection. Note that we still kept the original ranking value of each document even if some documents were excluded from the collection. The entire collection after pre-processing (stop word removal, stemming, etc.) resulted in 7,410 documents totalling 7.65M words, with an average of 74 documents per query, and an average vocabulary size of 8617 per query.

4.1. Topical group assignment

To examine bias in web search results, as opposed to image search results (Kay et al., 2015), we first clustered all documents to assign group labels. We used the collection of 30 queries from March 3, 2019 as a case study to analyze the presence and degree of bias in search results. To automatically assign each document to a topical group, we used k-means clustering algorithm with tf-idf representations of documents. Clustering is a commonly used technique in implicit SRD where subtopic groups are not explicitly defined. Similar to Celis et al. (2019), we obtained the diversity groups using k-means for the purpose of investigating diversity fairness. Note that there are various cluster-based approaches to represent the diversity groups. For diversity fairness, no matter

[1] https://trends.google.com/trends



(a) $2 \leq k$ (b) $2 \leq k \leq 10$ (c) $2 \leq k \leq 5$

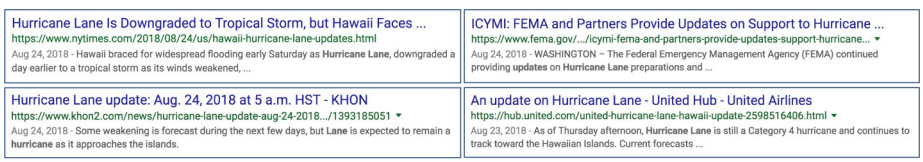
Fig. 3. Histogram of the best number of clusters decided by silhouette score. k is the maximum number of clusters. x-axis is the number of clusters.

which diversification approach is used, a fair ranking always satisfies the diversity requirement with respect to the obtained diversity groups. To best separate the documents, we used silhouette score to decide the number of clusters. The cluster number with the highest silhouette score was considered the best. Fig. 3 (a) is the histogram of the best cluster numbers for all the queries. We can see that the best cluster numbers vary greatly from query to query, with an average of 37.48. For the majority of the queries, the best cluster numbers lies between 30 to 70. We observe that the silhouette score tend to favor larger number of clusters for most of the queries. This is due to the fact that the search results are similar to each other when represented by tf-idf features. As a result, the distance between documents are so small that silhouette score increases as the number of clusters. This finding is consistent with previous works on clustering documents according to semantic meanings (e.g., Celis et al., 2019; Wang & Koopman, 2017). Consider the problem of selecting top-k among 100 search results returned by a search engine, where k is relatively small (e.g., $k = 10$ or $k = 20$), using silhouette score to decide cluster number with no constraint on the maximum becomes impractical for topical group assignment. To construct a fair top-k search results list, it is reasonable to assume that the number of groups or clusters is a parameter capped at k so that a fair solution is possible to exist. If the number of clusters is greater than k, then there is no way to construct a list of k items with at least one item from each cluster, thus the resulting list must be unfair for documents in the omitted cluster. On the other hand, if the number of groups is given, then a fair top-k list must have k no less than the number of groups. In this article, we set $k = 10$ for the purpose of simplicity in illustration and explanation. We experimented with the maximum number of clusters to be 10. Fig. 3 (b) shows the histogram of the best number of clusters capped at 10. We can see that most queries have the highest silhouette score with 9 and 10 clusters, approximating the maximum number of clusters allowed. The average number of clusters determined by silhouette score is 7.90. For number of clusters capped at 5, the average is 3.31 (see Fig. 3 (c) for the histogram).

Subtopic retrieval and topical group assignment algorithm is inherently different from the problem of fairness ranking. Fairness ranking aims to balance the number of documents from different groups where the group definition and membership assignment can be defined by any means according to the use case needs. Designing a perfect algorithm for topical group assignment is beyond the scope of this article. Therefore, in this article, we consider the groups determined by the k-means clustering algorithm to be good enough for topical group assignment. For the purpose of investigating the bias present in the top-10 search results, we limit the cluster number to be no greater than 5, and then decide the number of clusters for each query to be the one with the highest silhouette score.

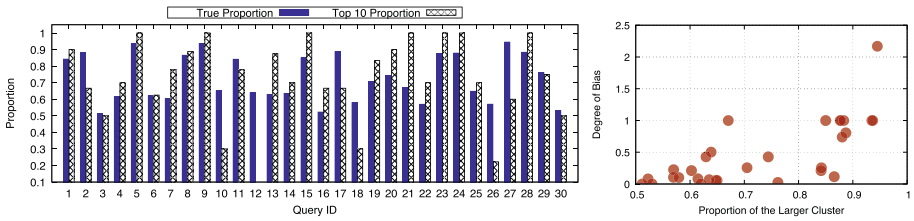
4.2. Analyzing bias

We started with the simplest case where cluster number is 2 to reveal bias in search engine results. Fig. 4 is an example of the clustering result for query “hurricane lane update”. As illustrated in the figure, cluster A tended to contain documents about severeness and forecast, while cluster B tended to contain documents that are more about government and travel agency announcements. After clustering, we compared the proportion of documents from the larger cluster in the top-10 results against the proportion in the entire document set. As shown in Fig. 5(a), for a majority of the queries, the proportion of cluster A in the top-10 results (orange bars) was not close to 0.5, meaning the two clusters did not have equal exposure in top-10. Thus, the top-10 results were biased under the statistical parity constraint. Meanwhile, for a majority of the queries, the proportion of cluster A in the top-10 results did not agree with the cluster distribution in the overall 100 search results (true proportion). In other words, the topic distribution in the top-10 results could not represent the entire 100 search results. Thus, the top-10 results were also biased under the disparate impact constraint. To see how biased the search results were under the disparate impact constraint, we used the definition in Eq. 7 to compute the degree of bias (shown in Fig. 5(b)).



(a) cluster A (b) cluster B

Fig. 4. Example of documents in clusters for query “hurricane lane update”.



(a) Proportion of cluster A in top-10 results vs overall results.

(b) Degree of bias in top-10 search results for each query.

Fig. 5. Bias in top-10 results from Google search.

To see how bias can be present across the overall search results, we investigated the top-k results for different values of k. Figs. 6(a) and 7(a) show the proportion of the larger cluster (denoted as cluster A) in the top-20 and top-30 results compared with that in the 100 search results for each query. We can see that the proportions of cluster A were very close for top-20 and top-30 results, and both of them presented bias under the statistical parity and disparate impact fairness constraints. Compared with the top-10 results, the top-20 and top-30 proportions became closer to the true proportions for some queries, while for others this was not the case. Again, we can compute the degree of bias for both sets of results. Figs. 6(b) and 7(b) reveal that there existed high degree of bias for many queries in both sets of results.

