How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval

Ruoyuan Gao
Department of Computer Science
Rutgers University, New Brunswick, NJ
ruoyuan.gao@rutgers.edu

Chirag Shah
Information School
University of Washington, Seattle, WA
chirags@uw.edu

ABSTRACT

The presence of bias in today's IR systems has raised concerns on the social responsibilities of IR. Fairness has become an increasingly important factor when building systems for information searching and content recommendations. Fairness in IR is often considered as an optimization problem where the system aims to optimize the utility, subject to a set of fairness constraints, or optimize fairness while guaranteeing a lower bound on the utility, or jointly optimize for both utility and fairness to achieve an overall satisfaction. While various optimization algorithms have been proposed along with theoretical analysis, in real world applications, the performance of different optimization algorithms often heavily depend on the data. Therefore, it is consequential to ask what is the solution space characterized by the data, what effect does introducing fairness bring to the system, and can we identify this solution space to help us trade-off different optimization policies and guide us to pick suitable algorithms and/or make adjustments on data? In this work, we propose a framework that offers a novel perspective into the optimization with fairness constraints problems. Our framework can effectively and efficiently estimate the solution space and answer such questions. It also has the advantage of simplicity, explainability, and reliability. Specifically, we derive theoretical expressions to identify the fairness and relevance bounds for data of different distributions, and apply them to both synthetic and real world datasets. We present a series of use cases to demonstrate how our framework is applied to facilitate various analyses and decision making.

ACM Reference format:
Ruoyuan Gao and Chirag Shah. 2019. How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval. In *Proceedings of The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval, Santa Clara, CA, USA, October 2–5, 2019 (ICTIR '19),* 8 pages.
https://doi.org/10.1145/3341981.3344215

1 INTRODUCTION

Fairness is a topic that is increasingly receiving attention in information retrieval (IR). IR systems should not only excel at helping users find what they need, but also bear the social responsibility of being fair [2, 10, 13, 29, 30]. An array of studies have shown that biased search and recommendation results can separate users from potentially nutritious information that differs from their personal preference (*filter bubble*) [7], impact users' perception of opinions and events [1, 6], reinforce social stereotypes [17, 26], manipulate users' decision making [12], and lead to unfair distribution of opportunities and resources [5, 18, 21, 28]. As a result, search and recommendation systems must consider the fair representation of results without sacrificing the system utility. For two-sided marketplace platforms that have customers on both the demand (e.g., users) and the supply side (e.g., retailers, artists) such as Amazon, Netflix, and Spotify, being fair to items recommended is particularly important for satisfying the suppliers and ensuring the exposure of and opportunities for less popular content providers, which is crucial for a diverse economical and social development [21]. Such systems must find a balance between optimizing the consumer satisfaction and the supplier fairness.

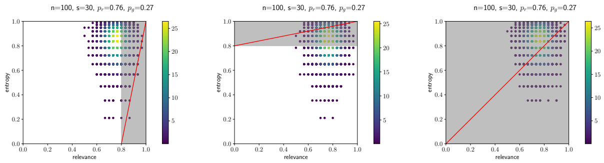
To incorporate fairness constraints into the algorithmic framework and an IR system, we can attempt different optimization policies depending on the emphasis on fairness or system utility. For applications where fairness constraints must be strictly enforced, optimizing for utility while subject to the fairness constraints may be the policy to opt for. In the case where a system aims to trade-off between various utility factors and fairness constraints, jointly optimizing for each factor and constraint seems more promising to achieve an overall good satisfaction. Ideally, researchers try to develop general algorithms and frameworks that are intended to work for all kinds of data. But in reality, adopting an algorithm or framework independent of data may be both impractical and unreasonable. Therefore, whether the objective is clear or not, it is generally a good idea to first associate with the data and application before selecting a policy.

Imagine a solution space of all possible utility values and degrees of fairness, we can depict each optimization policy in this space, and then analyze what solutions can we achieve with each policy. Figure 1 illustrates an example of different optimization policies against the solution space. It is easy to see that once the solution space is defined, the limitations of each policy in terms of optimal values are fixed. In other words, while selecting a policy, regardless of what algorithms are behind that policy, one cannot achieve a better result than the solution space defined by the data. Another advantage of this solution space analysis is that if we cannot improve on the algorithm side, maybe we should consider issues that reside within the data, and search for explanations and possible improvement from the data side. For example, in image search engines, the gender bias in occupational stereotypes may be a result of the ranking algorithm, but may also indicate the images in the machine-judged

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344215

relevant search space is unfairly distributed. Note that machine-judged relevance are different from user judgment due to various information need and characteristics of various users. An image that is considered relevant by a user may not be judged relevant by the system and thus, is likely to be eliminated from the search space. As a result, a possible solution to increase fairness is to improve on the search space, which requires adjusting the algorithm for relevance judgment in the system.



(a) optimize fairness with relevance ≥ 0.8

(b) optimize relevance with fairness ≥ 0.8

(c) jointly optimize fairness and relevance

Figure 1: Exploration space (shaded area) and a possible hypothetical algorithm (red line) for each different optimization policy. This is a density plot where points in this plot are random samples from the solution space.

