Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems

Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, Alex Beutel
Google Brain
{xuezhiw,nthain,aradhanas,fprost,edchi,jilinc,alexbeutel}@google.com

ABSTRACT

How can we build recommender systems to take into account fairness? Real-world recommender systems are often composed of multiple models, built by multiple teams. However, most research on fairness focuses on improving fairness in a single model. Further, recent research on classification fairness has shown that combining multiple "fair" classifiers can still result in an "unfair" classification system. This presents a significant challenge: how do we understand and improve fairness in recommender systems composed of multiple components?

In this paper, we study the compositionality of recommender fairness. We consider two recently proposed fairness ranking metrics: equality of exposure and pairwise ranking accuracy. While we show that fairness in recommendation *is not* guaranteed to compose, we provide theory for a set of conditions under which fairness of individual models *does* compose. We then present an analytical framework for both understanding whether a real system's signals can achieve compositional fairness, and improving which component would have the greatest impact on the fairness of the overall system. In addition to the theoretical results, we find on multiple datasets—including a large-scale real-world recommender system—that the overall system's end-to-end fairness is largely achievable by improving fairness in individual components.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Content ranking**; **Social recommendation**.

KEYWORDS

compositional fairness; recommender systems; ranking fairness

ACM Reference Format:
Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, Alex Beutel. 2021. Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441732

This work is licensed under a Creative Commons Attribution International 4.0 License.

WSDM '21, March 8–12, 2021, Virtual Event, Israel.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441732

1 INTRODUCTION

As recommender systems become more central in our lives, the importance of their fairness has become increasingly clear. Over the past few years, the research community has developed a variety of metrics for evaluating the impact of recommender systems on different groups of stakeholders, such as if relevant items get equal exposure, if demographic groups of item producers rank well conditioned on user interest, and if items from a diverse set of item producers or topics are shown [6, 9, 13, 17, 22, 47, 52, 56]. These are important values to consider in the *responsible* design of a recommender system, but how to build recommender systems to meet these goals is still difficult.

One fundamental challenge is that real-world recommendation products are often composed of many models, designed to capture different aspects of the user experience and sometimes even built and maintained by distinct teams [1, 12, 32, 51]. For example, one team may be responsible for predicting implicit feedback signals like clicks [32], another may be responsible for predicting explicit feedback like ratings or surveys [4, 5, 27, 53], and another may be responsible for predicting notions of item quality [35, 44]; these different predictions are typically then combined, often multiplicatively, to produce a final ranking [18, 21, 37, 58].

This multi-component design pattern poses a challenge for recommendation fairness: how should each team build their models to make the end-ranking seen by the user meet fairness goals? Nearly all of the research on training-time improvements for fairness has assumed that the "system" that we are improving is a single differentiable model [2, 6, 7, 42, 48, 55][1]. As a result, an obvious guess is to train each model in the recommender system to itself meet a ranking fairness goal. Unfortunately, recent research on fairness in classification showed that even if two classifiers are "fair," a combination of their predictions can still be "unfair" [23, 24]. Known as the compositional fairness problem, this presents a significant open question for recommenders: does recommendation fairness compose in our modular recommender systems?

In this paper, we study the compositional recommender fairness problem both from a worst-case lens and a data-driven average-case lens. First, we demonstrate that recommendation fairness, unfortunately, *is not guaranteed* to compose, which suggests a significant obstacle for building real-world recommender systems for fairness.

A motivating example. To make the problem concrete, consider the hypothetical example of a recommendation system for books (e.g., [26]) that wants to rank items by expected user satisfaction. It can be built with two components: (1) $pCTR = P(\text{click})$: one that predicts click-through rate on a book; (2) $pRating = E[\text{rating}|\text{click}]$: one that predicts the star rating given a book was clicked, akin to

[1]Note, we focus on training-time improvements because serving-time changes typically require knowing demographics at serving-time, which is often not possible [7].

[37]. With these components, items can then be ranked by:

$$E[\text{rating}] = P(\text{click}) \times E[\text{rating}|\text{click}] = pCTR \times pRating.$$

Let the fairness goal be demographic parity in *ranking exposure* [47]. That is, the ranking of the composite score should not systematically differ between *white* and *non-white* authors. Each component could be made "fair" with respect to author demographics through recent mitigation methods [6, 42, 48]. What does this mean for the demographic parity on the ranked composite scores? It may feel intuitive to assume that if each component gives equal exposure to each group, the overall system should as well. Here we show through an example that this is not the case. Assume we have 4 books with author demographics in the header of each column:

Component	non-white	non-white	white	white
$pCTR$	0.1	0.4	0.2	0.3
$pRating$	0.4	0.1	0.3	0.2
$pCTR \times pRating$	0.04	0.04	0.06	0.06

Each component exposes books from each group equally: if we rank the items by either $pCTR$ or $pRating$ individually, we get [non-white, white, white, non-white]. However, when the two components are multiplied together, the composite ranking ([white, white, non-white, non-white]) does not result in demographic parity[2]. In Section 3, we provide a broader theoretical analysis. We study two recently proposed ranking fairness metrics [6, 9, 22, 47], each capturing slightly different goals. For both, we consider the broader class of multiplicative composition, as it is common in production recommender systems [37, 58], and without loss of generality extends to additive composition by moving to the log-space, as is also common in practice [18, 21].

In response to these bleak results, we observe that these are worst-case guarantees, and thus flip to the constructive question: *when does* recommender fairness compose? What can we do in a real recommender system? We explore this both theoretically and empirically. We first prove a set of conditions about the models' predictions under which recommender fairness would compose across components. With this theoretical understanding, we develop an analytical framework for counterfactually testing a multi-model recommender system. That is, given an existing recommender system, we provide a method for measuring how well the system would meet a given recommender fairness goal if each model in the system was (independently) made fair under that metric. We find across multiple applications, including a real-world recommender, that these fairness goals can largely be met through this per-component approach, both through our analytical framework and then through modeling experiments. Further, we demonstrate that counterfactual testing enabled by the analytical framework can be used to uncover which components are most limiting fairness metrics, which could then be used to prioritize modeling improvements. Taken together, we believe this suggests a path toward improving fairness in recommender systems. In summary, our contributions in this paper are:

Theory: We provide theory showing a set of conditions under which fair components can compose into fair systems.

[2]One might ask: why not just rank by *either* pCTR or pRating? Unfortunately, doing that would result in either increased clickbait (if only focusing on pCTR) or unappealing-looking items (if ignoring pCTR). This is why recommender systems incorporate multiple signals and limits the flexibility of the composition function.

System Understanding: We provide an analytical framework of counterfactual testing to measure whether a system's signals can achieve compositional fairness and to diagnose which of these signals lowers the overall system fairness most.

Experiments: Although compositional fairness is theoretically not guaranteed, on multiple data-sets, including a large-scale production recommender system, we apply our counterfactual testing and find that the overall system fairness is largely achievable by improving fairness in individual components. We subsequently test a training-time regularization and find that while no individual component can be improved to make the system "fair," applying the approach to all of the components (separately) is highly effective.

