FlipTest: Fairness Testing via Optimal Transport

Emily Black*
emilybla@andrew.cmu.edu
Carnegie Mellon University

Samuel Yeom*
syeom@cs.cmu.edu
Carnegie Mellon University

Matt Fredrikson
mfredrik@cs.cmu.edu
Carnegie Mellon University

ABSTRACT

We present FlipTest, a black-box technique for uncovering discrimination in classifiers. FlipTest is motivated by the intuitive question: *had an individual been of a different protected status, would the model have treated them differently?* Rather than relying on causal information to answer this question, FlipTest leverages optimal transport to match individuals in different protected groups, creating similar pairs of in-distribution samples. We show how to use these instances to detect discrimination by constructing a *flipset*: the set of individuals whose classifier output changes post-translation, which corresponds to the set of people who may be harmed because of their group membership. To shed light on *why* the model treats a given subgroup differently, FlipTest produces a *transparency report*: a ranking of features that are most associated with the model's behavior on the flipset. Evaluating the approach on three case studies, we show that this provides a computationally inexpensive way to identify subgroups that may be harmed by model discrimination, including in cases where the model satisfies group fairness criteria.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing**;

KEYWORDS

fairness, machine learning, optimal transport, disparate impact

ACM Reference Format:
Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. FlipTest: Fairness Testing via Optimal Transport. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3351095.3372845

1 INTRODUCTION

With the recent introduction of machine learning in sensitive applications like predictive policing [17] and child welfare [40], the question of whether these algorithms can lead to unfair outcomes has gained widespread attention. These concerns are not merely hypothetical. Racial bias in the COMPAS recidivism prediction model [4] and gender bias in Amazon's hiring model [11] suggest that discriminatory models can have wide-reaching harmful effects.

A growing set of strategies have emerged for testing and detection of such discriminatory behaviors. A common approach that applies to group fairness criteria such as demographic parity [18] and equalized odds [22] is to measure aggregate statistics of the model's behavior on a targeted population. For example, this approach was taken with the COMPAS system for recidivism prediction by measuring false positive and negative rates across Caucasian and minority populations [17], and is supported by IBM's AIF360 toolkit for assessing model fairness [7]. However, there are several potential issues with this approach [9, 27, 44], among which is that models can potentially "pass" such audits while still behaving unfairly towards individuals, or even targeted subgroups [31]. Additionally, while aggregate statistics can reveal broad patterns of potential discrimination, they do not reveal additional information that sheds light on the underlying discriminatory mechanism at play, which is crucial when assessing whether the behavior is truly problematic.

Recent work [2, 19] instead searches for discrimination at the individual level, testing whether changes in the protected demographic status of an individual can cause changes in model outcome. However, to change the protected demographic status of an individual, these methods simply flip the value of the protected attribute (e.g., race or gender). While this can ensure that the model does not directly use the protected attribute to discriminate, it still allows the model to disproportionately harm a protected group by using features that are correlated with the protected attribute.

The framework of *counterfactual fairness* by Kusner et al. [29] takes these correlations into account by assuming a causal generative model for the relevant data. This approach has the advantage that instances of discrimination against individuals or small subgroups cannot "fly under the radar", and the causal generative model may lead to a more nuanced and granular understanding of how the model discriminates. However, the reliance on detailed causal information creates practical issues that may limit its applicability as well. Namely, it may not be feasible to assume access to a generative causal model in many applications, and if an inaccurate model is used, then the conclusions may be misleading. Moreover, the legal frameworks governing discrimination in many countries (e.g., *disparate impact* in the US [37] and *indirect discrimination* in the UK [32]) do not require a causal relationship with the protected status, so tests based on counterfactual fairness may fail to identify instances of legally actionable discrimination.

In this paper we present FlipTest, a *black-box, efficient, and interpretable* fairness testing approach that is motivated by the following intuitive question: *had an individual been of a different protected status, would the model have treated them differently?* In contrast to aggregate testing methods, FlipTest reasons about the model's behavior on individuals and subgroups to look for evidence of discrimination, and can thus uncover forms of discrimination to which group fairness measures are blind. However, unlike counterfactual fairness, FlipTest *does not rely on causal information*, and instead

*The first two authors contributed equally to this research.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372845

uses an *optimal transport mapping* [41] to answer the question above. Consequently, the goal of our test is not to demonstrate a causal link between the protected attribute and the model's output, but to showcase salient patterns in a model's behavior that may be indicative of discrimination. Importantly, this means that FlipTest is sensitive to both statistical and causal discrimination, and does not require strong causal assumptions about the data-generating process. Further, we show that the information computed in this process can provide insight into not just *whether* a model discriminates, but *how* it does and *who* is likely to be affected.

Problem setting. We consider a setting where a machine learning system is being audited for discriminatory behaviors, either by well-intentioned stakeholders who may have been involved in the model's construction, or by concerned practitioners outside of the development process. Ideally, the auditors include a domain expert who is familiar with the application and the subject population who will come into contact with the algorithm. We assume that these individuals and those responsible for training the model are not intentionally trying to evade a finding of discrimination.

Optimal transport. FlipTest uses an *optimal transport map* [41] to construct instances that may reveal whether a model's behavior is sensitive to changes in protected status. An optimal transport map transforms one probability distribution into another, while minimizing a given cost defined over their respective supports. For example, we might use an optimal transport map from the distribution of men to women in order to obtain a (female, male) pair of inputs with which to query the model. If the model's output differs for these two people, then it may be evidence that the model discriminates on the basis of gender. Using optimal transport to compare protected group outcomes is advantageous because it translates exactly from one distribution to another, generating inputs that are in the distribution of its image. When the image of an optimal map corresponds to a distribution that the model was trained on, the results will reflect characteristic model behavior that can be expected when the model is deployed. This is not necessarily true for other methods of generating alternate inputs on which to compare model outcomes, e.g. input influence measures [12]. Further, the mapping does not rely on causal information, and can reveal associative forms of discrimination that causal tests cannot while requiring fewer assumptions about that data.

A key challenge with this approach lies in constructing the mapping, which can be computationally demanding with large, high-dimensional datasets. In recent years there have been notable advances in methods for efficiently approximating optimal transport maps [34], and FlipTest's efficacy can benefit from ongoing work on this problem. In this paper, we present an approximation method based on generative adversarial networks (GANs) [20] (Section 3.1), and validate it by showing that it is feasible to construct good, stable approximations of known precise mappings (Section 5.1).

Finding evidence of discrimination. Beyond examining model behavior on individual pairs, we show how the information provided by the optimal transport map can be systematically evaluated for evidence of discrimination. In particular, we assume that the classifier in question produces binary outputs, one of which is seen as a favorable outcome and the other as unfavorable. We consider two sets of individuals under the optimal map: those whose prediction changes from favorable to unfavorable, and vice versa. We call these *flipsets*, and look to the relative size of each flipset for signs of potential discrimination; for example, we show how flipsets relate to well-known fairness criteria like demographic parity and equalized odds (Section 4). In addition, large flipsets can indicate subgroup-level discrimination that is not well described by these group fairness criteria (Section 5.3). By comparing the distribution of the flipsets to the distribution of the overall population, it is often possible to identify specific subgroups that the model discriminates against.