In particular, we are interested in the following questions that are motivated by the scenarios discussed above: 1) Do different fairness constraints affect user satisfaction, engagement, and how? 2) Given an IR system, what would fairness constraint bring into the system? For instance, do such constraints affect the choice of learning algorithms, parameter tuning, decision making, and evaluation metrics? 3) Consider the entire search space defined by data, what would be the relationship between fairness and relevance? Previous studies have found that user satisfaction tend to decrease as the amount of fairness increases when investigating strategies that jointly optimize for machine-judged relevance and supplier fairness [21]. But, does there exist a solution where a desired relevance and a desired fairness constraint can be achieved at the same time? 4) Given the trade-offs between each component of a system, for a particular dataset, can we identify the optimal solution values that the system can achieve? What are the best strategies to achieve those optimal values?

If we have the solution space discussed above available, those questions become straightforward to answer and analyze. Therefore, a more interesting question to ask is: given a dataset or search space, can we identify the solution space? What is the region that bounds the solution space and with what probability? In this paper, we propose a novel theoretical framework that estimates the solution space by sampling from the data space. Based on the estimated solution space, we then estimate the optimal values considering different weights on each dimension of the data. Specifically, our paper makes the following contributions.

- We propose a theoretical framework that offers a novel perspective into the fairness as optimization problems. With our framework, system designers can easily plug in their own utility functions and fairness constraints to get an overview of what is the solution space characterized by the data, evaluate the likelihood of achieving a predetermined optimization goal, identify the optimal results that can be achieved given the data, evaluate how well an algorithm may help achieve the goal, and trade-off between different optimization policies to select the most suitable policy. All of these benefits can be obtained before actually implementing any complex optimization algorithms or carrying out heavy experiments, thus saving considerable time and resources.
- We provide several theoretical analyses of our framework to demonstrate how such analyses can serve as guidelines for the applicability and effectiveness of our framework. Our framework not only offers hints on whether the data coverage and algorithms need adjustment, but it also highlights a direction of how such adjustment should be made, in order to achieve the desired fairness and utility. With theoretical results, our framework can be easily generalized to optimization problems over multiple dimensions, data of different distributions, and various personalized utility functions and fairness constraints.
- We demonstrate the application of our framework on both synthetic and real world data and show how this framework can be used to facilitate vast analyses and decision making.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes our problem and framework setting. Section 4 presents our framework with methodologies to analyze and several theoretical results. Section 5 demonstrates the use of our framework with synthetic and real world data. We conclude our work in Section 6.

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. A large number of studies have investigated possible solutions to address fairness in IR, which provide critical views of the fairness as optimization problems.

2.1 Bias in IR systems

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 IR systems, fairness is often concerned in terms of *individual fairness* and *group fairness*. With individual fairness requiring similar individuals being treated similarly, group fairness often requires that the protected groups (e.g., demographic categories, political perspectives, topical diversities and opinion polarities) are fairly presented in the retrieved items. Singh and Joachims [28] proposes to view group fairness from a perspective of exposure. Celis et al. [8], Dressel and Farid [11] study the threshold based fairness definitions in which the number of items from each group is bounded by a minimum and maximum threshold. Biega et al. [5], Chen et al. [9] investigate individual fairness in the ranking systems. See Narayanan [22] for a more comprehensive explorations of fairness definitions.

Bias in IR may arise from the source data, algorithmic or system bias, and cognitive bias [2, 23, 24]. Techniques and algorithms that learn from and mirror real world statistics may unavoidably carry social bias from the original data to the IR systems [4]. Baeza-Yates [2] discusses the presence, measurement, and cause of bias on the Web. Kulshrestha et al. [19] develops a framework that quantifies the amount of bias that arise from the data source and from the ranking system. Kay et al. [17], Otterbacher et al. [25, 26] show that bias can surface in search engine results because of the bias in data and algorithms. From a cognitive perspective, biased search or recommendation results can lead to unfair distribution of opportunities and resources [5, 18]. Users often rely on the presentation, especially ranking, of the search and recommendation results for credibility judgment, resource selection, and a belief and attitude shaping of information [1, 3, 6, 7, 14, 16, 23, 27] as well as preference and decision making [12]. Due to the differences in demographic and cultural background, purpose and expectations, users of an IR system may display different satisfactions [20]. Business to consumer online market platforms and online multimedia recommendation systems face the challenge of satisfying the demand of both the consumers and suppliers. Simply optimizing for consumer satisfaction may negatively impact supplier fairness [21].

Our framework is built upon addressing such bias issues in IR systems. Instead of limiting to target a particular system, we aim to develop a general framework that can be used across various types of IR systems. Our framework is able to capture the influences of a heavily utility-focused system that underweights the issue of bias from a theoretical perspective.

2.2 Fairness as an optimization problem

Fairness in IR is often modeled as an optimization problem with fairness constraint. There are primarily three types of optimization goals: to optimize utility (often represented by relevance) subject to a bounded fairness constraint, to optimize fairness while constraining utility with a lower bound, and to jointly optimize for both utility and fairness. The majority of the previous works are focused on the first type of optimization. Zehlike et al. [31] defines a fair top-k problem to ensure that the proportion of protected groups in the top-k ranking remains above a given threshold. Joseph et al. [15] introduces the multi-armed bandit problem with fairness constraints and shows the gap of regret between fair and unfair learning. Celis et al. [7] proposes a fast and low regret algorithm for the fairness-constrained bandit optimization in personalization. Celis et al. [8] formulates the fairness ranking as a constrained matching problem. The objective is to maximize the ranking score while bounding the number of items with each attribute on each position. Singh and Joachims [28] proposes a conceptual and computational framework for fairness ranking which maximizes the utility while satisfying some fairness constraints. For the second type of optimization, Biega et al. [5] proposes a fair ranking solution based on integer linear program. A few examples of work take on the perspective of joint optimization. Mehrotra et al. [21] considers supplier fairness versus user satisfaction in a two-sided marketplace setting where the goal is to jointly optimize for both fairness and relevance.