2 RELATED WORK

Fairness in Classification: The majority of the fairness metric definition literature focuses on classification. *Demographic parity* [14, 50, 54] is a common way of addressing discrimination against protected attributes. It requires a decision to be independent of the protected attribute. *Equalized odds* [31] require a predictor $\hat{Y}$ to be independent of the protected attribute, conditioned on the true label Y. Metrics have also been explored over continuous scores from a classifier: [11, 33, 42] all break down AUC of these scores into Mann-Whitney U-tests [39].

Fairness in Ranking: Recently, a few definitions have been proposed for fairness in the *ranking* setting [56]; in our work we focus on two recent framings. Singh and Joachims [47] propose measuring the exposure an item or group of items gets depending on what position they fall in a ranking. (Similar metrics were proposed by Biega et al. [9], and further studied by Diaz et al. [22].) The work offers multiple fairness goals, such as exposure proportional to relevance, but in our usage we build on this notion of exposure to measure group representation throughout a ranked list (as we ignore any label or relevance, this is philosophically closer to the principles of demographic parity above). Beutel et al. [6] focus on measuring accuracy in a recommender system based on *pairwise comparisons*. The accuracy of a ranking for a pair of items is defined as the probability that the clicked item is ranked higher than the un-clicked item. In this set-up, two items from different groups are used to create a pair, and the difference in accuracy for each group is used as a fairness metric. We use this metric to capture fairness in ranking more closely aligned with equal opportunity (as it is measuring accuracy with respect to a label—clicks). In Section 3, we will formalize the above two definitions within the same framework. For each of the ranking metrics, given per-component fairness, we will show conditions where compositional fairness holds (and counter-examples where it might not hold).

A closely related area to ranking fairness is ranking diversification [3, 15, 16, 29, 46, 49], where the goal is to diversify the ranking results to improve user satisfaction. In many cases, general-purpose diversification does not align with fairness for certain sub-groups.

Compositional Fairness: Dwork and Ilvento [23] studied general constructions for fair composition, and showed that classifiers that are fair in isolation do not necessarily compose into fair systems, for individuals or for groups. The authors studied the "Functional Composition" setting, where the assumption is that the binary outputs of multiple classifiers are combined through

logical operations to produce a single output for a single task. More recent work by Dwork et al. [25] explores composition of individual fairness in pipelines. In this paper, we focus on fairness in recommender systems and a general multiplicative compositional setting (or equivalently additive composition in the log-space), as is common in many real ranking systems [21, 37, 58].

Mitigation: Many mitigation approaches try to achieve fairness in the *single-task* (non-compositional) setting. The approaches can be partitioned by when they intervene in the creation of an ML system: pre-processing (e.g., [10]), post-processing (e.g., [28, 34, 43, 47]), and training, including adding constraints [2, 19, 20], regularization [7, 55], and adversarial learning [8, 36, 38, 57]. Post-processing approaches do not suffer from compositional challenges, but are often not feasible in practice due to not having demographic information at inference time. The regularization approaches are most similar to our analysis, encouraging matching the distribution of predictions [7, 55] or representation [8, 38] across groups.

3 METRICS AND THEORETICAL ANALYSIS

In this section, we formally define the problem: denote x as the item being ranked, and there are two groups of users being considered, Group $\mathcal{A}$ and Group $\mathcal{B}$. Assume a recommender system with K components, where the k-th component takes an input x and produces its own score $f_k(x)$; the broader class of multiplicative composition ([37, 58]) can be formulated as: $f(x) = \prod_{k=0}^{K-1} f_k(x)$. If all K components satisfy some fairness metric by themselves $F[f_k]$, we ask whether the overall function $f(x)$ satisfies the same fairness metric $F[f]$. Restated, we would like to know whether the system achieves compositional (end-to-end) fairness given that each component has achieved fairness independently.

In the sections below, we consider two commonly-used fairness metrics in recommender systems: *ranking exposure* [47], and *pairwise ranking accuracy* [6]. We describe each metric, and explore how the function composition affects end-to-end fairness.

3.1 Fairness in Ranking Exposure

3.1.1 Definition. We start with the ranking exposure [47] fairness metric. Formally, the ranking exposure for Group $\mathcal{A}$ is defined as:

$$\text{Exposure}(\mathcal{A}|r) = \sum\nolimits_{x\in\mathcal{A}} u(x|r),$$

under a ranking order r. Here u denotes the utility function, which is usually a monotonically decreasing function with respect to the rank of an item. One common choice is to use an exponent $w >= 0$, and $u(x|r) = [\text{rank}(x|r)]^{-w}$. The fairness exposure metric between Group $\mathcal{A}$ and Group $\mathcal{B}$ under a ranking order r is defined as:

$$\text{Gap}_{\text{exp}}(\mathcal{A},\mathcal{B}|r) = \frac{|\text{Exposure}(\mathcal{A}|r) - \text{Exposure}(\mathcal{B}|r)|}{\text{Exposure}(\mathcal{A}|r) + \text{Exposure}(\mathcal{B}|r)}, \quad (1)$$

which denotes the normalized gap between the exposure for Group $\mathcal{A}$ and $\mathcal{B}$. Note, when the two groups have the same size, i.e., $|\mathcal{A}| = |\mathcal{B}|$, the ideal exposure gap should reach zero in order to reach demographic parity [14, 50, 54]. This might not be the case when the two groups have different sizes.

Intuitively, this metric (here unconditioned on relevance) makes the goal to provide a diverse ranking with each group being well represented throughout the ranked list. While we build on [47] for framing, similar intuitions were also proposed in [9, 56] and used in job search applications [28].

3.1.2 Counter-example of Composition. We offer a counter-example and show that under the fair exposure metric, per-component fairness does not always guarantee compositional fairness. Consider a ranking system composed of two components f_0, f_1, two groups $\mathcal{A}, \mathcal{B}$, and each group has two items. For any $a > 0, \epsilon > 0$, suppose:

Component	$x_1^a \in \mathcal{A}$	$x_2^a \in \mathcal{A}$	$x_1^b \in \mathcal{B}$	$x_2^b \in \mathcal{B}$
$f_0(x)$	$a+\epsilon$	$a+4\epsilon$	$a+2\epsilon$	$a+3\epsilon$
$f_1(x)$	$a+4\epsilon$	$a+\epsilon$	$a+3\epsilon$	$a+2\epsilon$
Exposure of $f_0(x)$	e_1	e_0	e_1	e_0
Exposure of $f_1(x)$	e_0	e_1	e_0	e_1
Exposure of $f_0(x)\cdot f_1(x)$	e_1	e_1	e_0	e_0

Here we assume the first two positions receive a higher exposure e_0 than the exposure e_1 for the last two positions (i.e., $1 \geq e_0 > e_1 \geq 0$), and it is easy to see that each component by itself is fair, since $\text{Exposure}(\mathcal{A}|r_k) = \text{Exposure}(\mathcal{B}|r_k) = e_0 + e_1,\ k \in \{0,1\}$, r_k being the ranking order determined by $f_k(x)$. But when combined by $f(x) = f_0(x)\cdot f_1(x)$, Group $\mathcal{A}$ is always ranked below Group $\mathcal{B}$. Specifically, $\text{Gap}_{\text{exp}}(\mathcal{A},\mathcal{B}|r) = \frac{|e_0-e_1|}{e_0+e_1}$, and could be as large as 1 if $e_0 = 1, e_1 = 0$. We can also see that the magnitude of the scores does not play a role here, since we can make ϵ arbitrarily small to make the magnitude of the score for the two components arbitrarily close to each other ($\to a$), while keeping $\text{Gap}_{\text{exp}}(\mathcal{A},\mathcal{B}|r) = 1$.