To gain insight into how the model discriminates, we construct a *transparency report* that summarizes the most salient statistical differences in the features between the flipset individuals and their images under the optimal transport mapping (Section 4.2). Intuitively, the transparency report can serve as an overview of what features the model may be using to discriminate between populations.

However, it is not guaranteed that a person in the flipset is the victim of discrimination. For example, an inter-group disparity in the model's output may be permitted if there is a sufficient justification such as a "business necessity" [6, §II.B]. Therefore, we treat the flipset analysis and transparency report as a starting point for further investigation about the potential unfairness of the model, which can be followed up with more expensive and conclusive analyses that look at whitebox model information [12, 13].

Experiments. We empirically evaluate FlipTest on four different datasets (Section 5), demonstrating the testing workflow that we envision for it: Chicago Strategic Subject List [10], an illustrative dataset from Lipton et al. [31], the law school success dataset used to illustrate counterfactual fairness [29], and a synthetic dataset we construct to demonstrate differences from prior work. Our resuls show that FlipTest provides clear, interpretable evidence of discrimination in a range of settings, along with concrete diagnostic information that is useful when reasoning about the model behaviors that are responsible for the discrimination.

We compare FlipTest against two prior approaches: counterfactual fairness [29] and FairTest [38]. For counterfactual fairness, we examine the dataset used by the authors to evaluate the approach, and compare FlipTest's optimal transport-based results against those obtained by making comparable interventions on the generative causal model given by Kusner et al. [29, Section 5]. We find that the two lead to similar conclusions about the model's tendency to discriminate, despite the fact that FlipTest makes substantially fewer assumptions about the data. For FairTest, an approach based on statistical hypothesis testing of subgroup discrimination, we show that FlipTest can complement FairTest by detecting instances that FairTest misses.

Summary. To summarize, our main contributions are: *(1)* FlipTest, a black-box, efficient testing approach for detecting discrimination in classifiers; *(2)* the novel use of optimal transport for fairness testing; and *(3)* the application of FlipTest to two case studies involving predictive policing (Section 5.2) and hiring (Section 5.3), as well as comparisons to prior fairness testing methods (Sections 5.4, 5.5), which demonstrate that our approach can identify concrete examples of unfair model behavior in cases where prior testing methods do not.

2 AN ILLUSTRATIVE EXAMPLE

In this section, we illustrate the main concepts behind FlipTest with a running example, which uses a synthetic dataset created by Lipton et al. [31, §4.1]. This dataset consists of two features, hair length and work experience, and supposes a binary classifier that uses these features to decide whether a given person should be hired. We investigate possible gender bias in this model, asking whether the model's output would have been different had a given person been of a different gender. However, it is not sufficient to simply flip the gender attribute due to correlations in the data that the model may be use as a proxy for gender: in this data, gender is correlated with hair length and work experience. Additionally, flipping the gender attribute is not an option because the model does not directly use this attribute.

We instead map the set of women in the data to their male correspondents, and analyze cases where the model treats women differently from the men that they are mapped to. This raises the question of which specific man a given woman should be mapped to, for which we appeal to the intuition that a difference in treatment between two people is not by itself strong evidence of discrimination unless they are similar enough that the disparity cannot be justified. For example, when a man with 20 years of relevant experience is hired over a woman with no experience, this difference would likely be attributed to work experience rather than gender discrimination. This motivates our use of an *optimal transport mapping* [41], which minimizes the sum of the distances between a woman and the man that she is mapped to (her *counterpart*), where the distance quantifies how different a pair of people are.

We must now specify a distance function (or *cost function*) to operationalize the optimal transport mapping. Although there are no easy answers to the question of which people are similar for the purpose of establishing discrimination, the goal of this paper is to demonstrate a new technique for finding evidence of possible discrimination rather than to present a conclusive definition of discrimination. We find that the square of the L_1 distance leads to reliable results (Section 5), so we use it in this paper, but our approach is compatible with any cost function deemed suitable for the setting.

We then analyze the optimal transport mapping, which is depicted in the top plot of Figure 1. The analysis is through the *flipset* $F(h, G)$, which consists of all women whose outcomes were different from their counterparts'. We partition the flipset into the *positive flipset* $F^+(h, G)$, which contains the hired women whose counterparts were not, and the *negative flipset* $F^-(h, G)$, which is the set of rejected women whose counterparts were hired. Thus, in some sense the women in $F^+(h, G)$ were advantaged due to their gender, and those in $F^-(h, G)$ disadvantaged. Although this is not sufficient to establish that gender *caused* the difference in the way that some causal tests [12, 29] can, FlipTest has the advantage that it queries the model on in-distribution points only, so its response to these inputs is likely to be reliable.

We begin by examining the size of the positive and negative flipsets. Suppose that the model does not satisfy demographic parity, hiring disproportionately more men than women. Then, the flipsets will have different sizes, with the negative flipset larger than the positive. Therefore, a large difference in the sizes of the flipsets is evidence of possibly discriminatory behavior in the model. However, such differences may also be based on a justifiable reason,

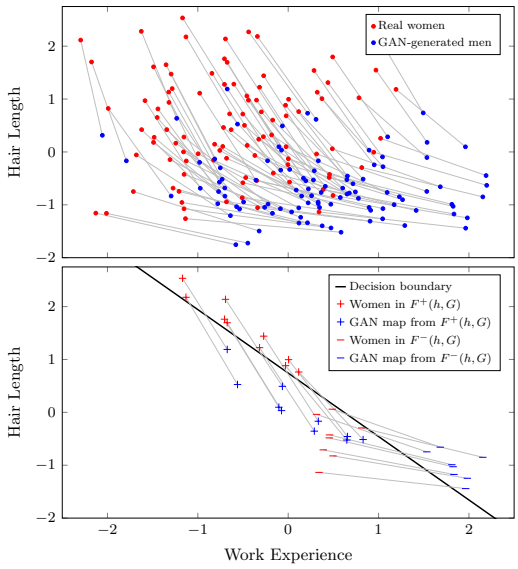

Figure 1: (Top) Optimal transport mapping from women to men in the Lipton et al. [31] synthetic dataset. This is an approximation generated by a GAN, as described in Section 3.1. (Bottom) Flipsets as defined by the model with the given decision boundary.

such as when the job in question has requirements that are more likely to be satisfied by a particular gender.