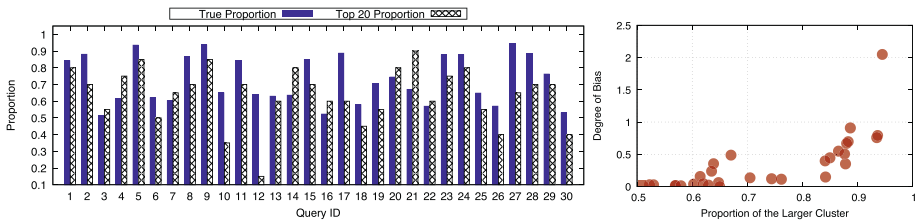
Next, we limited the number of clusters to be 5 and used silhouette score to decide the best number of clusters. We compared the degree of bias for different definitions of fairness. Fig. 8 shows the degree of bias for two different queries and the average degree of bias for all queries. We can see that the degree of bias under disparate impact constraint decreased as the rank lowered. This was because as the rank decreases, the set of top-k documents got closer to the entire collection of documents (100 results). As a result, the distribution of documents in top-k results got closer and eventually became the same as in the entire collection. Fig. 8(b) shows that the degree of bias for two fairness constraints were very different for the query "hurricane lane update", especially at higher ranks like the top-10 results. The statistical parity bias did not decrease as the rank decreased, indicating the documents from different clusters were not balanced at any rank. On the contrary, Fig. 8(a) for query "madden shooting" shows that the trends of the two bias were the similar. As the rank decreased, both degree of bias decreased. This was because in the entire collection of documents for this query, documents were relatively balanced. Hence the true distribution were close to fair under the statistical parity constraint. The average bias over all queries (Fig. 8(c)) shows that there existed high degree of bias in the search results under both fairness constraints, and in general the documents from different clusters were imbalanced.

Note that the bias revealed depends on the clustering algorithm or the group assignment of documents, so it is possible that, for some group definitions and assignments, the top-k results from Google search show less bias for some fairness definitions. But exploring such scenarios are out of the scope of this article. Here we aim to demonstrate that given a group assignment for topical diversity, a search engine may present certain degrees of bias. We are interested in how to address such bias and understanding the impact of bias on relevance.

5. Fairness ranking strategies

To achieve fairness in the top-k results, we explored several re-ranking strategies. The basic idea was to pick documents according to the fairness constraint $F = \{f_i | i = 1, 2, \cdots, g\}$ in Sections 3.1 and 3.2 from each group, and compose the new top-k ranking results. After settling the number of documents to pick from each group, we then worked on how to decide which documents to pick, and how to rank the picked documents. It is worth mentioning that picking from different groups is already enforcing some diversity. For easier illustration, consider $k = 10$, $g = 2$. We use the example in Fig. 9 to demonstrate how each strategy works. In Fig. 9, the numbers correspond to the original ranking positions returned by the search engine. Cells are colored according to the cluster membership. With $g = 2$, the ideal entropy (3) for statistical parity constraint becomes

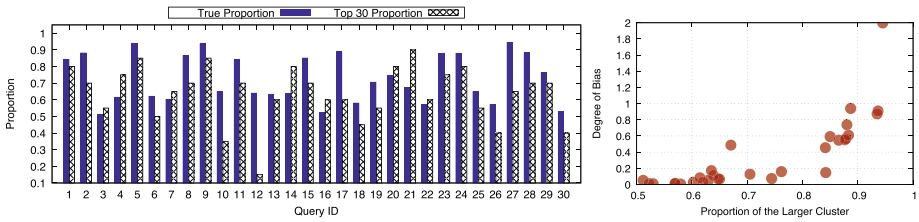
$$E(G) = \log_2 2 = 1, \tag{8}$$



(a) Proportion of cluster A in top-20 results vs overall results.

(b) Degree of bias in top-20 search results for each query.

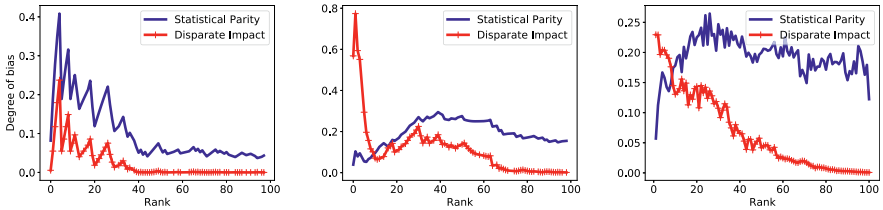
Fig. 6. Bias in top-20 results from Google search.



(a) Proportion of cluster A in top-30 results vs overall results.

(b) Degree of bias in top-30 search results for each query.

Fig. 7. Bias in top-30 results from Google search.



(a) Query "madden shooting" (b) Query "hurricane lane update" (c) Average of 30 queries

Fig. 8. Degree of bias for top-k results under different fairness constraints.

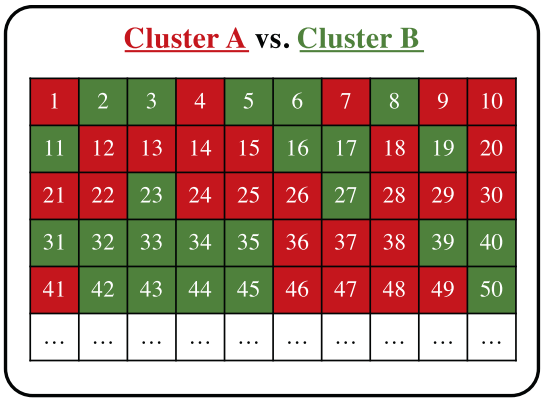


Fig. 9. Example of clusters for query "hurricane lane update". The numbers are document ranks in the original ranking $Rs(D)$. Documents in cluster A are in red cells, and documents in Cluster B are in green cells. E.g., the number 23 in green cell means the 23rd document returned in the search results is in Cluster B. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and for disparate impact constraint, let $p = g_i/n$, the ideal entropy (5) becomes

$$E(G) = -(p\log_2 p + (1-p)\log_2(1-p)). \tag{9}$$

5.1. Baseline

We consider three baseline strategies. Each strategy first picks the number of documents from each cluster strictly subject to the fairness constraints. Then within each cluster, each strategy either picks from the top or randomly.

Top-top: a naive approach is to simply pick the top ranked results from each group, then rank the documents in the same order of their original rank. Consider the example in Fig. 9 under statistical parity constraint; this strategy picks the top 5 from cluster A: [1,4,7,9,10] and top 5 from cluster B: [2,3,5,6,8]. The 10 documents are then ordered to compose the new ranking list [1,2,3,4,5,6,7,8,9,10].

Page-wise: While picking the top has the advantage of retaining good relevance, the disadvantage is obvious: the top documents close to each other are likely to be similar. As a result, the returned rank list has low diversity and novelty. To counteract this issue, we can try to increase the distance between the documents and bring up lower ranked documents. Given that Google search displays results in 10 pages, with 10 results per page, we can implement this idea by picking just one document from each cluster on each page, until the number of documents from each cluster satisfies the fairness constraint F. If a page does not contain any documents from the desired cluster, then we will pick instead from the previous page. When picking from a page, we pick the top document to

increase the chance of high relevance. In the example in Fig. 9, under statistical parity constraint, this strategy will pick [1,12,21,36,41] from cluster A and [2,11,16,31,42] from cluster B. We can order the selected documents the same way as in *top-top*, yielding [1,2,11,12,16,21,31,36,41]. Although we retrieved consecutive documents such as (1,2) and (11,12), the diversity is guaranteed through different group assignment in each pair.

Fair-random: the page-wise strategy has greatly increased the possibility of diversity and novelty of the re-ranked results. Yet, it is still not diverse enough to cover documents from even lower ranks. This can be addressed by introducing randomization into the re-ranking algorithm. Instead of always picking from the top, we pick uniformly at random from each cluster, then order the picked documents according to their original rank.