Our work is inspired by these efforts to achieve an optimal solution, whether an optimal utility, or an optimal fairness, or a joint optimal for both. While we do not propose algorithms for a particular optimization problem, we provide an illustration of how different optimization policies compare. Before delving deep into complex algorithms and theoretical analysis of a problem, we aim to develop a framework that envisions the possible solution space. The algorithms and framework proposed in previous works aim to address a set of optimization problems with little or no dependence on the data. Our work differs in that we associate data with objective functions, and emphasize the importance of considering the data factor when addressing fairness problems.

3 PROBLEM DEFINITION

Our problem definition and framework setting are motivated by [21] where upon receiving a user query, the system explores in the data space and returns a set of items that are considered relevant to the query. Formally, let $D = \{a_1, a_2, \ldots, a_N\}$ be a set of N items defining the dataset or search space. Each item a_i is associated with $k \geq 2$ properties denoted by vector $\langle p_1, p_2, \ldots, p_k \rangle$. Let $f_i(D)$ be a predefined function that summarizes the i-th dimension property values in set D.[1] Let $f(D) = \langle f_1, f_2, \ldots, f_k \rangle$ be the function that summarizes the set D. Assume property values on each dimension are independent identically distributed (i.i.d.) drawn from a distribution. The distribution of each dimension may or may not be the same. Let solution set $S_j \subset D$ denote a set of items returned by the system regarding a user query, $|S_j| = n$. The solution space $S = \{S_1, S_2, \ldots, S_j, \ldots\}$ is all subsets of D of cardinality n. Let $\min\{f_i(S_j)\}$ and $\max\{f_i(S_j)\}$ be the lower and upper boundaries of S on the i-th dimension. $R_i = [\min\{f_i(S_j)\}, \max\{f_i(S_j)\}]$ is the range of f_i. The goal of our framework is to estimate the region $R = \langle R_1, R_2, \ldots, R_k \rangle$ of S confined by boundaries on each dimension of the data.

4 DEPICTING BOUNDARIES

To begin with, assume each item is composed of two dimensions, i.e., $k = 2$, $a_i = \langle r_i, g_i \rangle$ and $f = \langle f_r, f_g \rangle$. The first dimension captures the relevance score of the item, the second dimension captures the protected group information. This simple assumption captures many real world application scenarios. It is also the assumption in previous works on fairness [5, 13, 28]. For example, consider each item as an image with respect to an occupational query in search engines; apart from the relevance score, each image can be associated with a gender property, in which case one may wish to balance the images of each gender group in the search results. In online question and answering communities, all the answers can be viewed as items where the group is topic aspect or opinion polarity. In two-sided marketplace platforms, the group associated with each item is often identified as producer or supplier, and the relevance denotes the consumer satisfaction.

4.1 Simple case

A simple case is that each dimension of the data $a_i \in D$ is from a Bernoulli distribution. Assume binary relevance score and group assignment (i.e., two groups), p_r be the probability that an item is relevant, g_i be the probability that an item is from group 1. Then r_i and g_i are distributed according to Bernoulli distribution, $r_i \sim \text{Bernoulli}(p_r)$, $g_i \sim \text{Bernoulli}(p_g)$. Given a set $S \subset D$ of n items, let f_r be the average relevance score of set S, f_g be the entropy of group memberships, $\bar{p_g}$ be the proportion of items in group 1,

$$f_r = \frac{\sum r_i}{n}, \tag{1}$$

[1] When the context is clear, we use f_i to denote the function value given a set.

$$
\begin{aligned}
f_g = H(\bar{p}) &= -\bar{p_g}\log_2\bar{p_g} - (1-\bar{p_g})\log_2(1-\bar{p_g}) \\
&= -\frac{\sum g_i}{n}\log_2\frac{\sum g_i}{n} - (1-\frac{\sum g_i}{n})\log_2(1-\frac{\sum g_i}{n}).
\end{aligned}
\tag{2}
$$

Since r_i and g_i are i.i.d. random variables, the sum of r_i and the sum of g_i are distributed according to Binomial distribution:

$$\sum r_i \sim \mathrm{B}\left(n, p_r\right), \; \sum g_i \sim \mathrm{B}\left(n, p_g\right).$$

Let X denote the the random variable $\sum r_i$, Y denote the random variable $\sum g_i$. Hence $X \sim \mathrm{B}\left(n, p_r\right)$, $Y \sim \mathrm{B}\left(n, p_g\right)$. According to Chebyshev's inequality, for any real number $t > 0$,

$$P(|x-\mu| \geq t\sigma) \leq \frac{1}{t^2},$$

where μ is the distribution mean, σ^2 is the distribution variance. Given a probability $1-q_r$ that X is within $t\sigma$ distance from the mean, we can bound X as follows: since $X \sim \mathrm{B}\left(n, p_r\right)$, $\mu = np_r$, $\sigma^2 = np_r(1-p_r)$, let

$$P\Big(|X - np_r| \geq t\sqrt{np_r(1-p_r)}\Big) \leq \frac{1}{t^2} \leq q_r, \tag{3}$$

we get $t \geq \sqrt{1/q}$, and

$$X \leq np_r - \sqrt{\frac{np_r(1-p_r)}{q_r}}\text{ , or } X \geq np_r + \sqrt{\frac{np_r(1-p_r)}{q_r}}.$$

Therefore, with probability $1-q_r$,

$$X \in [np_r \pm \sqrt{np_r(1-p_r)/q_r}], \tag{4}$$

hence

$$f_r \in [p_r \pm \sqrt{\frac{p_r(1-p_r)}{nq_r}}]. \tag{5}$$

Similarly, we can obtain the range of Y and f_g such that with probability $1-q_g$, f_g is within the derived range.