Alternatively, it could be that the positive and negative flipsets have the same size. If this is the case, i.e., the *net* flipset size is zero, then the model satisfies demographic parity. However, if the sizes of the individual flipsets are still large, then there may be discrimination at the subgroup level. To investigate which subgroups may be discriminated against (or unfairly advantaged), we can compare the distributions of the flipsets to that of the entire population. We plot the marginal distributions in Figure 5, and the results show that the advantaged ($F^+(h, G)$) women tend to have much longer hair than the disadvantaged ($F^-(h, G)$) women, suggesting that the model may be discriminating against shorter-haired women.

Finally, we can produce a *transparency report* to gain more information about why the model may behave in this way. The transparency report describes how the members of the flipsets are different from their counterparts, shedding light on which features may have contributed to the model's decision to classify the counterparts differently. As we can see in the bottom plot of Figure 1, the women in the negative flipset have much less work experience than their counterparts, and this suggests that they were not hired because of inadequate work experience. Although this method is not foolproof because work experience could have been a correlate of another feature that the model actually uses, it points to a specific aspect of model behavior for further investigation with tools such as QII [12] that can ascertain which feature is most responsible for the model's behavior. If it turns out that work experience causes the difference in the hiring decisions—and in our example it does—a practitioner can consult a domain expert to decide whether the use of work experience is justified. In many cases it would be justified, but it

may not if the disparity in work experience is due to discrimination.

In the rest of this paper, we solve the main technical challenges behind FlipTest and experimentally verify that it reaches the correct conclusions in many settings. We formally describe the optimal transport mapping in Section 3, and we show how to use a GAN to efficiently approximate an optimal transport mapping in a way that generalizes to unseen data. While the example in this section is based on demographic parity, we can split the dataset by the true label to test for a more individualized notion of equalized odds as well. Our experiments in Sections 5.2 and 5.3 show that FlipTest can identify possible subgroup discrimination in cases where the relevant group fairness objective (demographic parity or equalized odds) is met and in cases where it is not. Finally, in Sections 5.4 and 5.5 we compare FlipTest to prior similar methods.

3 OPTIMAL TRANSPORT MAPPING

In this section, we describe the optimal transport problem in more detail, and show in Section 3.1 how to solve a GAN objective to approximate the optimal transport mapping in a way that generalizes to unseen data points. In Section 3.2 we compare this GAN approximation to the exact optimal transport mapping and another approximation method by Seguy et al. [36], finding that the GAN tends to give more stable mappings. Although this GAN approximation is not the only way to operationalize FlipTest, we use the GAN approximation throughout the rest of this paper because it appeared to give more reliable results than the alternatives we considered.

We first introduce the notation. Let $\mathcal{S}$ and $\mathcal{S}'$ be two distributions defined over the feature space $\mathcal{X}$. In practice, we do not know these distributions, so we usually deal with observations of points drawn from these distributions instead. We will use the sets $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $S' = \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n\}$ to denote the observed points, where $n = |S| = |S'|$. Note that here we assume that $|S| = |S'|$ for ease of exposition, as this assumption allows the resulting exact optimal transport mapping to be deterministic. The general case where $|S| \neq |S'|$ can be handled through the use of randomized optimal transport mappings, and our approximation methods do not require that the two sets have equal size.

Let $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a cost function that describes the cost of moving between two points in the feature space $\mathcal{X}$. Intuitively, an optimal transport mapping from S to S' is a minimum-cost way to move the points in S such that the end result is S'. Thus, if more similar pairs of points have a lower cost, an optimal transport mapping describes how to match points in S with their similar counterparts in S'. Formally, an optimal transport mapping can be defined as a bijection $f : S \to S'$ that minimizes the expected cost $\mathbb{E}[c(\boldsymbol{x}, f(\boldsymbol{x}))] = \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{x}_i, f(\boldsymbol{x}_i))$.

3.1 Approximation via GANs

While the exact optimal transport mapping between S and S' can be solved through a linear program or the Hungarian algorithm [28], these methods do not scale well to large n. In addition, the exact mapping, as well as some approximations thereof [3, 35], is not defined for points outside of S. Therefore, we instead propose a generative adversarial network (GAN) [20] to approximate an optimal transport mapping in a way that avoids both of these issues.

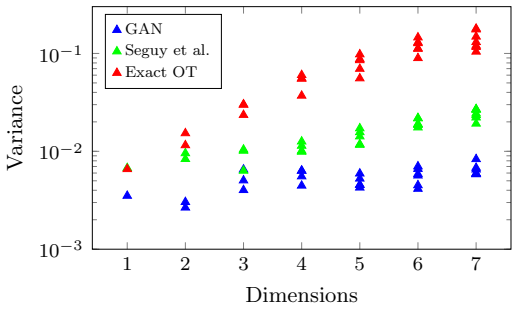

Figure 2: Variance of the GAN (blue), Seguy et al. [36] (green), and exact (red) optimal transport mappings over the different random draws of the observed points S and S'. Note the logarithmic vertical scale. The horizontal axis represents the number of dimensions in the feature space $\mathcal{X}$, and each plotted point represents the variance of one of the features. More details are given in Section 3.2.

Because we want to use the generator G as an optimal transport mapping, we assume that its inputs draw randomness from $\mathcal{S}$. For concreteness, we base our construction on the Wasserstein GAN [5], and note that our primary result (Proposition 1 below) can be extended to other types of GANs as well. When training a conventional Wasserstein GAN with the sets of observed points S and S', the generator's loss function is $(1/n)\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x}))$ for discriminator D, and the discriminator's loss function is $(1/n)\sum_{\boldsymbol{x}'\in S'} D(\boldsymbol{x}') - (1/n)\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x}))$. For the purpose of finding an optimal transport mapping, we modify the generator's loss function to take into account the cost of moving from a point in S to a point in S':

$$L_G = \frac{1}{n}\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x})) + \frac{\lambda}{n}\sum_{\boldsymbol{x}\in S} c(\boldsymbol{x}, G(\boldsymbol{x})) \quad (1)$$

Our modified generator has two objectives, with the parameter λ controlling their relative importance: generating the correct output distribution $\mathcal{S}'$, and minimizing the expected cost $c(\boldsymbol{x}, G(\boldsymbol{x}))$. Proposition 1 formalizes the intuition that these objectives are also those of an optimal transport mapping. The proof is given in Appendix A of the supplementary material.

Proposition 1. *Suppose that G^* is a minimizer of L_G among all G such that $G(S) = S'$. If $\lambda > 0$, G^* is an exact optimal transport mapping from S to S'.*

Although the generator G will not satisfy $G(S) = S'$ in practice, Proposition 1 motivates the use of this generator to approximate an optimal transport mapping. Our experimental results (Section 3.2) show that the approximate GAN mapping is more stable than the exact mapping, which is not very stable and changes drastically depending on which sets S and S' were drawn from $\mathcal{S}$ and $\mathcal{S}'$.