5.2. ε-greedy

As randomization does not consider the original rank, the availability of original ranking is wasted, and the algorithm is at risk of yielding low relevance results. To enjoy the diversity and novelty introduced by randomization, while benefiting from the clue of original ranking, we developed an ε-greedy fairness ranking algorithm. This algorithm is inspired by the exploitation and exploration idea from ε-greedy for the multi-armed bandit problem (Berry & Fristedt, 1985; Sutton & Barto, 2018). The multi-armed bandit problem is a stochastic scheduling problem where decisions must be made between exploitation and exploration to achieve maximum reward with limited resources. Imagine a player with a limited budget in front of $N \geq 2$ slot machines. Each machine has an unknown reward distribution. The player must decide which levers to pull, for how many times and in what order, to maximize the total rewards earned. The ε-greedy algorithm is a simple strategy that is commonly used in multi-armed bandit problems: with probability $1 - \varepsilon$, select the lever that historically yields the highest reward, i.e., exploitation; with probability ε, select a random lever, i.e., exploration. We took on the similar idea to derive the *naive ε-greedy* and *fair ε-greedy* for the ranking problem, where the rewards became relevance and agreement with fairness constraint.

Naive *ε-greedy*: without considering group assignments, the naive ε-greedy algorithm simply explores the entire document collection with probability ε; with probability $1 - \varepsilon$, it exploits the most rewarding document at the moment, i.e., the top ranked document in candidates. This algorithm aims to achieve a good relevance by taking advantage of the original ranking while introducing diversity. When $\varepsilon = 0$, it always picks the top k documents, yielding the same results as original top-k ranking. When $\varepsilon = 1$, it purely selects documents at random. This algorithm does not address fairness.

Fair *ε-greedy*: A fair ε-greedy algorithm considers group assignments of the documents (see Algorithm 1). The highest rewarding cluster at each moment is the one that pushes the results to satisfy fairness constraint in the current state. Due to the likelihood of exploitation, this algorithm guarantees fairness constraint with constant probability. The higher ε the more randomization and hence diversity, and the lower ε the more chances of being fair. When $\varepsilon = 0$, this algorithm behaves the same as *top-top*.

Despite the choices of ε, none of the above ε-greedy algorithms reduces to the *fair-random* strategy as *fair-random* always guarantees fairness but ignores original ranking.

6. Experiments

6.1. Evaluation metric

For the Google search dataset, we manually judged all documents' relevance with respect to each of the 100 queries. For each

Input: $k \geq 2$, $\varepsilon \in [0, 1]$, rank on documents $Rs(D)$, clusters $G = \{G_1, G_2\}$, fairness constraint $F = \{f_1, f_2\}$
Output: $Rf(D(k))$

Initialize $i = 1$, $D(k) = \{$ the top result in $Rs(D)$ $\}$
while $i < k$ **do**
  with probability ε:
  C = randomly pick from $\{G_1, G_2\}$
  with probability $1 - \varepsilon$:
  if $|\{d_j | d_j \in D(k) \cap G_1\}| < f_1 \cdot \frac{i}{k}$ **then**
    $C = G_1$
  else
    $C = G_2$
  end if
  add the top result in $Rs(C)$ to $D(k)$
  $i = i + 1$
end while
$Rf(D(k))$ = Sort $D(k)$ according to $Rs(D(k))$

Algorithm 1. Fair ε-greedy Ranking.

query, the annotator was presented with all the retrieved search results and asked to decide whether each result was relevant to the query. The order of the search results were shuffled so that the judgment was not influenced by the original ranking. The annotators were informed to judge each result simply based on the relevance in a normal search engine use scenario, i.e., they were not informed to consider factors like diversity or fairness. The relevance score was binary. The Cohen's Kappa coefficient for the inter-annotator agreement was 0.89.

In this work, we aim to explore the relationship between relevance, diversity and fairness. Previous works such as Celis et al. (2018); Geyik et al. (2019); Singh and Joachims (2018); Zehlike et al. (2017) explored ranking algorithms for achieving fairness while optimizing the relevance. However, such algorithms were only able to show the difference between the relevance achieved when certain fairness constraints were met compared to the optimal relevance score. They could not capture how diversity and relevance would change when different amount of fairness were introduced into the system. In addition, such algorithms required the availability of gold standard relevance judgment so that they could optimize for the relevance score computed with the judgment. However, for search engine results, the relevance judgment varies by users and the search intents. Hence optimizing relevance based on the system relevance judgment may not lead to the optimal relevance judged by users – this is also why diversity becomes a crucial factor when ranking results. In other words, the previous fairness ranking algorithms were not suitable as baselines for our purpose. Consequently, the original Google search ranking results were evaluated based on human judgment as the baseline. We evaluated the re-ranked top-10 list for the 100 queries using intent-aware metrics NRBP (Clarke, Kolla, & Vechtomova, 2009), ERR-IA (Chapelle, Metlzer, Zhang, & Grinspan, 2009), α-nDCG (Clarke et al., 2008), NRBP (Clarke et al., 2009), and our proposed Degree of Bias (DB) metric. α-nDCG, ERR-IA, and NRBP are commonly used metric for evaluating relevance that also account for diversity and novelty. The relevance scores were computed with the official TREC 2014 Web Track evaluation tool, *ndeval*[2]. With different amount of fairness introduced, measured by DB, we can analyze how diversity and novelty encoded relevance changes. The results for statistical parity and disparate impact fairness are reported in Table 2. For *fair-random* and ε-greedy strategies, we repeated experiments for 1000 runs and reported the average score on each metric. We also included the results of *naive ε-greedy* as a non-fairness strategy to study whether simply introducing randomness can improve the relevance and diversity, and the effect of incorporating fairness on relevance and diversity.

6.2. Fairness on diversity and relevance

First, let us consider the effect of randomization on relevance and diversity. Comparing the results produced by non-fair and fair algorithms, we can see that non-fair algorithms did not improve the diversity and novelty based relevance scores. In fact, in the non-fair case, as we introduced more randomness in hope that novelty would be increased by bringing search results from the lower ranked positions, the scores even dropped with the pure random algorithm *1-greedy*.

Second, all the fairness-aware ranking algorithms performed much better than the non-fair algorithms. This means that incorporating fairness helps improve diversity and relevance. Additionally, according to the scores of DB, the lower the degree of bias, the higher the relevance scores. The Google search's results were more biased regarding different subtopics, as we have seen in Section 4.2. This shows the important positive impact of fairness on diversity and relevance.