3.1.3 Condition for composition of ranking exposure. Now we present theory showing under what conditions we will achieve compositional fairness given we have per-component fairness, under the ranking exposure (§3.1.1) metric.

Consider a system with two components f_0 and f_1, and two groups $\mathcal{A}, \mathcal{B}$. Let X_A^0 represent the random variable defined by $\log f_0(x), x \in \mathcal{A}$, and X_B^0 for $\log f_0(x), x \in \mathcal{B}$. Similarly we define X_A^1, X_B^1 for f_1. Assume the top half and the bottom half of the items in the ranked list receive different exposure values $u(x|r)$ by sorting f_0 and f_1 in descending order, respectively. Assume we have per-component fairness in the log-space, i.e., $\text{median}[X_A^0] = \text{median}[X_B^0]$ for $\log f_0(x)$ and $\text{median}[X_A^1] = \text{median}[X_B^1]$ for $\log f_1(x)$. We have the following theorem:

THEOREM 1. *If $X_A^0, X_B^0, X_A^1, X_B^1$ are symmetric random variables such that $X_A^0+X_A^1$ and $X_B^0+X_B^1$ are also symmetric, then per-component fairness on $\log f_0$ and $\log f_1$ means we have compositional fairness for $\log f(x)$, where $f(x) = f_0(x)\cdot f_1(x)$.*

PROOF. By composing f_0 and f_1 we have: $\text{median}_{x\in\mathcal{A}}[\log(f_0(x)\cdot f_1(x))] = \text{median}_{x\in\mathcal{A}}[\log f_0(x) + \log f_1(x)] = \text{median}[X_A^0 + X_A^1]$, and since:

$\text{median}[X_A^0 + X_A^1] = \text{mean}[X_A^0 + X_A^1]$ (by symmetry of $X_A^0 + X_A^1$)

$=\text{mean}[X_A^0] + \text{mean}[X_A^1]$ (by linearity of expectation)

$=\text{median}[X_A^0] + \text{median}[X_A^1]$ (by symmetry of X_A^0 and X_A^1)

$=\text{median}[X_B^0] + \text{median}[X_B^1]$ (by per-component fairness)

$=\text{mean}[X_B^0] + \text{mean}[X_B^1]$ (by symmetry of X_B^0 and X_B^1)

$=\text{mean}[X_B^0 + X_B^1]$ (by linearity of expectation)

$=\text{median}[X_B^0 + X_B^1]$, (by symmetry of $X_B^0 + X_B^1$)

which equals to $\text{median}_{x\in\mathcal{B}}[\log f_0(x) + \log f_1(x)]$ and thus $\text{median}_{x\in\mathcal{B}}[\log(f_0(x)\cdot f_1(x))]$. □

To empirically test whether a random variable X is symmetric around a given mean θ, one could construct samples $(X, 2\theta - X)$ and test whether the two distributions are the same by a two-sample Kolmogorov–Smirnov test or a kernel two-sample test [30]. Further, for compositional fairness in the entire system, we need to convert $\log f_0, \log f_1, \log f$ back to the original space f_0, f_1, f. The median element remains unchanged by the monotonicity of the log function when there is an odd number of samples in each group, in which case we will have $\text{median}_{x\in\mathcal{A}}[f_0(x)\cdot f_1(x)] = \text{median}_{x\in\mathcal{B}}[f_0(x)\cdot f_1(x)]$. When there is an even number of samples per group, then we also need the median in the original space to be the same across groups in order for compositional fairness to hold.

3.2 Fairness as Pairwise Ranking Accuracy

3.2.1 Definition. Another commonly used fairness metric in ranking is *pairwise ranking accuracy* [6], where the idea is to compute the accuracy of a system ranking a pair of items correctly conditioned on the true outcome. The pair of items is constrained to come from two different groups, $\mathcal{A}$ and $\mathcal{B}$, through randomized experiments. Formally, the metric is defined as:

$$\begin{aligned}\text{PairAcc}(\mathcal{A} > \mathcal{B}|r) &= P(f(x_i) > f(x_j)|y_i > y_j, x_i \in \mathcal{A}, x_j \in \mathcal{B}) \\ &= \frac{P(f(x_i) > f(x_j), y_i > y_j|x_i \in \mathcal{A}, x_j \in \mathcal{B})}{P(y_i > y_j|x_i \in \mathcal{A}, x_j \in \mathcal{B})}.\end{aligned}$$

Here $y_i \in \{0, 1\}$ denotes the observed binary outcome for x_i (e.g., $y_i = 1$ means a recommended item is clicked, or a recommended product is purchased). Correspondingly, the Pairwise Ranking Accuracy Gap is defined as:

$$\text{Gap}_{\text{pair}}(\mathcal{A}, \mathcal{B}|r) = |\text{PairAcc}(\mathcal{A} > \mathcal{B}|r) - \text{PairAcc}(\mathcal{A} < \mathcal{B}|r)| \quad (2)$$

Remark. *Intuitively, this metric means that given a pair of items, one from group $\mathcal{A}$ and one from group $\mathcal{B}$, conditioned on one with $y_i = 1$ and the other with $y_i = 0$, we would like the system to have the same accuracy of ranking this pair of items correctly, regardless of which group the item with positive outcome ($y_i = 1$) is from.*

Further, let $X^k_{A_0}$ represent the random variable defined by $f_k(x_i)$ for $y_i = 0$ and $x_i \in \mathcal{A}$; and let $X^k_{A_1}$ represent the random variable defined by $f_k(x_i)$ for $y_i = 1$ and $x_i \in \mathcal{A}$. X_{B_0}, X_{B_1} are defined similarly. We can simply write the Pairwise Ranking Accuracy Gap metric as: $\text{Gap}_{\text{pair}}(\mathcal{A}, \mathcal{B}|r) = |P(X_{A_1} > X_{B_0}) - P(X_{B_1} > X_{A_0})|$.

3.2.2 Counter-example of composition. In the following we present a simple example that shows per-component fairness might not lead to compositional fairness, under the pairwise ranking accuracy gap metric. Consider the following system with two components, two groups, and two corresponding items within each group:

Component	$X^k_{A_1}$	$X^k_{B_0}$	$X^k_{B_1}$	$X^k_{A_0}$
$f_0(x)$	$\{1, 4\}$	$\{2, 3\}$	$\{1, 4\}$	$\{2, 3\}$
$f_1(x)$	$\{4, 1\}$	$\{3, 2\}$	$\{1, 4\}$	$\{3, 2\}$
PairAcc of $f_0(x)$	0.5		0.5	
PairAcc of $f_1(x)$	0.5		0.5	
$f_0(x)\cdot f_1(x)$	$\{4, 4\}$	$\{6, 6\}$	$\{1, 16\}$	$\{6, 6\}$
PairAcc of $f_0(x)\cdot f_1(x)$	0.0		0.5	

It is easy to see that each component is fair. When composed, we have $\text{PairAcc}(\mathcal{A} > \mathcal{B}|r) = 0.0$, because both items with $y_i = 1$ from $\mathcal{A}$ receive a lower prediction score $\{4, 4\}$ than items with $y_i = 0$ from $\mathcal{B}$: $\{6, 6\}$. Similarly we have $\text{PairAcc}(\mathcal{A} < \mathcal{B}|r) = 0.5$ and hence $\text{Gap}_{\text{pair}}(\mathcal{A}, \mathcal{B}|r) = 0.5$. In other words, the predictor f does not have equal treatment for ranking the items from $\mathcal{A}$ and $\mathcal{B}$.