3.2 Stability

Here, we compare the behavior of the exact optimal transport mapping to those of approximate mappings. This is not intended to be a comprehensive evaluation of all approximation methods, but rather an argument for the use of approximations over the exact mapping for our purpose. We note that FlipTest is compatible with any optimal transport method.

To measure stability, we fix a point $\boldsymbol{x} \in S$ and then draw multiple distinct samples of the other $n-1$ points from $\mathcal{S}$ and n points from $\mathcal{S}'$. Thus, we have different sampled sets S and S' each time, and we observe the variance of the point $f(\boldsymbol{x})$ over the random draws.

As with all experiments in this paper, we used the square of the L_1 distance as the cost function. The exact optimal transport mapping was computed as a linear program with Gurobi [21] implemented in Python 3, and for the GAN approximation (Section 3.1), we trained a Wasserstein GAN [5] using Keras [8] with the TensorFlow backend [1]. For these experiments, we mapped the standard multivariate normal distribution to itself. Because the size of the linear program for the exact optimal transport mapping increases at least quadratically with the size n of the dataset, we used $n = 500$. Each experiment was repeated with 100 random draws of the dataset.

In the first set of experiments, we set $\boldsymbol{x}$ to be the one vector and observed the mean of $f(\boldsymbol{x})$. Since we map a distribution to itself, f should roughly be the identity function, and the mean of $f(\boldsymbol{x})$ should be similar to $\boldsymbol{x}$. While this was the case for the GAN approximation, the exact mapping displayed a significant "regression-to-the-mean" effect that increased with the number of dimensions in the feature space.

In the second set of experiments, we set $\boldsymbol{x}$ to the zero vector and noted the variance of $f(\boldsymbol{x})$ under all three types of mappings. The result is plotted in Figure 2, showing that the GAN approximation is much more stable than the exact mapping and somewhat more stable than that obtained by the method described by Seguy et al. [36].

The differences in both mean and variance persisted when we changed the data distribution by making the features correlated with each other. These differences can be explained by the fact that the exact mapping tends to overfit to the observed points, since it has to map every point to another observed point. As a result, approximate mappings are better suited for evaluating the fairness of a model that is trained to generalize. Since the GAN mapping appears to be more stable than that of Seguy et al., for the rest of the experiments we will exclusively use GANs as the optimal transport method in FlipTest. At the time of writing, we have not been able to evaluate the very recent GAN-based optimal transport approximation method by Leygonie et al. [30] for use in FlipTest, but any advantage of their method over the construction given here would translate to an improvement for FlipTest's results.

4 FLIPSETS AND TRANSPARENCY REPORTS

We leverage the optimal transport mapping to gather two main pieces of information from a model: who may experience discrimination, and which features may be associated with this effect. In Section 4.1, we describe *flipsets*, which we use to answer the first question, and in Section 4.2 we show how to use them to construct *transparency reports*, which help answer the second question.

4.1 Flipsets

We begin by introducing the *flipset* (Definition 1), which is the set of points whose image under a transport mapping is assigned a different label by a binary classifier.

Definition 1 (Flipset). *Let $h : \mathcal{X} \to \{0, 1\}$ be a classifier and $G : \mathcal{S} \to \mathcal{S}'$ be an optimal transport mapping (or an approximation). The flipset $F(h, G)$ is the set of points in S whose mapping into $\mathcal{S}'$ under G changes classification.*

$$F(h, G) = \{\boldsymbol{x} \in S \mid h(\boldsymbol{x}) \neq h(G(\boldsymbol{x}))\} \tag{2}$$

The positive *and* negative *partitions of $F(h, G)$ are denoted by $F^+(h, G)$ and $F^-(h, G)$.*

$$F^+(h, G) = \{\boldsymbol{x} \in S \mid h(\boldsymbol{x}) > h(G(\boldsymbol{x}))\}$$
$$F^-(h, G) = \{\boldsymbol{x} \in S \mid h(\boldsymbol{x}) < h(G(\boldsymbol{x}))\}$$

In our experiments, $\mathcal{S}$ and $\mathcal{S}'$ will correspond to two groups with differing values for a protected attribute, and h will be a classifier with the potential to be unfair. For example, suppose that $\mathcal{S}$ and $\mathcal{S}'$ respectively correspond to female and male job applicants and that h is used to determine which applicants should proceed to further rounds of interview. Then $F^+(h, G)$ is the set of female applicants who proceed to the next round but whose male counterparts under G do not, and $F^-(h, G)$ is the women who do not proceed but whose male counterparts do.

Note that we can also create flipsets based on a mapping $G' : \mathcal{S}' \to \mathcal{S}$ in the opposite direction. Then, in our example $F^+(h, G')$ is the set of male applicants who proceed to the next round but whose female counterparts under G' do not, and $F^-(h, G')$ is the men who do not proceed but whose female counterparts do. If G' and G compose to the identity function, these flipsets would have the same size as $F^-(h, G)$ and $F^+(h, G)$, respectively, and we can test for this property as a sanity check of our GAN mappings. We expand this discussion in Appendix D of the supplementary material.

If the distributions $\mathcal{S}$ and $\mathcal{S}'$ are equal, we would expect G to be the identity function, leading to empty positive and negative flipsets. This corresponds to the setting where the input features are independent of the protected attribute, thereby ensuring that the model cannot discriminate on the basis of the protected attribute. Proposition 2 shows that demographic parity only provides a weaker guarantee of zero *net* flipset size. Thus, if the positive and negative flipsets are nonempty but have equal size, some individuals may be experiencing discrimination even though demographic parity holds. In this proposition, we use the exact optimal transport mapping to avoid any noise introduced by the GAN approximation, and the proof is provided in Appendix A of the supplementary material.

Proposition 2. *Let h be a binary classifier and $G : S \to S'$, with $|S| = |S'|$, be the exact optimal transport mapping. Then, $|F^+(h, G)| = |F^-(h, G)|$ if and only if the model satisfies demographic parity on the observed points, i.e.,*

$$|\{\boldsymbol{x} \in S \mid h(\boldsymbol{x}) = 1\}| = |\{\boldsymbol{x}' \in S' \mid h(\boldsymbol{x}') = 1\}|.$$

If we instead consider the distributions $\mathcal{S}|(y = 1)$ and $\mathcal{S}'|(y = 1)$, conditioned by the true label y, we can prove a similar result about equality of opportunity [22], and if we consider $\mathcal{S}|(y = 0)$ and $\mathcal{S}'|(y = 0)$ as well, we can extend the result to equalized odds [22].

When h is biased, the flipsets can provide several additional forms useful information about the model's behavior. First, the relative sizes of $F^+(h, G)$ and $F^-(h, G)$ can serve as a simple test of group fairness. Second, the absolute sizes of $F^+(h, G)$ and $F^-(h, G)$, if they are large, can indicate possible discrimination at the subgroup level. Third, if the distributions of the flipsets are different from $\mathcal{S}$, we gain information about *which* subgroup may be discriminated against. We illustrate these insights in greater detail in the case studies described in Section 5.