As we can see in Table 2, the *fair-random* and *naive 1-greedy* performed the worst. This was expected since they did not take advantage of the original ranking where higher ranked documents were more likely to be relevant. By always picking a document randomly, the relevance got discounted with lower ranked documents. As ε decreased, we exploited more and explored less. By reducing the amount of randomization due to exploration, we had a higher chance to select documents in the top rank. At $\varepsilon = 0$, the *naive ε-greedy* always picks the top ranked documents. So the *naive 0-greedy* algorithm achieved the same relevance scores as the original Google's top-10 results. However, since it did not account for fairness, the diversity scores were not as high as in the case of fairness ranking. For statistical parity fairness, we can see that the *top-top* (or *fair 0-greedy*), *fair 0.01-greedy*, and *fair 0.1-greedy* achieved the highest scores. This was because these strategies took advantage of the original ranking's high relevance while ensuring a good amount diversity fairness. When $\varepsilon = 1$, the *fair ε-greedy* always picks the top ranked document from a random cluster. Since each cluster is selected with the same probability, in expectation, each cluster will be selected the same number of times. Formally, let A, B be two clusters that partition the entire collection of documents, $x_i \in \{0, 1\}$ be the cluster assignment of document x_i. $x_i = 1$ if $x_i \in A$, and $x_i = 0$ if $x_i \in B$. Let X denote the random variable of cluster assignment in the top-k re-ranked results $D(k)$, $Y = |\{x_i | x_i \in A, x_i \in D(k)\}|$ be the number of cluster A documents in top-k results. For *fair 1-greedy* algorithm, $p(X = 1) = p(X = 0) = 1/2$, the expectation of random variable X

$$E[X] = p(x = 1)\cdot 1 + p(x = 0)\cdot 0 = \frac{1}{2}, \tag{10}$$

$$E[Y] = k\cdot E[X] = \frac{k}{2}. \tag{11}$$

The expected number of cluster B documents in top-k results is $k - E[Y] = k/2$. So, the result set should contain equal number of documents from each cluster in expectation. Consequently, the *naive 1-greedy* should produce similar results as *top-top* for statistical parity. However, because of the randomness associated with picking a cluster, perfect fairness was not guaranteed as in the case of *top-top*. Therefore, the diversity score was not as good as that of the *top-top*.

[2] https://github.com/trec-web/trec-web-2014

Table 2
Relevance scores on Google Search under different fairness constraints – (↑) indicates that the higher the better. (-) means not statistically significant compared with Google search's original ranking. All other relevance scores are statistically significant with $p < .001$.

Algorithm	DB↓	ERR-IA@10↑	α-nDCG@10↑	NRBP↑
Google search	0.2991	0.8523	0.8467	0.8581
non-fair				
0.0-greedy	0.2991	0.8523	0.8467	0.8581
0.01-greedy	0.2983	0.8520 (-)	0.8463 (-)	0.8578(-)
0.1-greedy	0.2875	0.8518 (-)	0.8458(-)	0.8574(-)
0.5-greedy	0.2597	0.8569	0.8498	0.8625
1.0-greedy	0.2516	0.8439	0.8379	0.8486
fair (statistical parity)				
top-top (0-greedy)	0.0003	**0.9410**	**0.9477**	**0.9411**
page-wise	0.0006	0.9317	0.9364	0.9322
fair-random	0.0005	0.9133	0.9189	0.9121
0.01-greedy	0.0004	**0.9411**	**0.9479**	**0.9411**
0.1-greedy	0.0018	**0.9411**	**0.9479**	**0.9411**
0.5-greedy	0.0130	0.9397	0.9457	0.9404
1.0-greedy	0.0758	0.9294	0.9314	0.9330
fair (disparate impact)				
top-top (0-greedy)	0.0021	0.9030	0.8962	0.9098
page-wise	0.0018	0.8853	0.8932	0.9075
fair-random	0.0021	0.8737	0.8666	0.8791
0.01-greedy	0.0046	0.9179	0.9148	0.9244
0.1-greedy	0.0076	0.9201	0.9176	0.9264
0.5-greedy	0.0436	**0.9297**	**0.9306**	**0.9340**
1.0-greedy	0.1530	**0.9291**	**0.9310**	0.9327

Third, let us compare relevance scores under different fairness constraints. The results of statistical parity were overall better than the disparate impact. This implies that producing a *balanced* ranking list regarding different subtopics yields better relevance at the subtopic level, hence better diversity. Consequently, it was not surprising to see that the *0.5-greedy* and *1-greedy* performed better comparing to the *top-top* in the case of disparate impact, while being worse than the *top-top* in the case of statistical parity – because the higher ε brought higher possibility of the results being balanced.

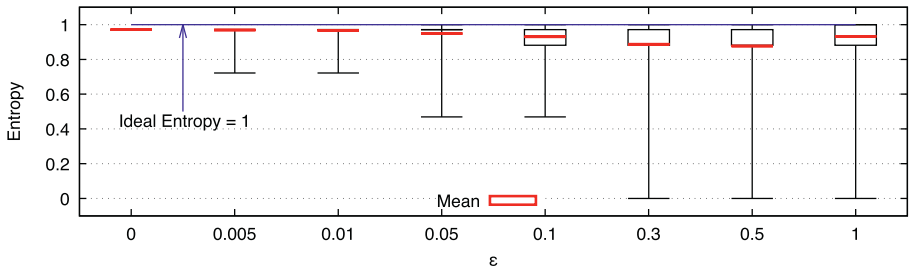
Forth, under the same fairness constraint, we can compare the strength and weakness of different re-ranking strategies. From the perspective of relevance metrics, the *page-wise* did slightly better than *fair-random* but was still worse than other strategies. This can be explained by the fact that, on the one hand, *page-wise* considered original ranking so that it was able to guarantee more relevant documents than *fair-random* in expectation; on the other hand, the *page-wise* strategy enforced more lower ranked documents than ε-greedy in expectation.

Finally, it seemed that all non-fair strategies produced similar results to the original Google search's results, while the fairness-aware strategies produced much better results. To understand whether the impact of fairness was statistically significant regarding the relevance scores, we performed the two-tailed *t*-test to compare each re-ranking strategy with the baseline (Google search's results). We then applied Tukey's HSD test (Jayasinghe, Webber, Sanderson, Dharmasena, & Culpepper, 2015) as the post-hoc correction. As shown in Table 2, all the fairness ranking algorithms performed statistically significantly different from the baseline ($p < .001$), while the non-fairness algorithms were not statistically significantly different ($p > .05$). For statistical parity fairness, the *0.01-greedy, 0.1-greedy*, and the *top-top* (*0-greedy*) algorithms did not yield statistically significant difference ($p > .05$) from each other. For disparate impact fairness, the *0.5-greedy* and *1.0-greedy* did not have statistically significant difference from each other for ERR-IA and α-nDCG ($p > .05$). All other fairness results across different fairness ranking algorithms were statistically significantly different from each other ($p > .05$).

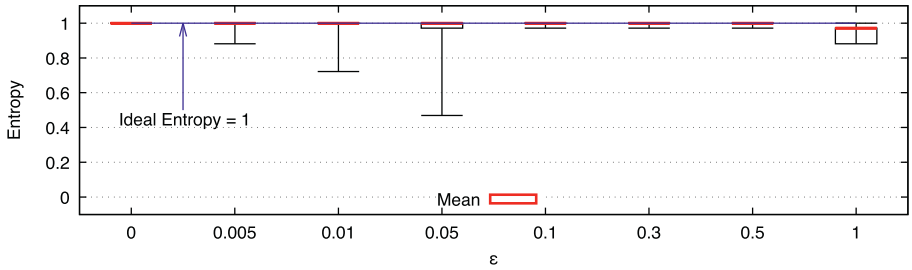
6.3. ε-Greedy vs fairness

To study the effect of exploration in ε-greedy algorithm's performance regarding fairness, we conducted simulated experiments showing the differences between algorithm entropy and the ideal entropy. For statistical parity, we randomly partitioned 100 integers into two clusters of equal size. We then ran the *naive* ε-*greedy* and *fair* ε-*greedy* respectively for 10,000 iterations. The observed entropy are illustrated in Fig. 10 using quartiles with means. From Fig. 10(a), we see that *naive* ε-*greedy* does not produce the same entropy as the ideal entropy in Eq. (8). As ε increases, the entropy deviates even farther. In Fig. 10(b), we see that the *fair* ε-*greedy* consistently produces entropy that is close to the ideal entropy. As ε increases, the entropy reliably remains close to ideal entropy, and the deviation is increasing at an negligible rate.