3.2.3 Condition for composition of pairwise ranking accuracy. We provide a theorem showing under which conditions compositional fairness holds under the pairwise ranking accuracy gap metric (Eq. (2)). We denote $\mathcal{A}_0$ and $\mathcal{A}_1$ to be the set of items from $\mathcal{A}$ with $y = 0$ and $y = 1$, respectively; we similarly define $\mathcal{B}_0$ and $\mathcal{B}_1$. We assume equal size on the groups, i.e., $|\mathcal{A}_\ell| = |\mathcal{B}_\ell|, \ell = \{0, 1\}$.[3] We define a delta term between all pairs from $\mathcal{A}_1$ and $\mathcal{B}_0$ and similarly between all pairs from $\mathcal{B}_1$ and $\mathcal{A}_0$, i.e.,

$$\begin{aligned}\Delta_k(\mathcal{A}_1, \mathcal{B}_0) &= \{f_k(x_i) - f_k(x_j) \mid x_i \in \mathcal{A}_1, x_j \in \mathcal{B}_0\}; \\ \Delta_k(\mathcal{B}_1, \mathcal{A}_0) &= \{f_k(x_i) - f_k(x_j) \mid x_i \in \mathcal{B}_1, x_j \in \mathcal{A}_0\}. \quad (3)\end{aligned}$$

Let $Z^0_{A_1,B_0}, Z^0_{A_0,B_1}$ represent the random variables defined by $\Delta_0(\mathcal{A}_1, \mathcal{B}_0), \Delta_0(\mathcal{B}_1, \mathcal{A}_0)$, respectively, for f_0. Similarly we define $Z^1_{A_1,B_0}, Z^1_{A_0,B_1}$ for f_1. Without loss of generality assume all component scores are positive, i.e., $X^0_{A_1} > 0, X^1_{B_0} > 0, X^0_{B_1} > 0, X^1_{A_0} > 0$ (if not we can shift the scores without changing the overall ranking). We have the following theorem for compositional fairness:

Theorem 2. *If the following hold:*

$$(X^0_{A_1} Z^1_{A_1,B_0} + X^1_{B_0} Z^0_{A_1,B_0})^{(+)} = (X^0_{A_1} Z^1_{A_1,B_0})^{(+)} + (X^1_{B_0} Z^0_{A_1,B_0})^{(+)}, \quad (4)$$

$$(X^0_{B_1} Z^1_{A_0,B_1} + X^1_{A_0} Z^0_{A_0,B_1})^{(+)} = (X^0_{B_1} Z^1_{A_0,B_1})^{(+)} + (X^1_{A_0} Z^0_{A_0,B_1})^{(+)}, \quad (5)$$

where $Z^{(+)} = P(Z > 0)$, then per-component fairness on f_0 and f_1 means we have compositional fairness for $f(x) = f_0(x)\cdot f_1(x)$.

Proof. With per-component fairness we have:

$$\text{for } f_0, P(X^0_{A_1} > X^0_{B_0}) = P(X^0_{B_1} > X^0_{A_0}), \text{ or } (Z^0_{A_1,B_0})^{(+)} = (Z^0_{A_0,B_1})^{(+)},$$
$$\text{for } f_1, P(X^1_{A_1} > X^1_{B_0}) = P(X^1_{B_1} > X^1_{A_0}), \text{ or } (Z^1_{A_1,B_0})^{(+)} = (Z^1_{A_0,B_1})^{(+)},$$

by composing f_0 and f_1 we have:

$$\begin{aligned}&(X^0_{A_1}\cdot X^1_{A_1} - X^0_{B_0}\cdot X^1_{B_0})^{(+)} \\ =&(X^0_{A_1}\cdot X^1_{A_1} - X^0_{A_1}\cdot X^1_{B_0} + X^0_{A_1}\cdot X^1_{B_0} - X^0_{B_0}\cdot X^1_{B_0})^{(+)} \\ =&(X^0_{A_1}\cdot Z^1_{A_1,B_0} + X^1_{B_0}\cdot Z^0_{A_1,B_0})^{(+)} \\ =&(X^0_{A_1}\cdot Z^1_{A_1,B_0})^{(+)} + (X^1_{B_0}\cdot Z^0_{A_1,B_0})^{(+)} \quad \text{(by Eq. (4))} \\ =&(X^0_{A_1}\cdot Z^1_{A_0,B_1})^{(+)} + (X^1_{B_0}\cdot Z^0_{A_0,B_1})^{(+)} \quad \text{(by per-component fairness)} \\ =&(Z^1_{A_0,B_1})^{(+)} + (Z^0_{A_0,B_1})^{(+)} \quad \text{(by } X^0_{A_1} > 0, X^1_{B_0} > 0) \\ =&(X^0_{B_1}\cdot Z^1_{A_0,B_1})^{(+)} + (X^1_{A_0}\cdot Z^0_{A_0,B_1})^{(+)} \quad \text{(by } X^0_{B_1} > 0, X^1_{A_0} > 0) \\ =&(X^0_{B_1}\cdot Z^1_{A_0,B_1} + X^1_{A_0}\cdot Z^0_{A_0,B_1})^{(+)} \quad \text{(by Eq. (5))} \\ =&(X^0_{B_1}\cdot X^1_{B_1} - X^0_{B_1}\cdot X^1_{A_0} + X^0_{B_1}\cdot X^1_{A_0} - X^0_{A_0}\cdot X^1_{A_0})^{(+)} \\ =&(X^0_{B_1}\cdot X^1_{B_1} - X^0_{A_0}\cdot X^1_{A_0})^{(+)},\end{aligned}$$

i.e., $P(X^0_{A_1}\cdot X^1_{A_1} > X^0_{B_0}\cdot X^1_{B_0}) = P(X^0_{B_1}\cdot X^1_{B_1} > X^0_{A_0}\cdot X^1_{A_0})$, hence compositional fairness holds. □

[3] Note, in order to form pairs, the definition of pairwise ranking accuracy additionally implies $|\mathcal{A}_1| = |\mathcal{B}_0|$ and $|\mathcal{A}_0| = |\mathcal{B}_1|$.

One scenario for Eq. (4) to hold is when there is a perfect match on the signs of $X^0_{A_1} Z^1_{A_1,B_0}$ and $X^1_{B_0} Z^0_{A_1,B_0}$ (similarly for Eq. (5)). Alternatively, if the change of signs by comparing $(X^0_{A_1} Z^1_{A_1,B_0} + X^1_{B_0} Z^0_{A_1,B_0})$ to $(X^0_{A_1} Z^1_{A_1,B_0}, X^1_{B_0} Z^0_{A_1,B_0})$ adds up to zero, Eq. (4) also holds.