4.2 Transparency Reports

A *transparency report* (Definition 2) identifies features that change the most, and most consistently, under G between members of a given flipset ($F^+(h, G)$ or $F^-(h, G)$). These features are likely candidates for the underlying reasons for the observed discrimination, and can be examined further using more costly causal influence methods [12] to make a final determination.

Definition 2 (Transparency Report). *Let $h : \mathcal{X} \to \{0, 1\}$ be a classifier, $G : \mathcal{S} \to \mathcal{S}'$ be an optimal transport mapping (or approximation), and $F(h, G)$ be the corresponding flipset. If $\mathcal{X} \subseteq \mathbb{R}^d$, we can compute the following vectors, each of whose coordinate corresponds to a feature in $\mathcal{X}$:*

$$\frac{1}{|F^\star(h,G)|} \sum_{\boldsymbol{x} \in F^\star(h,G)} \boldsymbol{x} - G(\boldsymbol{x}), \text{ and}$$
$$\frac{1}{|F^\star(h,G)|} \sum_{\boldsymbol{x} \in F^\star(h,G)} \operatorname{sign}(\boldsymbol{x} - G(\boldsymbol{x}))$$

Here, $\star \in \{+, -\}$. Together, these vectors define a transparency report, *which consists of two rankings of the features in $\mathcal{X}$, each sorted by the absolute value of each coordinate.*

Intuitively, the features ranked highest by the transparency report are those that are most associated with the model's differences in behavior on the flipset. As we show in Sections 5.2 and 5.3, these often align closely in practice with the features used by the model to discriminate.

5 EXPERIMENTS

We now apply FlipTest to real and synthetic datasets, illustrating its use in finding discrimination in models. We begin by providing additional validation of the GAN optimal transport approximation (Section 5.1), and then move on to two case studies: a biased predictive policing model (Section 5.2), a well as a hiring model (Section 5.3) that contains a subtle form of subgroup discrimination. In Sections 5.4 and 5.5, we compare FlipTest to prior fairness testing methods, namely counterfactual fairness-based auditing [29] and FairTest [38]. In all of the experiments, we trained GANs using the configuration described in Section 3.

5.1 GAN Validation

In general, the GAN does not converge to the target distribution. Moreover, limitations in the amount of data available to train the GAN will reduce the accuracy of the approximation. To evaluate the effect of these factors on flipsets, we trained a GAN with samples from two identical distributions. In this setting, as the sample size approaches infinity, the exact optimal transport mapping will lead to empty flipsets. Therefore, the results of this experiment would indicate how many flips we can expect due to the noise in the GAN approximation and the finite sample size.

Because we map a distribution to itself, in order to simulate a more typical application of GAN training, we added additional random features that are *dependent* on the protected attribute. We drew 10,000 points drawn from each distribution, and to amplify any errors in the GAN mapping, we trained a very complex model by fitting an SVM with RBF kernel to random labels. Further details on the experimental procedure are given in Appendix B of the supplementary material. In the end, the flipsets were small despite the above steps that were intended to increase the size of the flipsets: we observed $|F^+(h, G)| = 167$ and $|F^-(h, G)| = 148$, each accounting for approximately 3% of the data. This serves a benchmark for comparison with the models in the following sections, which have larger and more unbalanced flipsets. This difference is striking given that we would expect larger flipsets from the complicated SVM classifier here than the simpler models in the later experiments.

Table 1: Validations for GANs included in the experiments section. KS refers to the Kolmogorov–Smirnov two-sample test statistic. MSE Diff refers to the difference, between real and generated data, of the mean squared error of a linear regression model trained on the real data to predict a feature from the rest. For KS and MSE, we ran 10 trials; the mean is given first, with standard deviation in parentheses. Dist: OT, GAN refers to the average of the squared L_1 distance between a data point x and its counterpart $G(x)$ under an exact optimal transport mapping (OT) or an approximated GAN mapping (GAN).

Experiment	Features	KS (std)	MSE Diff (std)	Dist: OT, GAN
SSL: Dem Parity	Age	0.054 (0.001)	0.070 (0.048)	2.04, 1.64
	Gang Aff	0.018 (0.006)	0.094 (0.034)	
	Narc Arr	0.025 (0.002)	0.298 (0.058)	
SSL: Eq Odds (Neg)	Age	0.019 (0.021)	0.019 (0.057)	1.52, 1.42
	Gang Aff	0.007 (0.004)	−0.009 (0.053)	
	Narc Arr	0.019 (0.007)	0.094 (0.078)	
Lipton: Dem Parity	Work Exp	0.072 (0.014)	0.150 (0.204)	2.19, 2.09
	Hair Len	0.074 (0.043)	0.141 (0.114)	
Law School	LSAT	0.057 (0.012)	3.757 (0.462)	10.99, 11.10
	GPA	0.110 (0.027)	−0.040 (0.005)	

To further validate our GAN mappings, we computed the exact optimal transport mapping f on a 2,000-member subset of the data and computed the average squared L_1 distance between x and $f(x)$. Then, we compared this quantity to the average squared L_1 distance between x and $G(x)$ for the GAN generator G. If the GAN mapping closely matches the optimal transport mapping, we would expect these numbers to be similar.

We also examined the fit of the GAN-based approximation on the datasets used in the evaluation using the Kolmogorov–Smirnov (KS) two-sample test on the marginal distributions for each feature between the real target data S' and the generated data $G(S)$. The output of this test corresponds to the largest difference in the empirical distribution functions of the two samples, with a small KS-statistic corresponding to the case where the two distributions are similar. In Table 1, we only report the KS-statistic and its variance, and not the associated p-value, since we do not require or expect the two samples to be from the exact same distribution. We instead use the statistic as a metric to judge how far apart the distributions are, aiming for them to be as close as possible.

Since similarity tests on the marginals of a distribution do not account for correlation between features, we further validate the output by training a linear regression model to predict each feature from the others (e.g., in the SSL dataset, age from the other seven features). We train these regression models on the real target data S', and compare the accuracies of these models on S' to those on the generated data $G(S)$. If the mean squared error is similar between predicting all features from the true data and the generated data, we take this as evidence that the GAN has captured correlation

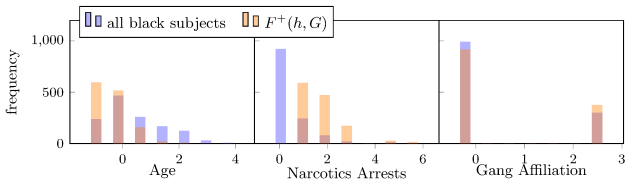

Figure 3: Distribution of the positive flipset for a black-to-white mapping (orange) and the overall black subject population (purple). On the x-axis we have (normalized) feature values; y-axis is number of individuals. We did not plot the flipset in the other direction, as it only has 4 individuals.