For disparate impact, we partitioned 100 integers into two clusters of different sizes and conducted experiments for varying ε with varying p. We observed the same effectiveness of achieving fairness as under statistical parity constraint. In Fig. 11(a), taking $p = .3$ as an example, we see that the algorithm entropy stayed close to ideal entropy. With an increasing ε, we observed a slight increase of deviation. This was because the higher the ε the more the exploration, and thus less guarantee on fairness. Note that when $\varepsilon = 0$, the *fair* ε-*greedy* always selected documents that satisfied the fairness constraint, hence produced the same entropy as *fair-random*

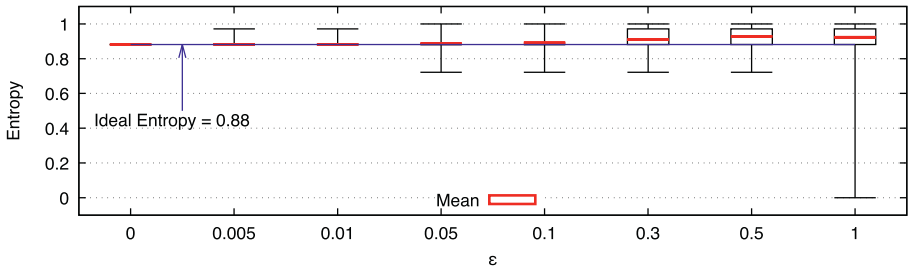


(a) Naive ε-greedy

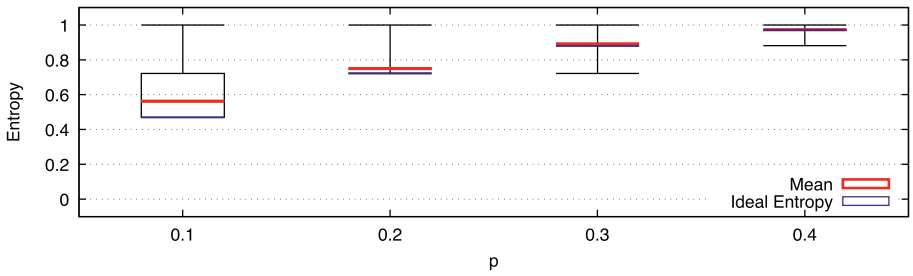


(b) Fair ε-greedy

Fig. 10. Entropy of ε-greedy under statistical parity.



(a) $p = 0.3$



(b) $\varepsilon = 0.1$

Fig. 11. Entropy of fair ε-greedy under disparate impact constraint. The entropy deviates from ideal entropy with varying group proportion p and is affected by ε.

strategy. When $\varepsilon = 1$, the *fair ε-greedy* always picked clusters at random. Therefore, the expected entropy should be the same for both fairness constraints.

To examine the relationship between fairness and cluster distribution, we took a fixed $\varepsilon = 0.1$ as an example, and varied the proportion p of integers assigned to a specified cluster. We observe that the ideal entropy increases as p increases (blue bars in Fig. 11(b)). The *fair ε-greedy* reflects the same trend as ideal. As p increases, the difference in size of the two clusters are decreased, and the algorithm entropy gets closer to the ideal entropy. When $p = .5$, the two clusters have exactly the same size. Therefore, statistical parity can be viewed as a special case of disparate impact when two groups are of equal size. From this we can infer that *fair ε-greedy* works the best to ensure fairness when sizes of different groups are close.

Since ε-greedy algorithms do not guarantee fairness at all times for $\varepsilon > 0$, we wanted to measure the degree of bias when fairness

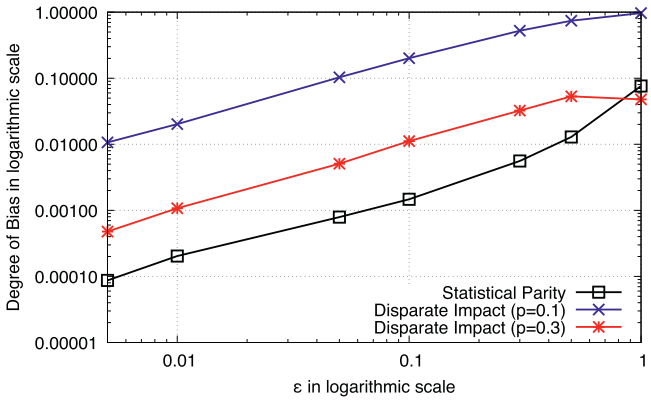


Fig. 12. Degree of bias of fair ε-greedy algorithm.

constraints were not met. Theoretically, the larger ε for exploration, the higher degree of bias. To compare the affect of ε on different fairness constraints, we performed 1000 runs for each choice of ε and fairness. The average degree of bias results are depicted in Fig. 12. Again, we used $p = .3$ as an example to illustrate the effectiveness of *fair ε-greedy* under disparate impact constraint. Fig. 12 shows that the degree of bias grows with ε, and the closer p is to 0.5 the lower degree of bias. This is consistent with our observation in Fig. 11(b).

6.4. Robustness

To evaluate the robustness of each fairness ranking strategy, we conducted comparison experiments on the TREC 2017 Common Core Track[3] dataset. This dataset contains graded relevance judgment on 27,781 documents in The New York Times Annotated Corpus (NYTcorpus)[4] for 50 queries. This corpus has 26.39M words in total, and an average vocabulary size of 43,710 per query. Different from the Google search data, there is no ranking on the documents for each query. Therefore, we used BM25 to compute a baseline ranking. Then for each query, we selected the top-100 results ranked by BM25 algorithm. Again, we ran 1000 iterations for various algorithms with randomization, and report the average degree-of-bias and relevance scores for each metric. The scores of each algorithm were presented in Table 3.

Similar to the results on Google search data, the non-fair strategies and the *fair-random* performed the worst. The performance for statistical parity were overall better than the disparate impact fairness. This again confirmed our previous observation that balanced results with respect to subtopics could lead to better relevance and diversity. However, we did observe that introducing more randomness improved the score in the case of non-fairness, implying that BM25 did not account for much novelty. In addition, the *page-wise* performed the best for α-nDCG and NRBP for statistical parity, implying that the evaluation results may differ by different datasets.

Comparing Tables 2 and 3, we can see that the intent-aware relevance scores were higher on the Google search dataset. The reasons may be complex, but we try to explain what happens from the perspective of the corpus overall relevance, quality of baseline ranking, mechanism of the *fair ε-greedy* algorithm, and the availability of graded relevance judgment. To begin with, we examined the clusters generated by k-means algorithm on each dataset. With cluster A being the larger size cluster, we found a correlation of $0.8 \sim 0.9$ between the ratio of A's size against the total size of the two clusters, and the ratio of relevant documents in A against total number of relevant documents. This indicated that the clusters were not separating documents based on relevance so the clustering algorithm should not have major impact on the relevance results. Secondly, only less than 30% documents were marked as 'relevant' among all judged documents in NYTcorpus whereas Google search data had 80% documents being relevant. Meanwhile, it was highly likely that the first 100 search results returned by Google were the most relevant among an enormous amount of Web data. The ranking mechanism of Google was presumably significantly better than a simple BM25, hence the most relevant results were concentrated at the top of the rank list.