Assume relevance and group property are independent, i.e., random variables X and Y are independent. We can compute the joint distribution to depict a region for the range of X and Y, consequently a region for $\langle f_r, f_g\rangle$ values. Let R_r denote the range of f_r with probability $1-q_r$, R_g denote the range of f_g with probability $1-q_g$, then

$$P[f_r \in R_r, f_g \in R_g] \geq (1-q_r)(1-q_g). \tag{6}$$

The independence is a reasonable assumption since in real world applications, the protected attributes that require fairness are often assumed to be independent of the system's decision making. For example, in the criminal risk assessment task where the system predicts the likelihood of future recidivism for a particular defendant, the predicted results should be independent of the dependent's gender [11]. This assumption of independence must also be applied to resume search systems. A fair resume search engine should not rank candidates with a preference related to gender [9]. Question and answering platforms are another example; if the query is an opinion question, then whether an answer is relevant or not should be independent of its opinion polarity.

4.2 Unknown distribution

When we do not know the exact underlying distributions of r_i and g_i, but we know that r_is are i.i.d. random variables drawn from a distribution with mean μ_r and finite variance σ_r^2, g_is are i.i.d. random variables drawn from a distribution with mean μ_g and variance σ_g^2, we can try to apply the Central Limit Theorem. Let $\bar{r} = \sum r_i/n$ and $\bar{g} = \sum g_i/n$, $s^2 = \frac{1}{n-1}\sum_i^n (r_i - \bar{r})^2$, where $\bar{r}$ and $\bar{g}$ are the mean of the sampling distribution, s^2 is the the unbiased estimator of sample variance. According to the Central Limit Theorem, as the sample size n goes to infinity, the distributions of $\sqrt{n}(\bar{r}-\mu_r)$ and $\sqrt{n}(\bar{g}-\mu_g)$ approximate normal distributions each, despite r_i and g_i's underlying distributions. Specifically, we have

$$\frac{\left(\bar{r}-\mu_r\right)}{\sigma_r/\sqrt{n}} \xrightarrow{d} N\left(0,1\right), \; \frac{\left(\bar{g}-\mu_g\right)}{\sigma_g/\sqrt{n}} \xrightarrow{d} N\left(0,1\right).$$

Now we can easily compute the confidence intervals by referencing the z-values for the standard normal distribution to bound the sample mean $\bar{r}$ and $\bar{g}$. Consequently, we can bound the range of f_r and f_g given the range of $\bar{r}$ and $\bar{g}$. Note that the Central Limit Theorem only applies when the sample size n is sufficiently large or when the underlying distribution is normal. If none of such conditions are met, or if the underlying distribution variance is unavailable, we can use sample variance s^2 to estimate the population variance. Then the variables

$$\frac{\bar{r}-\mu_r}{s\sqrt{n}}\text{ , and }\frac{\bar{g}-\mu_g}{s\sqrt{n}}$$

each has a student t-distribution. With t-distribution, we can reference the t-values to compute the confidence intervals. Specifically, with the corresponding z-value for the standard normal distribution and t-value for the t-distribution, we can obtain the following confidence intervals respectively with a $100(1-q)$ confidence level,

$$\left(\bar{x} - z\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z\cdot\frac{\sigma}{\sqrt{n}}\right) \tag{7}$$

if the underlying distribution is normal or the sample size n is sufficiently large, and

$$\left(\bar{x} - t\cdot\frac{s}{\sqrt{n}}, \bar{x} + t\cdot\frac{s}{\sqrt{n}}\right) \tag{8}$$

if the Central Limit Theorem cannot be applied, and the underlying distribution is bell-shaped. Here $\bar{x}$ is to be replaced by $\bar{r}$ and $\bar{g}$, and σ replaced by σ_r and σ_g accordingly. This means with probability $1-q$, the above confidence interval will contain the true population mean μ. So with probability $1-q$,

$$\bar{x} \in \left(\mu \pm z\cdot\frac{\sigma}{\sqrt{n}}\right), \bar{x} \in \left(\mu \pm t\cdot\frac{s}{\sqrt{n}}\right), \tag{9}$$

where μ is the true population mean, which is replaced by μ_r and μ_g in our case. Now that we have the range of $\bar{r}$ and $\bar{g}$, we can directly compute the range of f_r and f_g. To apply the t-distribution here, we are making the assumption that the true population is a normal distribution, which is a reasonable assumption for many real world large scale data.

4.3 Optimal points

Once we have the boundary depicted according to some given probability q, we can estimate the optimal values for both dimensions that can be achieved within this region. The optimal values depend on the weights of each dimension which can be customized as needed. Let $w_r \neq 0$ and $w_g \neq 0$ be the weights of f_r and f_g, respectively, representing the importance of each dimension that a system wishes to address. Then within the range defined by R_r and R_g, we can identify the optimal value by maximizing f_r (or equivalently f_g) subject to

$$\frac{f_r}{f_g} = \frac{w_g}{w_r}, f_r \in R_r, f_g \in R_g. \tag{10}$$

Note that assuming the data contains at least one item of relevance score 1, and each group has at least one item (otherwise there will not be a cluster of two groups), then $f_r \in (0, 1]$ and $f_g \in (0, 1]$. We must address that the optimal points derived here are only theoretically possible considering a sufficiently large sample space. It is highly likely that these optimal points are not observed in the available dataset. However, this gives a direction to push the optimization algorithms that jointly optimizes for both dimensions.