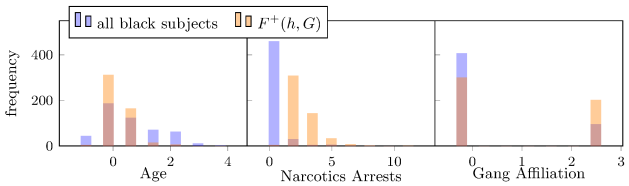

Figure 4: Distribution of the positive flipset for a black-to-white mapping of ground-truth negative subjects (orange) and the overall black ground-truth negative population (purple). The x-axis displays (normalized) feature values, and the y-axis displays frequency.

between the features well. For each of the experiments in the following sections, the validation results are also reported in Table 1. The KS and MSE statistics are reported over 10 trials.

5.2 Testing a Biased Model

The Chicago Strategic Subject List (SSL) dataset [10] consists of arrest data collected in Chicago for the purpose of identifying which individuals are likely to be involved in a violent crime, either as a victim or a perpetrator. We used the following eight features that are also used by the SSL model: number of times as victim of a shooting incident, age during last arrest, number of times as victim of aggravated battery or assault, number of prior arrests for violent offenses, gang affiliation, number of prior narcotic arrests, trend in recent criminal activity and number of prior unlawful use of weapon arrests. The target feature corresponds to the risk of being involved in a shooting, ranging in value from 0 to 500 (low to high risk). We normalized all input features to zero mean and unit variance, and a standard least-squares regression model for this data primarily relies on age ($w_{age} \approx -50$), giving the remaining features coefficients of magnitude less than 10.

We create a classification task from this dataset by setting a threshold on score (345); 10% of the dataset has a score above this threshold. As prior work [46] has shown that models trained on this data do not appear to exhibit significant bias, we deliberately bias the classifier by making it rely more heavily on the number of prior narcotics arrests, which is correlated with race ($r = 0.12$) and is arguably less predictive of involvement in violent crime than, say, number of prior arrests for violent offenses. The resulting linear classifier exclusively uses age and narcotics arrests, classifying a person as high risk if and only if $-53 \cdot age + 25 \cdot narc > 65$. As a result, the model is expected to assign black subjects higher average scores without necessarily being accurate.

We perform two tests on this model: in Section 5.2.1, we map between all black and white subjects, testing for criteria related to demographic parity, and in Section 5.2.2, we construct one map between the ground-truth positive (high-risk) black and white subjects, and another map between the ground-truth negative (low-risk) subjects, testing for criteria related to equalized odds.

5.2.1 FlipTest Demographic Parity. Additional details on experimental setup are found in Appendix C of the supplementary material.

The positive flipset $F^+(h, G)$ shows that 1,290 black subjects that are marked by the model as high risk are marked as low risk when sent through the black-to-white mapping (out of 3,683 black subjects marked as high risk). On the other hand, the negative flipset $F^-(h, G)$ consists of only 4 people, which is an extremely small fraction of the 37,877 black subjects marked as low risk. The size of $F^+(h, G)$ and the significant asymmetry of the flipsets suggest that the model discriminates on the basis of race, which is consistent with how the model was constructed.

To gain more insight about which subgroups may be discriminated against, we investigate the distributions of the flipsets as compared to that of the general black population. The histograms of the marginal distributions of select features is given in Figure 3, with the rest of the features plotted in Figure 8 in the supplementary material. For age and narcotic arrests, the flipset subjects skew away from the full population, towards younger people with more narcotic arrests. This suggests the bias of the model affects younger people with more narcotic arrests most. On the other hand, the marginals largely overlap for gang affiliation, which the model does not directly use.

We then look at the transparency report (Figure 8 in the supplementary material) to learn which features may be responsible for this apparently biased behavior. The black subjects in $F^+(h, G)$ changed the most, and most consistently, in narcotics arrests under the GAN mapping, which is again consistent with the type of bias that we introduced into this model. Thus, although the transparency reports only provide direct insight into statistical correlations between the features and model behavior, in this case they identified precisely the feature that is responsible for the model's bias. In general, based on the information from the transparency report, an practitioner testing a model can decide whether they would like to investigate the model further for bias based on the features in the transparency report. A more definitive evidence, such as that given by QII [12], would be required to conclude that the feature causes the model's discriminatory behavior.

In Appendix F and Figure 9 in the supplementary material, we present an additional set of experiments on this data using demographic parity, using a model that discriminates against black subjects based on a proxy involving several correlated features.

5.2.2 FlipTest Equalized Odds. We now analyze the SSL model, targeting a somewhat different fairness criteria related to equalized odds. We do so by training two optimal transport approximations: one to map only black ground-truth negatives to white ground-truth negatives, and one to map only black ground-truth positives

to white ground-truth positives. Both GANS were trained with $\lambda = 10^{-4}$. Additional experimental details are given in Appendix C of the supplementary material.

When mapping ground-truth negatives, we find that $F^+(h, G)$ contains 499 individuals, out of 5,002 ground-truth negative black subjects marked as high risk by the model. Meanwhile, $F^-(h, G)$ was empty. Prompted by the size and asymmetry of the flipsets, we compare the marginals of $F^+(h, G)$ compared to those of the overall black ground-truth negative population (Figure 4) to identify the subpopulation most affected by this potential discrimination.

$F^+(h, G)$ largely consists of younger subjects who are much more likely to have narcotics arrests in their record, and be a part of a gang, but have a lower trend in recent criminal activity. There is not much difference in the rest of the features. The complete histograms of the marginals and the transparency report are given in Figure 10 in the supplementary material. Figure 10 also contains the transparency report, which has narcotics arrests at the top of both lists, i.e., narcotics arrests is both the largest average change and the most consistent change under the GAN mapping for individuals in $F^+(h, G)$. Thus, the transparency report correctly identifies the source of bias that was injected into the model.

When mapping ground-truth positives, we find that $F^+(h, G)$ contains 216 individuals out of 1,568 black ground-truth positive subjects labeled as high risk, and that $F^-(h, G)$ has 102 individuals out of 3,767 black ground-truth positive subjects labeled as low risk. The transparency report for $F^+(h, G)$ lists narcotics arrests as one of its main features, correctly identifying this source of bias.

The transparency report for $F^-(h, G)$ lists a positive change in age as one of the most consistent features. This suggests that some older black subjects were mapped to younger white subjects, and since the model relies most heavily on age, the white counterparts were marked as high risk while the black individuals were marked as low risk. Since the original SSL model, which generated the ground-truth labels, also relies heavily on age, the 629 ground-truth positive white subjects were all quite young, while the age distribution of the 5,335 ground-truth positive black subjects had a larger spread. Thus, the transparency report on $F^-(h, G)$ is a result of optimal transport, which inherently tries to match the generated distribution to the target distribution. Practitioners should be cognizant of this distribution-matching behavior when mapping between starkly different distributions, as the model may be justified in using some features that are associated with the protected attribute. This flipset points out the model's heavy reliance on age, and a domain expert may weigh in on whether or not this is justifiable. Full results are available in Figure 11 in the supplementary material.