Finally, we performed significance tests to see whether the increase in relevance scores on NYTcorpus were significant. We found similar results that the scores by fairness ranking algorithms were statistically significantly better than the baseline ($p < .001$).

6.5. Correlation and convergence

To better understand the relationship between relevance, diversity, novelty and fairness, we performed the correlation analysis on the degree-of-bias metric and each of the relevance metrics. Specifically, we computed the correlation between DB and ERR-IA@10, DB and α-nDCG@10, DB and NRBP. The results are shown in Tables 4 and 5 (overall).

[3] https://trec.nist.gov/data/core2017.html
[4] https://catalog.ldc.upenn.edu/LDC2008T19

Table 3
Relevance scores on NYTcorpus under different fairness constraints – (↑) indicates that the higher the better. (-) means not statistically significant compared with Google search's original ranking. All other relevance scores are statistically significant with $p < .001$.

Algorithm	DB↓	ERR-IA@10↑	α-nDCG@10↑	NRBP↑
BM25	0.3096	0.4828	0.5068	0.4718
non-fairness				
0.0-greedy	0.3096	0.4828	0.5068	0.4718
0.01-greedy	0.3085	0.4831(-)	0.5074(-)	0.4718(-)
0.1-greedy	0.2966	0.4887	0.5148	0.4758
0.5-greedy	0.2674	0.5067	0.5343	0.4923
1.0-greedy	0.2528	0.5015	0.5251	0.4883
statistical parity				
top-top (0-greedy)	0.0	0.5447	0.5805	**0.5286**
page-wise	0.0	**0.5487**	**0.5894**	0.5261
fair-random	0.0	0.5469	0.5775	**0.5292**
0.01-greedy	0.0001	0.5447	0.5805	**0.5286**
0.1-greedy	0.0015	0.5448	0.5808	**0.5286**
0.5-greedy	0.0831	0.5451	0.5818	0.5284
1.0-greedy	0.0759	0.5376	0.5726	0.5219
disparate impact				
top-top (0-greedy)	0.0026	0.5285	0.5591	0.5160
page-wise	0.0026	0.5245	0.5569	0.5071
fair-random	0.0026	0.5254	0.5483	0.5140
0.01-greedy	0.0038	0.1388	0.5592	0.5161
0.1-greedy	0.0065	0.5296	0.5606	0.5170
0.5-greedy	0.0416	**0.5364**	0.5670	**0.5221**
1.0-greedy	0.1531	**0.5377**	**0.5727**	**0.5220**

Table 4
Correlation analysis between DB and each of the relevance metrics on Google search data. Under each correlation analysis method, the reported values are the correlation coefficient. "Overall" is considering the correlation regardless of the fairness constraints. Significance level (two-tailed p-value): *:$p < .05$, **:$p < .01$, ***:$p < .001$.

Metric	Pearson	Kendall
overall		
ERR-IA@10	-0.7856***	-0.3871*
α-nDCG@10	-0.7607***	-0.3871*
NRBP	-0.7949***	-0.3988*
statistical parity		
ERR-IA@10	-0.9486***	-0.6061**
α-nDCG@10	-0.9545***	-0.6061**
NRBP	-0.9412***	-0.6364**
disparate impact		
ERR-IA@10	0.5488	0.7807*
α-nDCG@10	0.5957	0.8783**
NRBP	0.5051	0.7807*

On the Google search dataset, we observed that the Pearson's coefficient between DB and each of the relevance metrics ranged from -0.76 to -0.79 for different relevance metrics, with $p < .001$; and the Kendall's τ coefficient ranged from -0.39 to -0.40, with $0.01 < p < .05$. On the NYTcorpus dataset, the Pearson's coefficient ranged from -0.80 to -0.88 for different relevance metrics, with $p < .001$; and the Kendall's τ coefficient ranged from -0.57 to -0.40, with $p < .05$. The correlation results indicated that according to the Pearson's coefficient, fairness, relevance, diversity and novelty were correlated. The correlation implies that, on the one hand, as explained before, diversity helps improve fairness. On the other hand, this observation confirmed that increasing fairness could improve the relevance and diversity, and addressing fairness could help improve relevance, diversity and novelty.

Although there was a correlation, diversity fairness was not the same as diversity. To see this, we performed the correlation analysis on the statistical parity fairness and the disparate impact fairness separately. Specifically, we computed the correlation between DB and ERR-IA@10, DB and α-nDCG@10, DB and NRBP under each fairness constraint (see Tables 4 and 5). In the case of statistical parity, the Pearson's coefficient is around -0.95 for all fairness ranking algorithms on Google search data, and around -0.95 to -0.97 on NYTcorpus, with $p < .001$. While in the case of disparate impact, according to the Pearson's coefficient analysis, there was no statistically significant correlation between DB and any of the intent-aware relevance metrics, on either dataset. According to the Kendall's analysis, we observed the correlation for statistical parity was more statistically significant than for the disparate impact. This shows that only statistical parity fairness is correlated with diversity and relevance, and the disparate impact fairness does not. This implies that the more balanced the search results concerning different subtopics, the higher diversity and intent-aware relevance

Table 5
Correlation analysis between DB and each of the relevance metrics on NYTcorpus data. Under each correlation analysis method, the reported values are the correlation coefficient. "Overall" is considering the correlation regardless of the fairness constraints. Significance level (two-tailed p-value): *:$p < .05$, **:$p < .01$, ***:$p < .001$.

Metric	Pearson	Kendall
overall		
ERR-IA@10	-0.8367***	-0.4529**
α-nDCG@10	-0.7963***	-0.4059*
NRBP	-0.8770***	-0.5706***
statistical parity		
ERR-IA@10	-0.9572***	-0.5954**
α-nDCG@10	-0.9467***	-0.5038*
NRBP	-0.9720***	-0.8397***
disparate impact		
ERR-IA@10	0.6305	0.5855
α-nDCG@10	0.6988	0.6831*
NRBP	0.4535	0.4880

can be expected.

All randomized fair algorithms converge reasonably fast. Within 500 iterations, the algorithms converges to a stable performance regarding relevance.

7. Conclusion

In this work, we investigated the topical diversity bias presented in search engine results. We used Google search data as a lens to surface the existence of bias in top search results. We proposed entropy-based metrics that can measure the degree of bias and effectively evaluate different fairness ranking algorithms. To study the trade-off between fairness ranking and the relevance score, we explored several fairness re-ranking strategies. Our experimental results show that fairness can greatly benefit the relevance and diversity. Algorithms such as *top-top* can meet the fairness requirement while retaining good relevance.

Our work provides the following lessons and implications for future work.

1. When provided with a good relevance-based ranking as baseline ranking, re-ranking strategies that make use of the baseline ranking tend to yield higher relevance;
2. When fairness is of the highest concern, non-randomized algorithms suffice to satisfy the fairness constraint while retaining good relevance;
3. Based on the system relevance judgment, introducing randomness that brings the lower ranked results to the top does not necessarily increase diversity and novelty. Diversity fairness helps improve the relevance and diversity. There is a strong correlation between statistical parity fairness and diversity, implying that these two can benefit each other. The more balanced the results regarding different subtopic groups, the higher relevance and diversity can be expected. Yet the disparate impact fairness does not correlate with diversity, although improving disparate impact fairness can help improve diversity and relevance;
4. In the case of disparate impact fairness constraint, it becomes harder to achieve a high relevance and diversity while satisfying the fairness constraint. In this case, randomized algorithms work the best to introduce personalized amount of diversity and novelty at a low cost of relevance and fairness; and
5. One fairness algorithm does not fit all – the performance a re-ranking algorithm varies depending on the dataset (topics, ranking algorithms, quality of relevance), constraints on fairness and relevance, and evaluation metrics. In addition, the trade-offs between diversity, novelty, and relevance varies by datasets and the choice of fairness ranking algorithms. One must choose wisely according to these factors.