4 ANALYTICAL FRAMEWORK AND MODELING SOLUTIONS

As we see in the theoretical results, improving the fairness of individual components *sometimes* benefits the fairness of the composite score, depending on the components and the relationship between them. Therefore we ask: if we have a multi-component system where we observe fairness issues, how much will improving the fairness for each component help the overall system's fairness?

Taking this data-driven view of the problem, we find there are multiple questions that we can tractably answer:

(1) Given a system with fairness issues, improving which components would yield the greatest benefit?
(2) If all components were independently "fixed", what would be the resulting fairness metrics for the combined system?

4.1 Per-Component Fixes

To discover which components would be most beneficial in improving fairness, we explore the impact of two realistic classes of methods.

4.1.1 Distribution matching for ranking exposure. A significant amount of academic literature [30, 40] and publications on what is used practice [6, 7], takes the perspective of regularizing the model such that the distribution of predictions from each group (sometimes conditioned on the label) is matching. Under different formulations this has been done by comparing the covariance [55], correlation [6, 7], and Maximum Mean Discrepancy [30, 41] between the distributions. Therefore, we consider whether matching the groups' distributions of predictions for each model has the desired effect on the combined fairness metrics.

As this is an analytical framework, in contrast to a training framework, we can easily do this offline by directly changing the predictions over our dataset. We consider distribution matching for a component f_k. In order to match the distributions, we sort all examples in each group by their scores f_k. We define by $\mathbf{a}^{(k)}$ a sorted vector of scores for examples in Group $\mathcal{A}$ and by $\phi_{a,k}$ the mapping of examples to positions in this sorted list, i.e. $\mathbf{a}^{(k)}_{\phi_{a,k}(x)} = f_k(x)$ and $\mathbf{a}^{(k)}_j \leq \mathbf{a}^{(k)}_{j+1}$ for all j; we similarly define $\mathbf{b}^{(k)}$ and $\phi_{b,k}$ for examples from Group $\mathcal{B}$. Assuming $|\mathcal{A}| = |\mathcal{B}|$, when matching the distributions, we define the "fixed" component $\tilde{f}_k$ as follows:

$$\tilde{f}_k(x) = f_k(x), \text{ if } x \in \mathcal{A}; \quad \tilde{f}_k(x) = \mathbf{a}^{(k)}_{\phi_{b,k}(x)}, \text{ if } x \in \mathcal{B}. \tag{6}$$

That is, for examples in $\mathcal{B}$, $\tilde{f}_k$ returns the score for a similarly ranked item from $\mathcal{A}$ such that the empirical distribution over $\mathcal{A}$ and $\mathcal{B}$ exactly matches. Note, $\mathbf{a}$ and $\mathbf{b}$ are the empirical cumulative distribution function (CDF) for f_k over $\mathcal{A}$ and $\mathcal{B}$ respectively, and as such if $|\mathcal{A}| \neq |\mathcal{B}|$ then simple interpolation to match the empirical CDFs can be used.

Theorem 3. *$\tilde{f}_k$ as defined by Eq. (6) has an exposure gap (defined by Eq. (1)) of zero, assuming the ranking order r based on $\tilde{f}_k$ gives the exact same rank of x given the same score of $\tilde{f}_k(x)$.*[4]

Proof. Given the definition in Eq. (1), it is easy to see that

$$\begin{aligned}\text{Gap}_{\text{exp}}(\mathcal{A}, \mathcal{B}|r) &= \frac{|\text{Exposure}(\mathcal{A}|r) - \text{Exposure}(\mathcal{B}|r)|}{\text{Exposure}(\mathcal{A}|r) + \text{Exposure}(\mathcal{B}|r)} \\ &= \frac{|\sum_{x\in\mathcal{A}}[\text{rank}(x)]^{-w} - \sum_{x\in\mathcal{B}}[\text{rank}(x)]^{-w}|}{\text{Exposure}(\mathcal{A}|r) + \text{Exposure}(\mathcal{B}|r)} = 0.\end{aligned}$$

Because there is an exact one-to-one correspondence of $\tilde{f}_k(x_i), x_i \in \mathcal{A}$ and $\tilde{f}_k(x_j), x_j \in \mathcal{B}$ that gives exactly one pair of $\text{rank}(x_i) = \text{rank}(x_j)$ from each group, which cancels each other given $|\mathcal{A}| = |\mathcal{B}|$ and thus results in a zero gap. □

4.1.2 Label-conditioned distribution matching for pairwise ranking accuracy. The above approach only recalibrates the predictions by group but does not necessarily align with any labels for the task. As such, for the pairwise ranking accuracy gap metric (Eq. (2)) the method as described is not guaranteed to give per-component pairwise fairness. For that, we provide a slight modification of the algorithm above, i.e., we exactly match the empirical distribution between $\Delta_k(\mathcal{A}_1, \mathcal{B}_0)$ and $\Delta_k(\mathcal{B}_1, \mathcal{A}_0)$ (Eq. (3)), which aligns with the regularization proposed in [6]. In the following we show this label-conditioned distribution matching suffices to achieve pairwise ranking fairness for each component k.

Theorem 4. *$\hat{f}_k$ as defined by matching $\Delta_k(\mathcal{A}_1, \mathcal{B}_0)$ and $\Delta_k(\mathcal{B}_1, \mathcal{A}_0)$ has a pairwise ranking accuracy gap, defined in Eq. (2) of 0.*

Proof. The Pairwise Ranking Accuracy $\text{PairAcc}(\mathcal{A} > \mathcal{B}|r)$ is given by $\frac{\sum_{x_i\in\mathcal{A}_1, x_j\in\mathcal{B}_0} I[\hat{f}_k(x_i) > \hat{f}_k(x_j)]}{|\mathcal{A}_1|\cdot|\mathcal{B}_0|}$, i.e., it is equal to the percentage of positive deltas in $\hat{\Delta}_k(\mathcal{A}_1, \mathcal{B}_0) = \{\hat{f}_k(x_i) - \hat{f}_k(x_j) \mid x_i \in \mathcal{A}_1, x_j \in \mathcal{B}_0\}$ by definition. Similarly, $\text{PairAcc}(\mathcal{B} > \mathcal{A}|r)$ is equal to the percentage of positive deltas in $\hat{\Delta}_k(\mathcal{B}_1, \mathcal{A}_0)$.