5.3 Testing a Group-Fair Model

Previously, Lipton et al. [31] argued that some fair learning algorithms employ a "problematic within-class discrimination mechanism". To support this argument, they create a synthetic data distribution that consists of two features: work experience and hair length. They use this distribution to train a learning algorithm by Zafar et al. [47] that seeks to equalize hiring rates across genders. They then find that the resulting linear model uses hair length as a proxy for gender, thereby unfairly benefiting long-haired men and harming short-haired women. In this section, we replicate this experimental setting to demonstrate that our method can detect unfair behavior in a model that appears to be fair at the population level.

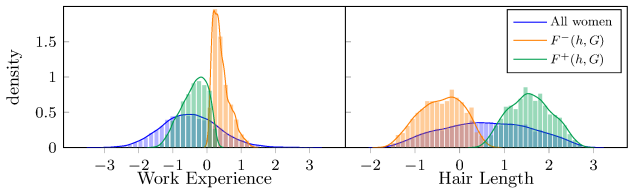

Figure 5: Distribution of women in the hiring data generated by Lipton et al. [31] (Section 5.3). The green distribution is that of the women who were hired while their male statistical counterparts were not, and the orange distribution represents those who were not hired while their male statistical counterparts were, according to a model of Zafar et al. [47] that seeks to equalize hiring rates across genders.

We trained a linear model on 10,000 men and 10,000 women drawn from the male and female distributions, respectively, after scaling each feature to zero mean and unit variance. Then, we trained approximate optimal transport mappings ($\lambda = 10^{-4}$) in both directions and evaluated the fairness of the model on a test set of 10,000 men and 10,000 women drawn from the same distributions. Here we present the women-to-men mapping, and the results on the men-to-women mapping is given in Appendix D of the supplementary material.

At the population level, the model treated the two groups similarly, hiring 27% of the men and 30% of the women. However, when we mapped the women to the men 715 women were disadvantaged by the model (rejected with hired male counterpart) whereas 1215 were advantaged (hired with rejected male counterpart). These flipsets comprise a much larger portion of the population than those encountered in Section 5.1, suggesting some discriminatory behavior at the subgroup level. Looking more closely at distributions of these flipsets (Figure 5) provides insight into the subgroups experiencing discrimination: the disadvantaged women tend to have shorter hair and longer work experience than the average woman, and the advantaged women have the opposite characteristics. This is consistent with the observation of Lipton et al. [31] that the model penalizes people with a masculine characteristic (short hair) in order to equalize hiring rates.

In addition, the transparency report shows that the disadvantaged women have much less (1.3 standard deviations) work experience and slightly longer (0.5 SD) hair than their counterparts, while the advantaged women have slightly less (0.6 SD) work experience and much longer (1.6 SD) hair than their counterparts. This suggests that the disadvantaged women are disadvantaged due to their short work experience and that the advantaged women are advantaged due to their hair length. We note that in general, this is not sufficient to establish *causal* claims about why the model behaves this way; for example, the difference in hair length may be due to the model's use of some other feature that is correlated with hair length. In this case, however, the model's weights—which define its causal behavior—and the result of the transparency report agree. The model, which is a linear regressor, weights the two features almost

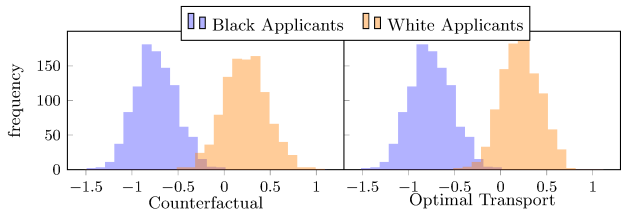

Figure 6: Results of regression model predicting first year law school success on a) counterfactuals generated with a causal model and b) alternate inputs generated with optimal transport. The x-axis is predicted first year average at law school, and the y-axis is frequency.

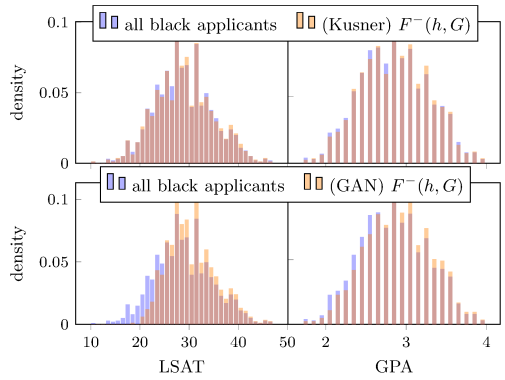

Figure 7: Flipset distributions for the counterfactual (Kusner) and the GAN-generated data. Transparency report results are included in Figure 12 in the supplementary material.

equally ($w_{hair} = 1.4$, $w_{work} = 1.2$). Since work experience, but not hair length, is a legitimate factor in most hiring decisions, after this deeper investigation into the model we may be able to conclude that this apparently fair model in fact discriminates against some men.

5.4 Comparison with Counterfactual Fairness

In this section, we compare FlipTest with the auditing technique described by Kusner et al. [29] on a law school dataset [43]. We test for fairness in two models: a regression model that predicts the first year average grade (FYA) in law school based on LSAT score, undergraduate GPA, race, and gender; and a binary classification model that predicts whether an individual will be in the top or bottom 50% of class using the same attributes. We create this second model by using the median FYA as a threshold.

Kusner et al. construct a causal model that postulates how race and gender are related to LSAT scores and GPAs, so that they can intervene on these attributes to construct counterfactuals. Their model assumes that GPA is distributed according to $\mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G)$ and LSAT as Poisson$(b_L + w_L^K K + w_L^R R + w_L^S S)$, where R is race, S is sex, and K is knowledge, an underlying parameter the model postulates is independent of race and sex. $w_\star^R$ and $w_\star^S$ are weights for race and sex specific to the GPA and LSAT distributions, and knowledge is distributed as $\mathcal{N}(0,1)$. Using this causal model, they create counterfactuals by intervening on the protected attributes, i.e., generating the underlying knowledge value for a section of the law school data, and then re-generating LSAT and GPA scores with flipped race attributes.

Kusner et al. use these counterfactual distributions to audit the law school regression model. They determine that a model is counterfactually fair if the distributions of the model's prediction are identical for the pre-intervention and post-intervention inputs. We perform this same test, using both their causally generated inputs and the pairs generated by an approximate optimal transport mapping ($\lambda = 5 \times 10^{-5}$). Since Kusner et al. do not normalize their data for their models or counterfactual generation process, we also do not normalize this data. We see in Figure 6 that the two methods of data generation give similar results on this test, both suggesting a conclusion of discrimination by the model.