That being said, we must also point out the limitations of the work reported here and the approach taken. The main shortcoming is that our work did not use the state-of-the-art framework or dataset for studying search result diversification. However, this is not simply out of convenience or disregard for repeatability; instead, it is due to the fact that our goal in this work is to pursue *fairness* in search results. While diversifying the results is one way to achieve this goal, it is not the only way and often not the correct way to do so. Creating a set of results with uniform sampling from different aspects of a topic may provide a diversified set, but not fair if we care about how people are exposed to information. For instance, the top 10 image results for query "CEO" often show all men. That is not diversified, nor is it fair from a socio-educational standard. Having images of 5 men and 5 women is diversified, but not fair if the images with men still show up before the images with women. A fair ranking should use not only the diversification of the results, but also be aware of other constraints and preferences, such as the position bias and exposure. At a higher level, the difference between diversity and fairness is that of equality and equity. Not all diverse sets are fair. Fairness is also a more inclusive concept that takes into account not only the side of the information receiver, but also the information provider. We acknowledge that our work related to fairness in this article perhaps ended up being close to addressing diversity and not as broad as just explained, but we hope that the

new framework proposed here allows us and others to continue pushing the envelop by creating and testing new methods of bringing true fairness in search. And while we have provided some discussions on the associations that the concepts of diversity, novelty, and fairness share, there is much to be discussed and debated going forward.

On the practical side, since the Google search results are time sensitive, and the subtopic group assignments may be different as the clustering results may differ. The degree-of-bias and relevance results reported in this article may not be repeatable. Further experiments with gold-standard test collections and state-of-the-art result diversification approaches may help improve the inclusiveness of our findings. However, our evaluation and correlation analysis on both Google search data and the NYTcorpus data showed the robustness of our findings and implications. While not perfect or complete, we believe we have provided at least the first few steps in this new direction exploring diversity fairness in search engines, and the relationships between fairness, diversity, novelty and relevance.

When the relevance and fairness constraint are less emphasized, and diversity is more encouraged, our choices on algorithms become more flexible. A possible direction for future work is to formulate the fairness ranking in such scenarios as an optimization problem and develop efficient re-ranking strategies.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2019.102138

References

Allam, A., Schulz, P. J., & Nakamoto, K. (2014). The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *Journal of medical Internet research, 16*(4), https://doi.org/10.2196/jmir.2642.

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—a user study. *Journal of the Association for Information Science and Technology, 60*(1), 135–149. https://doi.org/10.1002/asi.20941.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). 5*https://doi.org/10.1007/978-94-015-3711-7.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). *Equity of attention: Amortizing individual fairness in rankings. The 41st international acm sigir conference on research development in information retrieval*. New York, NY, USA: ACM405–414. https://doi.org/10.1145/3209978.3210063.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Proceedings of the 30th international conference on neural information processing systems*. USA: Curran Associates Inc4356–4364.

Bråten, I., Strømsø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction, 21*(2), 180–192. https://doi.org/10.1016/j.learninstruc.2010.02.002.

Browne Graves, S. (1999). Television and prejudice reduction: When does television as a vicarious experience make a difference? *Journal of Social Issues, 55*(4), 707–727. https://doi.org/10.1111/0022-4537.00143.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230.

Carbonell, J., & Goldstein, J. (1998). *The use of mmr, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval*. New York, NY, USA: ACM335–336. https://doi.org/10.1145/290941.291025.

Carpineto, C., D'Amico, M., & Romano, G. (2012). Evaluating subtopic retrieval methods: Clustering versus diversification of search results. *Information Processing & Management, 48*(2), 358–373. https://doi.org/10.1016/j.ipm.2011.08.004.

Celis, L. E., Kapoor, S., Salehi, F., & Vishnoi, N. (2019). *Controlling polarization in personalization: An algorithmic framework. Proceedings of the conference on fairness, accountability, and transparency*. New York, NY, USA: ACM160–169. https://doi.org/10.1145/3287560.3287601.

Celis, L. E., Straszak, D., & Vishnoi, N. K. (2018). *Ranking with fairness constraints. 45th international colloquium on automata, languages, and programming107. 45th international colloquium on automata, languages, and programming* Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik28:1–28:15. https://doi.org/10.4230/LIPIcs.ICALP.2018.28.

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). *Expected reciprocal rank for graded relevance. Proceedings of the 18th acm conference on information and knowledge managementCIKM '09*New York, NY, USA: ACM621–630. https://doi.org/10.1145/1645953.1646033.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). *Investigating the impact of gender on rank in resume search engines. Proceedings of the 2018 chi conference on human factors in computing systems*. New York, NY, USA: ACM651:1–651:14. https://doi.org/10.1145/3173574.3174225.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153–163.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). *Novelty and diversity in information retrieval evaluation. Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval*. New York, NY, USA: ACM659–666. https://doi.org/10.1145/1390334.1390446.

Clarke, C. L., Kolla, M., & Vechtomova, O. (2009). *An effectiveness measure for ambiguous and underspecified queries. Proceedings of the 2nd international conference on theory of information retrieval: Advances in information retrieval theoryICTIR '09*Berlin, Heidelberg: Springer-Verlag188–199. https://doi.org/10.1007/978-3-642-04417-5_17.

Demartini, G., & Siersdorfer, S. (2010). *Dear search engine: What's your opinion about…?: Sentiment analysis for semantic enrichment of web search results. Proceedings of the 3rd international semantic search workshop*. New York, NY, USA: ACM4:1–4:7. https://doi.org/10.1145/1863879.1863883.

Drosou, M., & Pitoura, E. (2010). Search result diversification. *SIGMOD Record, 39*(1), 41–47. https://doi.org/10.1145/1860702.1860709.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). *Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference*. New York, NY, USA: ACM214–226. https://doi.org/10.1145/2090236.2090255.

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences, 112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112.

Fortunato, S., Flammini, A., Menczer, F., & Vespignani, A. (2006). Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences, 103*(34), 12684–12689. https://doi.org/10.1073/pnas.0605525103.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347. https://doi.org/10.1145/230538.230561.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). *Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data miningKDD '19*New York, NY, USA: ACM2221–2231. https://doi.org/10.1145/3292500.3330691.

Haas, A., & Unkel, J. (2017). Ranking versus reputation: Perception and effects of search result credibility. *Behaviour & Information Technology, 36*(12), 1285–1298. https://doi.org/10.1080/0144929X.2017.1381166.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning. Proceedings of the 30th international conference on neural information processing systems*. USA: Curran Associates Inc3323–3331.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The information society, 16*(3), 169–185.

Jayasinghe, G. K., Webber, W., Sanderson, M., Dharmasena, L. S., & Culpepper, J. S. (2015). Statistical comparisons of non-deterministic ir systems using two dimensional variance. *Information Processing & Management, 51*(5), 677–694. https://doi.org/10.1016/j.ipm.2015.06.005.