Given that we exactly matched the delta terms in $\hat{\Delta}_k(\mathcal{A}_1, \mathcal{B}_0)$ and $\hat{\Delta}_k(\mathcal{A}_1, \mathcal{B}_0)$, and since $|\mathcal{A}_0| = |\mathcal{A}_1| = |\mathcal{B}_0| = |\mathcal{B}_1|$, we have $\text{PairAcc}(\mathcal{A} > \mathcal{B}|r) = \text{PairAcc}(\mathcal{B} > \mathcal{A}|r)$, i.e., the pairwise ranking accuracy gap (as defined in Eq. (2)) is 0. □

4.1.3 Distribution Normalization. While the above procedure is provably guaranteed to achieve per-component fairness, under the definitions given previously, in practice we would want to use a regularization on the model for this goal, which will be noisier. As such, we consider a lighter-weight approach: per-group normalization:

Definition 1. *Per-Group Normalization. We modify component f_k to incorporate per-group normalization by:*

$$\bar{f}_k(x) = \frac{f_k(x) - \mu_{x\in\mathcal{G}}[f_k(x)]}{\sigma_{x\in\mathcal{G}}[f_k(x)]}, \quad \mathcal{G} \in \{\mathcal{A}, \mathcal{B}\}, \tag{7}$$

where μ, σ is the empirical mean and standard deviation on $f_k(x)$, for $x \in \mathcal{A}$ or $x \in \mathcal{B}$, respectively. While $\bar{f}$ is not guaranteed to provide even per-component fairness under either definition, we find in practice it too can significantly improve end-to-end fairness.

[4] Note in real applications this is usually not the case since a tie-breaking strategy is needed for two items with the same score. Assuming the tie-breaking strategy is random, $\tilde{f}_k$ should achieve an exposure gap close to zero.

4.2 Counterfactual Testing

How can we use the modified functions described above to understand the system's end-to-end fairness properties? All of the questions given at the beginning of this section are counterfactual questions: what would happen if we succeeded in fixing a component or set of components? With the above methods for simulating a fixed component (without actually changing the model training), we can do this headroom analysis.

Per-Component Effect. As before, we assume we have K components which are multiplied together such that the overall score given to an example by the system is $f(x) = \prod_{k=0}^{K-1} f_k(x)$. Even when improving the fairness of one component, it is not guaranteed to improve the fairness of the overall system. For example, two components could be equally biased in opposite directions such that improving only one actually worsens the end-to-end fairness metrics.

Therefore, we use the above per-component modifications to test the effect of independently improving individual components. We will use g_κ to characterize a modified component as described above, i.e., $g_\kappa \in \{\tilde{f}_\kappa, \hat{f}_\kappa, \bar{f}_\kappa\}$. With this we can simulate how the system would behave if we improve a given component κ:

Definition 2 (κ-Improved System). *Given a system f with K components f_k, and a simulated improved component g_κ for component κ, we define the improved system as:*

$$f^{(\bar{\kappa})}(x) = g_\kappa(x) \prod_{k \neq \kappa} f_k(x) = f(x) \frac{g_\kappa(x)}{f_\kappa(x)}. \tag{8}$$

With this we can understand the fairness of this counterfactual system $f^{(\bar{\kappa})}$, using either Eq. (1) or (2): if we improved the fairness of component κ, what would be the resulting end-to-end fairness?

4.3 Modeling Solution

The above analytical framework is a low-cost way to test if per-component fix leads to compositional fairness before we implement any real algorithmic changes. In practice, after we have identified that a system's overall fairness can be improved by per-component fixes, as well as which components should be prioritized to fix, we need an algorithmic change over the existing per-component model. Within each component, it is easy to see that the problem is a standard single-task optimization problem, where the objective can in general be written as $L_{ranking} + \lambda \cdot L_{fairness}$, where $L_{ranking}$ is the original ranking loss (e.g., loss over CTR predictions), and $L_{fairness}$ is a fairness loss which has many instantiations in existing works [7, 45, 55]. In the experiment section we will explore modeling solutions and show that the results are consistent with the solutions we provided through our analytical framework.

5 EXPERIMENTS

5.1 Synthetic Data

We begin with presenting experiments on synthetic datasets to demonstrate the relationship between per-component fairness and compositional fairness, connecting to our theoretical analysis. Again we assume the system has two components f_0 and f_1, and we evaluate the fairness metrics with respect to two groups $\mathcal{A}$ and $\mathcal{B}$.

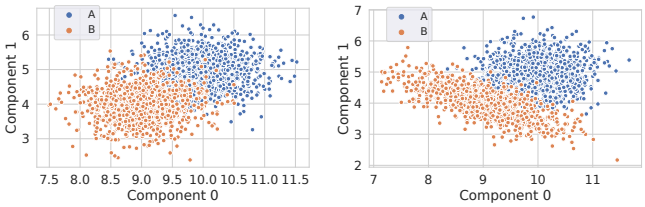

Figure 1: Data distribution for the two components on synthetic dataset 1 (left) and 2 (right).

Dataset with independent Gaussian distributions. Assume

$$f_0(x) \sim \mathcal{N}(10, 0.5), x \in \mathcal{A}; \quad f_0(x) \sim \mathcal{N}(9, 0.5), x \in \mathcal{B}.$$
$$f_1(x) \sim \mathcal{N}(5, 0.5), x \in \mathcal{A}; \quad f_1(x) \sim \mathcal{N}(4, 0.5), x \in \mathcal{B}.$$

We draw 1000 examples from each group, as shown in Figure 1 (left), where the x-axis is for $f_0(x)$ and y-axis for is $f_1(x)$. The two colors represent the two groups, respectively.

Table 1 shows the *ranking exposure* metric (as defined in Eq. (1), with $w = 0.65$) on this synthetic dataset. We see that Group $\mathcal{A}$ gets a significantly more exposure ($> 50\%$ more) than Group $\mathcal{B}$. We start by applying the fix on each component by *distribution matching* (Eq. (6)). From Table 1 we see that fixing only one component has very limited effect on overall system's fairness, and the end-to-end fairness can only be achieved by fixing both components (i.e., per-component fairness), which is consistent with Theorem 1 since both distributions are symmetric and independent of each other. Second, we test fixing each component by *distribution normalization* (Eq. (7)), and it is also effective in reducing the gap between the two groups while fixing only one component at a time.

Dataset with anti-correlated Gaussian distributions. In this experiment, we follow the same setting as the previous experiment except changing $f_0(x) = \mathcal{N}(13, 0.5) - f_1(x)$ for $x \in \mathcal{B}$ (we choose $\mu = 13$ for the first Gaussian such that $\mu[f_0(x)] = 9$, same as the first dataset) to create an anti-correlation between f_0 and f_1 for group $\mathcal{B}$. Again 1000 examples are sampled for each group, as shown in Figure 1 (right). Table 1 (right three columns) shows the fairness metrics on Synthetic Dataset 2. Compared with the results on Synthetic Dataset 1, we can see that the anti-correlation (thus breaking the symmetry requirement in Theorem 1) makes the end-to-end fairness metric much harder to achieve (larger overall gap).

5.2 German Credit Data

In this section, we demonstrate our analytical framework on a public academic dataset: the German Credit data[6], as another example to illustrate the effect of score composition on the end-to-end fairness. This dataset provides a set of attributes for each person, including credit history, credit amount, installment rate, personal status, gender, age, etc., and the corresponding credit risk.