4.4 Generalization

Our framework can be generalized to allow for various summarizing functions and multiple dimensions. Due to the space limitation, we cannot elaborate on every single scenario. Instead, we demonstrate the methodology for generalization with a few examples.

Customized functions. Depending on the specific application scenarios, users can plug in customized functions instead of the mean relevance and the entropy described here. For instance, the utility function can be recall, discounted cumulative gain, ads click through rate, user engagement time, to name a few. The fairness measurement function can be distance to the constrained fairness conditions. For example, with g_i dimension representing a binary membership of a protected group, p being the proportion of items of membership 1 and $1-p$ being the proportion of items of membership 0, the **statistical parity** fairness constraint can be written as $p = 1 - p$ or $p = 0.5$. In our framework, the fairness degree of any set of items can be defined in terms of the distance to $p = 0.5$,

$$f_g = |\bar{p} - 0.5| = |\frac{\sum g_i}{n} - 0.5|. \tag{11}$$

Then, the optimal points regarding the g_i dimension are those that minimizes f_g. The **disparate impact** is defined as

$$\frac{\bar{p}}{1-\bar{p}} = \frac{P(r_i = 1|g_i = 1)}{P(r_i = 1|g_i = 0)},$$

where the left hand side probabilities are from the sample estimation, and the right hand side probabilities are from the population distribution. Again, we can write f_g as a distance between the left hand side and right hand side such that the optimal value is achieved when the distance is 0,

$$f_g = |\frac{\bar{p}}{1-\bar{p}} - \frac{P(r_i = 1|g_i = 1)}{P(r_i = 1|g_i = 0)}|. \tag{12}$$

Multiple dimensions. Our framework can also be generalized to allow for multiple dimensions. This is particularly useful when a system needs to consider more than two factors. For instance, a resume search engine may need to consider the gender, racial and other demographic fairness along with the candidates' qualification. A two-sided marketplace platform may need to design multiple metrics for measuring user satisfaction while considering the fairness for content providers. Here we show how to generalize to the case $k > 2$ with each dimension having a Bernoulli distribution as an example. For other distributions, we can generalize in the same way using the results from Section 4.2. Assume we need to estimate the region R with probability $1 - q$. We can estimate each of the k dimensions with a confidence level $1 - q/k$ according to the Bonferroni correction. Applying the analysis in Section 4.1 to first derive the range in Equation 4 for each dimension with probability $1 - q/k$. Then compute the range of each dimension's summarizing function similar to Equation 5. Now that we have the range of f_i with probability $1 - q/k$ for each dimension $i \in [1, k]$, we can get

$$P[R] = P[\{f_i \in R_i | i = 1, 2, \ldots, k\}] \geq (1 - q). \tag{13}$$

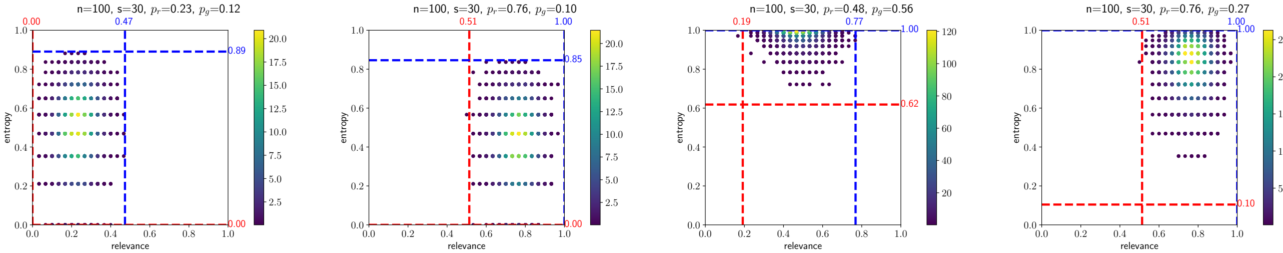
5 DEMONSTRATION OF THE FRAMEWORK

In this section, we demonstrate how to apply our framework to facilitate various analysis and decision making with a few synthetic data examples and real world data examples. We illustrate this with 2-dimensional data where the first dimension of the data point $a_i = \langle r_i, g_i \rangle$ denotes relevance $r_i \in \{0, 1\}$, and the second dimension is group assignment $g_i \in \{0, 1\}$. Here we assume the relevance and group takes on binary values for simplicity. The analysis for multiple dimensions and other distributions can be performed similarly, and is therefore eliminated due to space limitation.

5.1 Datasets

Synthetic dataset. Given a pair of probabilities (p_r, p_g), we randomly generate 100 points of parameters $\langle r_i, g_i \rangle$ where $r_i \sim \text{Bernoulli}(p_r)$, $g_i \sim \text{Bernoulli}\left(p_g\right)$, $i \in [1, 100]$. Denote these 100 points as dataset $D(p_r, p_g)$. We compute the population mean μ_r and μ_g on the relevance and fairness dimension, respectively. For each pair of (p_r, p_g), we repeat this process multiple times and observe that the generated population distribution does not deviate far from the true distribution. Subsequently, for each pair of parameter choices, we randomly select one generated dataset for analysis. Due to space limitation we only demonstrate on dataset with 4 pairs of parameters: $(0.3, 0.1)$, $(0.7, 0.1)$, $(0.5, 0.5)$, $(0.7, 0.3)$.