Jiang, Z., Dou, Z., Zhao, W. X., Nie, J., Yue, M., & Wen, J. (2018). Supervised search result diversification via subtopic attention. *IEEE Transactions on Knowledge and Data Engineering, 30*(10), 1971–1984. https://doi.org/10.1109/TKDE.2018.2810873.

Joachims, T. (2002). *Optimizing search engines using clickthrough data. Proceedings of the 8th acm sigkdd international conference on knowledge discovery and data mining*. New York, NY, USA: ACM133–142. https://doi.org/10.1145/775047.775067.

Kamiran, F., & Calders, T. (2009). *Classifying without discriminating. 2009 2nd international conference on computer, control and communication*. IEEE1–6. https://doi.org/10.1109/IC4.2009.4909197.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). *Discrimination aware decision tree learning. Proceedings of the 2010 ieee international conference on data mining*. Washington, DC, USA: IEEE Computer Society869–874. https://doi.org/10.1109/ICDM.2010.50.

Kammerer, Y., & Gerjets, P. (2012). *Chapter 10 how search engine users evaluate and select web search results: The impact of the search engine interface on credibility assessments. Web search engine research*. Emerald Group Publishing Limited251–279. https://doi.org/10.1108/S1876-0562(2012)002012a012.

Kay, M., Matuszek, C., & Munson, S. A. (2015). *Unequal representation and gender stereotypes in image search results for occupations. Proceedings of the 33rd annual acm conference on human factors in computing systems*. New York, NY, USA: ACM3819–3828. https://doi.org/10.1145/2702123.2702520.

Keane, M. T., O'Brien, M., & Smyth, B. (2008). Are people biased in their use of search engines? *Communications of the ACM, 51*(2), 49–52. https://doi.org/10.1145/1314215.1314224.

Kearns, M., Roth, A., & Wu, Z. S. (2017). *Meritocratic fairness for cross-population selection. Proceedings of the 34th international conference on machine learning*1828–1836.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). *Quantifying search bias: Investigating sources of bias for political searches in social media. Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. New York, NY, USA: ACM417–432. https://doi.org/10.1145/2998181.2998321.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). *Counterfactual fairness. Advances in neural information processing systems*. Curran Associates, Inc.4066–4076.

Lahoti, P., Weikum, G., & Gummadi, K. P. (2019). *ifair: Learning individually fair data representations for algorithmic decision making. 35th IEEE International Conference on Data Engineering*1334–1345. https://doi.org/10.1109/ICDE.2019.00121.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1600008. https://doi.org/10.1002/gch2.201600008.

Ludolph, R., Allam, A., & Schulz, P. J. (2016). Manipulating google's knowledge graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *Journal of medical Internet research, 18*(6), https://doi.org/10.2196/jmir.5430.

Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., & Yilmaz, E. (2017). *Auditing search engines for differential satisfaction across demographics. Proceedings of the 26th international conference on world wide web companion*626–633. https://doi.org/10.1145/3041021.3054197.

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). *Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. Proceedings of the 27th acm international conference on information and knowledge management*. New York, NY, USA: ACM2243–2251. https://doi.org/10.1145/3269206.3272027.

Narayanan, A. (2018). 21 fairness definitions and their politics. Retrieved April 26, 2019, from https://www.youtube.com/watch?v=jIXIuYdnyyk.

Novin, A., & Meyers, E. (2017). *Making sense of conflicting science information: Exploring bias in the search engine result page. Proceedings of the 2017 conference on conference human information interaction and retrieval*. New York, NY, USA: ACM175–184. https://doi.org/10.1145/3020165.3020185.

Otterbacher, J., Bates, J., & Clough, P. (2017). *Competent men and warm women: Gender stereotypes and backlash in image search results. Proceedings of the 2017 chi conference on human factors in computing systems*. New York, NY, USA: ACM6620–6631.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication, 12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). *Discrimination-aware data mining. Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*. New York, NY, USA: ACM560–568. https://doi.org/10.1145/1401890.1401959.

Purcell, K., Rainie, L., & Brenner, J. (2012). Search engine use 2012. Retrieved April 26, 2019, from https://www.pewinternet.org/2012/03/09/search-engine-use-2012/.

Santos, R. L. T., Macdonald, C., & Ounis, I. (2015). Search result diversification. *Foundations and Trends in Information Retrieval, 9*(1), 1–90. https://doi.org/10.1561/1500000040.

Shokouhi, M., White, R., & Yilmaz, E. (2015). *Anchoring and adjustment in relevance estimation. Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. New York, NY, USA: ACM963–966. https://doi.org/10.1145/2766462.2767841.

Singh, A., & Joachims, T. (2018). *Fairness of exposure in rankings. Proceedings of the 24th acm sigkdd international conference on knowledge discovery and data mining*. New York, NY, USA: ACM2219–2228. https://doi.org/10.1145/3219819.3220088.

Snow, J. (2018). Bias already exists in search engine results, and it's only going to get worse. Retrieved April 26, 2019, from https://www.technologyreview.com/s/610275/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* Cambridge, MA, USA: MIT press.

Tavani, H. (2016). *Search engines and ethics. The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics, 111*(2), 1017–1031. https://doi.org/10.1007/s11192-017-2298-x.

Wang, X., Bendersky, M., Metzler, D., & Najork, M. (2016). *Learning to rank with selection bias in personal search. Proceedings of the 39th international acm sigir conference on research and development in information retrieval*. New York, NY, USA: ACM115–124. https://doi.org/10.1145/2911451.2911537.

Weber, I., Garimella, V. R. K., & Borra, E. (2012). *Mining web query logs to analyze political issues. Proceedings of the 4th annual acm web science conference*. New York, NY, USA: ACM330–334. https://doi.org/10.1145/2380718.2380761.

Wu, Y., Zhang, L., & Wu, X. (2018). *On discrimination discovery and removal in ranked data using causal graph. Proceedings of the 24th acm sigkdd international conference on knowledge discovery and data mining*. New York, NY, USA: ACM2536–2544. https://doi.org/10.1145/3219819.3220087.

Xu, J., Xia, L., Lan, Y., Guo, J., & Cheng, X. (2017). Directly optimize diversity evaluation measures: A new approach to search result diversification. *ACM Trans. Intell. Syst. Technol. 8*(3), 41:1–41:26. https://doi.org/10.1145/2983921.

Yang, K., & Stoyanovich, J. (2017). *Measuring fairness in ranked outputs. Proceedings of the 29th international conference on scientific and statistical database management*. New York, NY, USA: ACM22:1–22:6. https://doi.org/10.1145/3085504.3085526.

Yu, H.-T., Jatowt, A., Blanco, R., Joho, H., Jose, J., Chen, L., & Yuan, F. (2017). *A concise integer linear programming formulation for implicit search result diversification. Proceedings of the tenth acm international conference on web search and data miningWSDM '17*New York, NY, USA: ACM191–200. https://doi.org/10.1145/3018661.3018710.

Yu, H.-T., Jatowt, A., Blanco, R., Joho, H., Jose, J. M., Chen, L., & Yuan, F. (2018). Revisiting the cluster-based paradigm for implicit search result diversification. *Information Processing & Management, 54*(4), 507–528. https://doi.org/10.1016/j.ipm.2018.03.003.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). *Fa*ir: A fair top-k ranking algorithm. Proceedings of the 2017 acm on conference on information and knowledge management*. New York, NY, USA: ACM1569–1578. https://doi.org/10.1145/3132847.3132938.