We assume the final score for assessing credit risk is composed by the following four factors: 1) credit amount; 2) age; 3) number of existing credits at this bank (denoted as "num_credits" in the following), 4) number of people being liable to provide maintenance for (denoted as "num_liable"). We consider the problem of ranking all people in this dataset by the above score composition, and we

[5]The gap cannot be exactly 0 from discretization effects at the top of the list.
[6]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

	Dataset 1			Dataset 2		
Fixed Component(s)	Group $\mathcal{A}$	Group $\mathcal{B}$	Overall Gap	Group $\mathcal{A}$	Group $\mathcal{B}$	Overall Gap
None (baseline)	0.7640	0.2360	0.5281	0.7699	0.2301	0.5398
Distribution Matching						
Component 1	0.7433	0.2567	0.4865	0.7602	0.2398	0.5205
Component 2	0.6856	0.3144	0.3712	0.7318	0.2682	0.4636
Both	0.4818	0.5182	**-0.0365**[5]	0.6262	0.3738	**0.2524**
Distribution Normalization						
Component 1	0.5472	0.4528	0.0943	0.6156	0.3844	0.2312
Component 2	0.5470	0.4530	0.0940	0.5765	0.4235	**0.1529**
Both	0.4858	0.5142	**-0.0285**	0.6950	0.3050	0.3899

Table 1: The ranking exposure metric (over individual component and after composition) on Synthetic Dataset 1 and 2. Dataset 1 has independent distributions, and dataset 2 has anti-correlated distributions between components.

Fixed Component(s)	Male Rep.	Female Rep.	Overall Gap
None (baseline)	0.6081	0.3919	0.2162
credit amount	0.5852	0.4148	0.1704
age	0.5865	0.4135	0.1731
num_credits	0.5986	0.4014	0.1972
num_liable	0.5953	0.4047	0.1907
credit amount & age	0.5652	0.4348	0.1304
credit amount & age & num_liable	0.5392	0.4608	0.0783
All components	0.5352	0.4648	0.0705

Table 2: Compositional fairness by distribution matching for each component on the German Credit dataset.

consider the end-to-end fairness metric to be the ranking exposure (Eq. (1), $w = 0.65$) with respect to *gender* groups: male, female[7].

In the first setting, we assume the same group size, i.e., $|\mathcal{A}| = |\mathcal{B}|$, and in order to achieve demographic parity, the top N people within each gender group should receive the same ranking exposure. In this case the ideal exposure gap should reach zero. In Table 2, we show the effect on the end-to-end fairness, in terms of the percentage of male and female representation in the end ranking, and the exposure gap between them. We use *distribution matching* (Eq. (6)) to improve the system, and we use the counterfactual testing (Section 4.2) to test the effect of fixing each component alone, and the effect of fixing different combinations of the components.

In Table 2 we see that distribution matching for each component independently can help on the compositional fairness (smaller "Overall Gap") to different extents. Fixing multiple components simultaneously effectively further improves compositional fairness, and the overall gap is reduced most when all components are fixed. This headroom analysis provides us guidance on which components should be prioritized for improving end-to-end fairness.

In the second setting, we do not assume the same group size, and we rank $\mathcal{A}$ and $\mathcal{B}$ proportional to their respective sizes ($|\mathcal{A}| = 690$ for male, $|\mathcal{B}| = 310$ for female on this dataset). We vary the number of top positions t and plot the exposure gap metric with respect to the positions. As a reference, we also plot the exposure gap under random ordering (denoted as "Gap (random)" in the figures, by averaging over 100 runs), which ranks each person regardless of their gender. In Figure 2, we show the results by *distribution normalization* on multiple components simultaneously. The title of each sub-figure indicates the components that we have applied

[7] As that is how gender is categorized in the dataset.

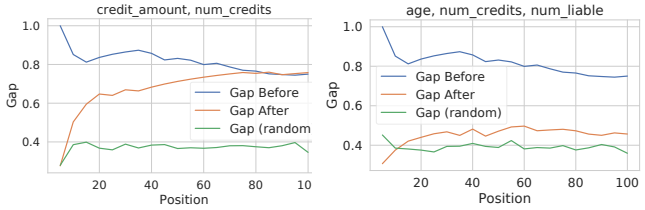

Figure 2: Gap between gender groups with respect to each position, with different group sizes, by distribution normalization on the German Credit dataset.

distribution normalization on. We can see that under this more realistic setting, our proposed fixes over per-component (and over combinations of components) can still effectively lead to significant improvements on the end-to-end fairness metric.

Empirically the results indicate that fairness does compose approximately on this dataset, we further analyzed the data distribution and found that most of the component scores (and the sum of many component combinations) are Gaussian-distributed (and thus symmetric), hence following Theorem 1 compositional fairness holds relatively well on this dataset.

5.3 Case Study on A Real Production Recommender System

In this section, we describe the results on a large-scale real-world recommender system. On an abstract level, the system mainly consists of three different components, one predicting the probability of click (denoted as "CTR"), and two other components predicting different signals of user satisfaction, denoted as "Satisfaction 1 (S1)" and "Satisfaction 2 (S2)", similar to [6]. In the following, we present results by fixing each individual component by:

- Matching the marginal distribution of $f_k(x), x \in \mathcal{A}$ and $f_k(x), x \in \mathcal{B}$, as in Eq. (6).
- Matching the conditional distributions of $X^k_{A_0}$ and $X^k_{B_0}$, and the conditional distributions of $X^k_{A_1}$ and $X^k_{B_1}$, building on (6).
- Matching the distribution on the delta terms: $\Delta_k(\mathcal{A}_1, \mathcal{B}_0)$ and $\Delta_k(\mathcal{B}_1, \mathcal{A}_0)$, as in Theorem 4.

The results are shown in Table 3, the first column denotes the "fixed" component(s), and column 2-4 show the Pairwise Ranking Accuracy Gap (Eq. (2)) for each component (abbreviated to "CTR", "S1", "S2"), respectively. Column 5-7 show the overall (compositional) Pairwise Ranking Accuracy for Group $\mathcal{A}$ and $\mathcal{B}$ and the

Fixed component(s)	Gap (CTR)	Gap (S1)	Gap (S2)	Group $\mathcal{A}$ Acc.	Group $\mathcal{B}$ Acc.	Overall Gap
None (baseline)	0.1024	0.1781	0.1078	0.5862	0.7408	0.1546
Matching on Marginal Distributions						
CTR	**0.0049**	0.1781	0.1078	0.6198	0.7084	0.0886
Satisfaction 1	0.1024	**0.0103**	0.1078	0.6292	0.7054	0.0762
Satisfaction 2	0.1024	0.1781	**0.0202**	0.6021	0.7270	0.1248
All	0.0049	0.0103	0.0202	0.6781	0.6546	0.0236
Matching on Conditional Distributions						
CTR	**0.0057**	0.1781	0.1078	0.6164	0.7048	0.0884
Satisfaction 1	0.1024	**0.0092**	0.1078	0.6258	0.7039	0.0781
Satisfaction 2	0.1024	0.1781	**0.0197**	0.6003	0.7258	0.1255
All	0.0057	0.0092	0.0197	0.6697	0.6472	0.0225
Matching on Delta Distributions						
CTR	**0.0000**	0.1781	0.1078	0.6473	0.7408	0.0935
Satisfaction 1	0.1024	**0.0000**	0.1078	0.6669	0.7408	0.0739
Satisfaction 2	0.1024	0.1781	**0.0000**	0.6197	0.7408	0.1211
All	0.0000	0.0000	0.0000	0.7630	0.7408	**0.0222**

Table 3: Compositional fairness by distribution matching in each component, on a large-scale real-world recommender system.

overall Pairwise Ranking Accuracy Gap. We see that first, compared to matching on marginal/conditional distributions, matching on the delta distributions is the only method that achieves zero gap on the per-component gap metric (Column 2-4 in Table 3). This is consistent with our theory (Theorem 4). Second, although marginal/conditional distribution matching does not provably ensure per-component fairness, empirically they still lead to a significant gap reduction (all close to zero), and effectively help on the compositional fairness (Column "Overall Gap" in Table 3). Finally, compositional fairness is better achieved when all the components are fixed, and fixing per-component alone only helps to a certain extent on the compositional fairness.