***YOW* RSS feeds dataset** [32]: We take the same dataset as used in [7, 28]. This dataset is a collection of 21 users' feedback on RSS news feeds. Each feed is associated with a source identifier which we consider as group assignment. Each news article is judged for relevance on a scale from from 1 to 5, with 5 being the most relevant. We select all news feeds from topics that contain "people" keyword and are of the top two sources (with source identifier 14 and 8,157). We consider the two sources as two groups that we wish to consider for fairness. For relevance, roughly half of the selected feeds are rated with a score 4, and the rest with scores 2 and 3. We convert news feeds with relevance score 4 to relevance 1, and the rest to relevance 0. This results in 48 news feed data points.



(a) $D(0.3, 0.1)$

(b) $D(0.7, 0.1)$

(c) $D(0.5, 0.5)$

(d) $D(0.7, 0.3)$

Figure 2: Solution space of synthetic dataset generated from different distributions. p_r is the probability of an item being relevant, p_g is the probability of an item belonging to group 1. The solution space is illustrated by density plot after randomly sampling 10^4 times. The horizontal dashed lines bound the entropy and the vertical lines bound the relevance, each with probability 0.9. The red lines are lower bounds and blue lines are upper bounds.

5.2 Analysis

We consider the task of imposing fairness as an optimization problem. Given a dataset D, an IR system performs a retrieval task by returning a solution set of $S \subset D$ items to the user. The goal is to maximize the utility of S while ensuring that S presents items of different groups in a fair fashion. Let $|S| = s$. For demonstration, we use the average relevance $f_r = \sum r_i/s$ as the utility metric, and entropy $f_g = H(\sum g_i/s)$ as the measure of group fairness. Note that the utility function and degree of fairness can be replaced with any customized functions.

With the proposed framework, we test a few ideas of possible analysis that can be carried out. On the synthetic data, let us say we want to select 30 items from the 100 candidates, $s = 30$. On the *YOW* RSS feeds dataset, we select $s = 20$ items from the 48 items.

1. What are the possible relevance and fairness scores of a solution set S (the *solution space*)? How does the distribution of the dataset affect the solution space?

Since the solution space is the set of all subset $S \subset D$ with a predetermined cardinality s, the number of solution sets is $C(N, s)$ (N choose s). This number can be extremely large. For instance, on the synthetic data, $C(100, 30)$ is approximately 3×10^{25}. For a dataset that is as large as one million items, $C(10^6, 100)$ will be approximately 10^{442}. For quick overview and analysis on a large dataset, we can randomly sample from the solution space and plot the density, successfully avoiding the time and effort enumerating all possible solutions. Note that the solution space is determined by the dataset and the summarizing functions only. Therefore, the theoretical bounds are independent of the sampling approach.

For the synthetic data, we first compute the range of random variables $\sum r_i$ and $\sum g_i$ according to Equation 4. For example, in dataset $D(0.5, 0.5)$, all points are generated from Bernoulli(0.5) for both dimensions. Using the true distribution mean $p_r = p_g = 0.5$, let $q_r = q_g = 0.1$, we have $\sum r_i \in [6.34, 23.66]$ and $\sum g_i \in [6.34, 23.66]$. So $f_r \in [0.21, 0.79]$, $f_g \in [0.74, 1]$, each with probability 0.9. [2] This defines a regional boundary for relevance and fairness (entropy) with probability 0.81. Figure 2 (c) plots the probability density and boundary. We can see that most of the solution sets have a high entropy approaching the optimal 1 and a relevance around 0.5. For solutions that have a high relevance close to 1, we can expect a high entropy as well but the number of such solutions is going to be very small. In other words, it is less likely that we are able to observe such solutions from the small dataset. For a different distribution, for instance, $D(0.3, 0.1)$, the solution distribution will look a lot different (Figure 2 (a)). It is almost impossible to observe a solution with $f = \langle 1, 1 \rangle$ on the data generated from this distribution. For the *YOW* RSS feeds data, we can perform the same analysis (Figure 3). The means of dimensions r_i and g_i are 0.56 and 0.50 respectively. We can see that the solution sets are concentrated around the mean relevance and entropy 1. We can expect that with high probability, the solution sets from this dataset will have high entropy, but we may not be able to find a solution set that has relevance close to 1.

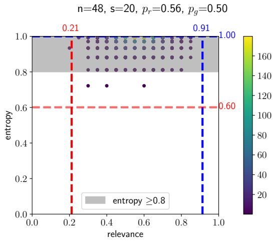


Figure 3: Solution space of *YOW* RSS feeds dataset. The shaded area is the exploration space of the optimization policy that optimizes relevance subject to a minimum fairness (entropy) threshold $T = 0.8$.

The examples above demonstrate that the proposed framework can aid in depicting the solution space given a particular dataset. This provides clues for what are the best relevance and entropy of an item set that the system can return, and how likely such an item set can be observed in the data, all despite the choice of the optimization algorithms.

2. What is the effect of introducing fairness into the system? What is the trade-off between fairness and relevance?

Many previous works have shown that as we increase the amount of fairness, we observe a decrease in the system performance in terms of retrieval accuracy, relevance score, user satisfaction on personalization and recommendation satisfaction, etc. [5, 7, 21, 28]. While this is true for the optimization algorithms and the dataset examined in their works, it may not be the case

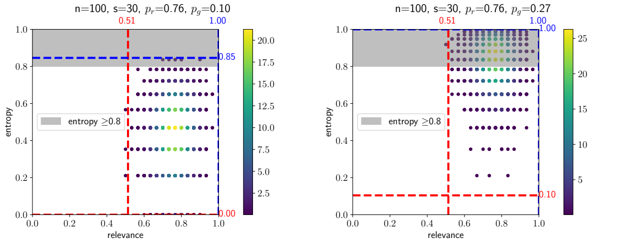
[2] Numbers are rounded to the nearest hundredth.

for all algorithms and all dataset. From Figure 2 (c-d), we can see that fairness does not necessarily have a detrimental effect on relevance. For some data such as $D(0.3, 0.1)$, and $D(0.7, 0.1)$ in Figure 2 (a-b), optimizing for fairness does lead to solution sets with lower relevance. But for data like $D(0.5, 0.5)$ and $D(0.7, 0.3)$ in Figure 2 (c-d), we are able to optimize fairness while also optimizing relevance. In other words, the effect of fairness on system performance hinges on the both optimization algorithm and dataset. With a dataset that tends to present a good fairness overall, we can achieve optimal fairness and maximize relevance at the same time without worrying much about the trade-offs between them. The benefit of our proposed framework is that we can visualize the relationship between fairness and relevance on a given dataset without having to implement complex algorithms and conduct extensive experiments. The impact of introducing fairness and the trade-offs between fairness and system utility can be easily analyzed through our framework.

3. What is the exploration space of each optimization policy? Which policy should one pick?

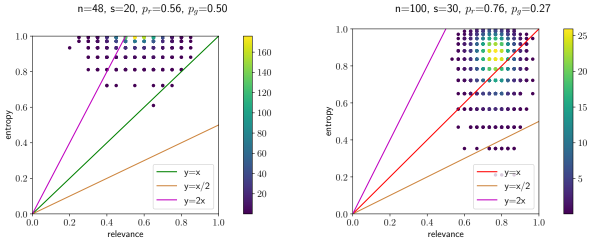
We now demonstrate how to use the proposed framework to facilitate decision making regarding optimization policies. First, consider the policy that optimizes for relevance subject to a minimum threshold T of fairness constraint. Let us draw a horizontal line that represents this fairness threshold $y = T$. The exploration space for this type of policy is then the area above this horizontal line. For example, let $T = 0.8$, then the shaded area in Figure 3 is the exploration space on the *YOW* RSS feeds data. Since $y = 0.8$ is close to the lower boundary of entropy, this policy is basically exploring in the major solution space. In the solution space, with high probability, a solution set with optimal entropy exists in the dataset, but high relevance that is close to 1 is unlikely to be observed. Meanwhile, the solution space shows that for high relevance, the entropy is also likely to be high. Because this optimization policy optimizes for relevance, it can be expected to find solution sets that enjoys both high relevance and entropy if such sets exist. For such dataset where solution sets are concentrated around entropy 1, frameworks like the one developed in [28] will be very suitable. On a different dataset, however, this policy may not be able to reach the solutions that are optimal for both dimensions. Consider the synthetic data $D(0.7, 0.1)$, in Figure 4 (a), the minimum threshold $T = 0.8$ is obviously above the entropy upper bound of the solution space. So the policy is essentially exploring a space that cannot be achieved with the given dataset. Another example in Figure 4 (b) is the dataset $D(0.7, 0.3)$. Since the goal of this optimization policy is to achieve the highest relevance while subject to the minimum fairness threshold, we can see that the solution sets found by this policy will have optimal relevance but not optimal entropy by simply observing the density plot. Such analyses are more obvious when the policy is subject to the optimal fairness constraint. In this special case, the exploration space collapses to a horizontal line $y = 1$. The policy will find whatever set that has the maximum relevance alone this line, and such a set may not even exist in the dataset if the solution space around this line is extremely sparse.

The analysis for the policy that optimizes fairness while subject to a minimum relevance threshold is similar to the analysis above; thus, we eliminate the demonstrations here due to space limitation.



(a) $D(0.7, 0.1)$ **(b)** $D(0.7, 0.3)$

Figure 4: Exploration space (shaded area) of the optimization policy that optimizes relevance subject to a minimum fairness (entropy) threshold $T = 0.8$, on synthetic dataset.



(a) ***YOW*** **RSS feeds** **(b) Synthetic data** $D(0.7, 0.3)$

Figure 5: Exploration space (a single line) $y = \frac{w_g}{w_r}x$ of the optimization policy that jointly optimizes relevance and fairness, with different weights on each dimension.

Second, we consider the optimization policy that jointly optimizes for both dimensions. We can use weights to denote the trade-off between the two dimensions. From Equation 10, we can draw a line $y = \frac{w_r}{w_g}x$ in the solution space, as shown in Figure 5. We plot $y = 2x$, $y = x$, $y = x/2$ as examples. $y = x$ is a special case where we put equal weights on relevance and fairness. This is also a guideline for quick view of the trade-offs between relevance and fairness which can help decide which dimension to put more weights on according to the specific application need. We see that for the synthetic data $D(0.7, 0.3)$, this policy can lead to a high relevance and high entropy solution set. Yet for the *YOW* RSS feeds data, this policy probably finds nothing because the entropy concentrates around 1 but relevance concentrates around 0.5.