Analysis over the component scores shows that in this system, most components are correlated with each other, hence it is easier for the conditions in Theorem 2 to hold. This is consistent with the results in Table 3 where we see that fairness composes well after making each component independently fair.

5.4 Improving the Real Recommender

Last, we test if applying model regularization approaches to each component successfully improves the fairness of the overall recommender system. We implemented the fairness loss similar to [45], since it is most similar to our offline label-conditioned distribution matching analysis (Section 4.1.2). Specifically, we minimize the Maximum Mean Discrepancy (MMD) [30] between the two delta distributions: $\Delta_k(\mathcal{A}_1, \mathcal{B}_0)$ and $\Delta_k(\mathcal{B}_1, \mathcal{A}_0)$. The results are shown in Table 4[8]. As we see there, applying the MMD regularization to any individual component leaves a significant pairwise accuracy gap, but by applying it to all of the components (still independently to each one) significantly reduces the pairwise ranking accuracy gap. This is highly encouraging in that it demonstrates that significant progress can be made on fairness even in multi-component recommender systems and suggests that the insights gained from our counterfactual testing could be very useful in deciding when per-component fixes should be applied.

[8]Note, as these results are from continuous training, the dataset and thus absolute numbers differ slightly from the previous analysis.

Fixed components	CTR	S1	S2	All
Distribution matching via MMD [45]	0.080	0.088	0.067	0.013

Table 4: Pairwise ranking accuracy gap on the real-world recommender system from modeling-based approaches (label-conditioned MMD regularization).

6 CONCLUSION

As most industrial recommender systems are composed of many models and tasks, understanding how and when fairness composes is crucially important to enabling the application of fairness principles in practice. In this paper, we explore, both theoretically and empirically, the question: given a recommender system where the end ranking score is the product of scores from each component, does making each component fair independently improve the full system's fairness? We formalize this problem in two recently proposed fairness metrics for ranking, *fairness of exposure*, and *pairwise ranking accuracy gap*. For both metrics we find the unfortunate challenge that recommender fairness is not guaranteed to compose, aligned with prior work in classifiers [23].

To overcome this obstacle, we focus on studying when *does* fairness compose. We first present theory showing conditions under which per-component fairness does compose. Because the theory shows that composition is distribution-dependent, we propose taking a data-driven approach to this problem. We offer an analytical framework for measuring how much improving per-component fairness will improve the recommender system's fairness and diagnosing which components we should prioritize improving to have the most impact. By applying our analytical framework to multiple datasets, including a large real-world recommender system, we find that in practice most of the end-to-end exposure or accuracy gaps can be addressed through independently applying per-component improvements.

Our results highlight that while guarantees do not hold in the worst-case, there is more nuance over realistic data distributions. We believe this framework is a strong foundation for future research on compositional fairness, with clear extensions for classification [11, 33, 42] and opportunities to generalize to more fairness metrics and compositional functional forms.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In *TKDE*.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*.

[3] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *WSDM*.

[4] Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. 2016. Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the 25th International Conference on World Wide Web*. 761–770.

[5] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, 35.

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *KDD' 19*.

[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. *arXiv preprint arXiv:1901.04562* (2019).

[8] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[9] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*. 405–414.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS 2016*.

[11] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 491–500.

[12] Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. In *User Modeling and User-Adapted Interaction, Volume 12, Issue 4*. 331–370.

[13] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[14] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *ICDMW '09*.

[15] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. 2011. Efficient Diversification of Web Search Results. In *VLDB*.

[16] J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR'98*.

[17] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[18] Mark Claypool, Anuja Gokhale, Tim Miranda, Paul Murnikov, Dmitry Netes, and Matthew Sartin. 1999. Combing content-based and collaborative filters in an online newspaper. *ACM SIGIR'99. Workshop on Recommender Systems: Algorithms and Evaluation* (1999).

[19] Andrew Cotter, Michael Friedlander, Gabriel Goh, and Maya Gupta. 2016. Satisfying Real-world Goals with Dataset Constraints. (2016).

[20] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *ICML*.

[21] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and et al. 2010. The YouTube Video Recommendation System. In *RecSys*.

[22] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. *arXiv preprint arXiv:2004.13157* (2020).

[23] Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. In *arXiv preprint arXiv:1806.06122*.

[24] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition. In *FATML*.

[25] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual Fairness in Pipelines. *arXiv preprint arXiv:2004.05167* (2020).

[26] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *RecSys '18*. 242–250.

[27] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *SIGIR*. 55–64.

[28] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *KDD '19*. 2221–2231.

[29] Sreenivas Gollapudi and Aneesh Sharma. 2009. An Axiomatic Approach for Result Diversification. In *WWW*.

[30] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. In *The Journal of Machine Learning Research, Volume 13*. 723–773.

[31] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*.

[32] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 1–9.

[33] Nathan Kallus and Angela Zhou. 2019. The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric. *arXiv:1902.05826* (2019).

[34] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *AIES '19*.

[35] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

[36] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The Variational Fair Autoencoder. In *ICLR*.

[37] Xiao Ma, Liqin Zhao, Guan Huang, Wang Zhi, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *SIGIR*.

[38] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*.

[39] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. In *Annals of Mathematical Statistics*.

[40] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. 2017. McGan: Mean and Covariance Feature Matching GAN. In *ICML*.

[41] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. 2017. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning* 10, 1-2 (2017), 1–141.

[42] Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Serena Wang. 2020. Pairwise Fairness for Ranking and Regression. (2020), 5248–5255.

[43] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *NIPS 2017*.

[44] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European Conference on Information Retrieval*. Springer, 810–817.

[45] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779* (2019).

[46] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning Diverse Rankings with Multi-Armed Bandits. In *ICML*.

[47] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD' 18*. 2219–2228.

[48] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*. 5426–5436.

[49] Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. 2010. Learning optimally diverse rankings over large document collections. In *ICML*.

[50] Indrė Žliobaitė. 2015. On the relation between accuracy and fairness in binary classification. *ArXiv* abs/1505.05723 (2015).

[51] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 105–114.

[52] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*. 2921–2930.

[53] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 113–120.

[54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Learning Fair Classifiers.

[55] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Fort Lauderdale, FL, USA.

[56] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *CIKM'17*. 1569–1578.

[57] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Association for the Advancement of Artificial Intelligence*.

[58] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending What Video to Watch next: A Multitask Ranking System. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*.