In this case, one may wish to adopt the first policy which is to set the minimum threshold for fairness and optimize for relevance.

To sum up, if the solution space is concentrated around high relevance and high fairness, all three types of policy are suitable to find a solution set that is optimal for both dimensions. If the majority solution sets are of high relevance yet low fairness, it is better to select the policy that optimizes for fairness while subject to a high relevance constraint. Similarly, if the majority solutions are of high fairness yet low relevance, it is better to opt for optimizing for relevance while setting a high fairness constraint. In the case where solution space is concentrated around low relevance and low fairness, the policy that jointly optimizes for both dimensions seems more practical to find a reasonably good relevance and fairness solution set.

4. What does it mean if we cannot achieve optimal fairness on the given dataset? What are the implications of the solution space?

If the solution space shows that with high probability, the solution set with optimal fairness does not exist, then one may need to consider the bias in the data collection and retrieval algorithms. From the perspective of data, if the data is highly unbalanced for different groups, or is too biased according to other fairness definitions, then no matter what algorithms we use, we are unlikely to surface enough items from the minority for a fair exposure. From the perspective of system design and algorithms, the core retrieval framework depends on many basic system components such as data collection, pre-processing, indexing, searching, ranking, and personalization. As a result, any steps that fail to consider the minority groups (e.g., dialects compared to the standard language) and subsequently fail to capture enough representations from the minority, will lead to the biased search space, which may contribute to the low possibility of achieving optimal fairness.

6 CONCLUSION

In this paper, we proposed a novel perspective of analyzing the fairness problems in IR. We presented a framework that first depicts the solution space on any given dataset by estimating theoretical boundaries and optimal solution values, and then we utilized the solution space to facilitate a variety of analysis and decision making which would otherwise be considerably time and resource consuming. This framework has the advantage of being simple to deploy, explain, is reliable with theoretical guarantees, and it is easy to generalize to account for various applications. Researchers and system designers can plug into this framework customized utility functions and fairness constraints, and apply to data of different distributions. We demonstrated the application of our framework with synthetic and real world data. We hope that through our exploration of examples of research questions that can be answered with this framework, we can inspire a broad range of analysis that could potentially benefit from this framework.

The theoretical results presented in this paper do call for some diligence while putting them in practice. The problem setting requires some assumptions that may not always hold. For example, the independence assumption between multi-parties or between relevance and fairness may not be valid for some real world situations. To address this, one can first perform the correlation analysis between different components and then depict the solution boundary by analyzing the characteristics of the joint distribution. We emphasize the contribution of this framework concept and leave more inclusive theoretical analysis such as multinomial distributions and multi-dimension optimizations for future work.

REFERENCES

[1] Mahmoudreza Babaei, Abhijnan Chakraborty, Juhi Kulshrestha, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. 2018. Analysing Biases in Perception of Truth in News Stories and their Implications for Fact Checking. In *Proceedings of FAT'18*.

[2] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results – A user study. *JASIST* 60, 1 (2009), 135–149.

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[5] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proceedings of SIGIR'18*.

[6] Ivar Bråten, Helge I Strømsø, and Ladislao Salmerón. 2011. Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction* 21, 2 (2011), 180–192.

[7] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K Vishnoi. 2018. An Algorithmic Framework to Control Bias in Bandit-based Personalization. *arXiv preprint arXiv:1802.08674* (2018).

[8] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *ICALP'17*.

[9] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of CHI'18*.

[10] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne. *SIGIR Forum* 52, 1 (2018), 34–90.

[11] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018).

[12] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS* 112, 33 (2015), E4512–E4521.

[13] James Grimmelmann. 2011. Some skepticism about search neutrality. *The Next Digital Decade: Essays on the Future of the Internet* (Jan 2011).

[14] Alexander Haas and Julian Unkel. 2017. Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology* 36, 12 (2017), 1285–1298.

[15] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Proceedings of NeurIPS'16*.

[16] Yvonne Kammerer and Peter Gerjets. 2012. Chapter 10 How Search Engine Users Evaluate and Select Web Search Results: The Impact of the Search Engine Interface on Credibility Assessments. In *Web search engine research*. Emerald Group Publishing Limited, 251–279.

[17] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of CHI'15*.

[18] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. 2017. Meritocratic fairness for cross-population selection. In *Proceedings of ICML'17*.

[19] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of CSCW'17*.

[20] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of WWW'17*.

[21] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of CIKM'18*.

[22] Arvind Narayanan. 2018. 21 fairness definitions and their politics. Retrieved April 26, 2019, from https://www.youtube.com/watch?v=jIXIuYdnyyk.

[23] Alamir Novin and Eric Meyers. 2017. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of CHIIR'17*.

[24] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *SSRN Electron. J.* (2016), 1–47.

[25] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of CHI'17*.

[26] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *Proceedings of SIGIR'18*.

[27] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and adjustment in relevance estimation. In *Proceedings of SIGIR'15*.

[28] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of SIGKDD'18*.

[29] Jackie Snow. 2018. Bias already exists in search engine results, and it's only going to get worse. https://www.technologyreview.com/s/610275/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/.

[30] Herman Tavani. 2016. Search Engines and Ethics. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

[31] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of CIKM'17*.

[32] Yi Zhang. 2005. Bayesian Graphical Models for Adaptive Information Filtering. https://users.soe.ucsc.edu/~yiz/papers/data/YOWStudy/