We also provide a comparison with the FlipTest method, again using both methods of generating target distributions. The flipset sizes are very similar for both models: both models have $|F^+(h,G)| = 0$, and $|F^-(h,G)|$ is 811 and 835 for the counterfactual and GAN data, respectively. Figure 7 shows that the distributions of the flipsets are similar, but the GAN-generated inputs tend to lead to a flipset that contains black applicants with higher GPA and LSAT scores. The transparency reports (Figure 12 in the supplementary material) for the two methods also largely agree with each other, with LSAT ranked higher than GPA, and gender having essentially no effect. Since gender has little effect on the flipset, we do not plot the gender distributions of the two flipsets, and we do not report the transparency report results for the race feature, since by definition of the flipset, race will always change. However, other features can lead to bias as well—the transparency report clearly shows that LSAT score changes significantly under the GAN mapping (> 1 standard deviation), and this may indicate another source of bias in the model.

We now investigate the weights of our model, comparing the transparency report to what we know about the model. First, there is discrimination in the model based on race, as evidenced by the fact that the feature *race=black* has a negative weight and the feature *race=white* has a positive weight. Second, after correcting for the fact that the data in this experiment is not normalized, we see that the weight for LSAT is higher than that for GPA. The transparency report is consistent with this observation.

We conclude this section by observing that, even without access to a causal model, FlipTest can give nearly identical results as causally generated counterfactuals.

5.5 Recognizing Biased Features

Here we demonstrate a major conceptual difference of FlipTest from FairTest by Tramèr et al. [38]. FairTest searches for subgroups experiencing possible discrimination by using a decision tree to iteratively search for subgroups within which there is high correlation between the model output and the protected attribute. It then reports the subgroups with has the highest correlations.

To clearly illustrate the difference between FairTest and FlipTest, we use a synthetic data distribution with only one feature: the number of prior arrests. This feature is distributed as Geometric$(1/4) - 1$

for people in $\mathcal{S}$ and Geometric$(1/2)-1$ for people in $\mathcal{S}'$. In addition, we assume a simple model, which classifies a person as low risk if the number of prior arrests is zero, high risk if it is two or more, and either low risk or high risk uniformly at random if there is exactly one prior arrest. Because the members of $\mathcal{S}$ tend to have more prior arrests than those in $\mathcal{S}'$, this model classifies disproportionately higher number of people in $\mathcal{S}$ as high risk.

We let S and S' be sets of 10,000 points drawn from $\mathcal{S}$ and $\mathcal{S}'$, respectively, and ran FairTest and FlipTest on these sets. FairTest correctly identified possible discrimination at the group level, reporting a confidence interval of $[0.2452, 0.2941]$ for the correlation between the model's output and the protected attribute. However, the correlation decreased in all subgroups that FairTest subsequently considered. This is because subgroups in FairTest are defined by the values of the input features. Thus, FairTest compares a subgroup of S to a subgroup of S' *that has similar numbers of arrests*.

On the other hand, FlipTest recognizes that the distributions of the number of arrests is different, and adjusts the comparisons accordingly. For example, the set of people with 1 arrest in S is compared to the set with 2–4 arrests in S'. As a result, the flipsets are very large and unbalanced, with $|F^+(h, G)| = 2572$ and $|F^-(h, G)| = 0$. In addition, the transparency report explains a reason for this difference, showing that 100% of the people in $F^+(h, G)$ had more arrests than their statistical counterparts in $\mathcal{S}'$, with an average of 1.44 more arrests.

We see this phenomenon on real data as well. On the SSL data (Section 5.2), FairTest notes an overall bias against the entire black population, but does not report that the discrimination is based on narcotics arrest since the feature itself is biased, with higher levels for black subjects than white subjects. The full results are given in Figure 4 in the supplementary material.

Thus, the choice of the appropriate fairness test may depend on the setting. If the number of prior arrests is a strong indicator for recidivism risk, then it makes sense to compare subgroups of people with similar numbers of arrests. On the other hand, if the model had used a feature that is completely unrelated to crime, it would be harder to justify comparing people who are similar with respect to that feature. Our experiments in this section show that FairTest is better suited for settings where the input features can justify potential differences in the model's output.

6 RELATED WORK

Counterfactual Fairness Testing. Counterfactual fairness [29] compares the model's behavior on a real input and a counterfactual, causally generated input. Similarly, Datta et al. [12] perform causal interventions on input features to study which features are influential in changing in the output of the model, Wachter et al. [42] generate simple L_1-nearest counterfactuals as a form of explanations for model outputs, and Ustun et al. [39] develop a method that outlines what actions individuals can take to change their classification outcome in linear models. However, these methods, like those discussed in the introduction [2, 19], generate potentially unrealistic, out-of-distribution points, which can jeopardize their conclusions. By contrast, the points that we generate conform to the data distribution.

Optimal Transport. Others have proposed using the optimal transport map in the context of fairness, but to the best of our knowledge, it has not yet been used as a fairness testing mechanism. Del Barrio et al. [14] use the optimal transport mapping to extend a previous method [18] of data pre-processing, which obfuscates protected attribute information, to the multivariate setting. Concurrent work from Yang et al. [45] also develop a method of approximating an (unbalanced) optimal transport mapping using GANs. Their formulation is closely related to ours, but they do not consider its application to fairness testing. Altschuler et al. [3] and Quanrud [35] present efficient methods for approximating the optimal transport mapping, but the resulting mappings are not defined for previously unseen data points. By contrast, the mappings produced by Leygonie et al. [30], Perrot et al. [33], and Seguy et al. [36] do generalize to unseen points, making them suitable for use with FlipTest.

Individual Fairness. Individual fairness criteria [15, 16, 24, 48] bind guarantees about the fairness of a model's behavior to every individual, as opposed to an aggregated statistic. Dwork et al. [16] note that, in some cases, the model must essentially be a constant function to satisfy individual fairness and group fairness at the same time. They propose an alternative that applies an optimal transport mapping to one of the groups, obtaining a transformed dataset on which they solve individual fairness. FlipTest is motivated by this approach, but we specifically look for potentially discriminatory differences between pairs of individuals *that belong to different groups*, and use optimal transport to construct pairs of individuals that exemplify these differences.

Subgroup Fairness. One application of this work is uncovering *subgroup unfairness* [23, 26], i.e., identifying subgroups that are possibly harmed as a result of their group membership. FairTest [38] uses a decision tree to find subgroups with high discriminatory association, while taking care to ensure that the association is statistically significant. However, as we show in Section 5.5, FairTest does not handle the case where the input feature itself is biased. Zhang et al. [49] also find a computationally faster way to search the exponentially large number of subgroups, but their method also suffers from the same issue that FairTest does. Kearns et al. [25] prove that checking for subgroup fairness is equivalent to weak agnostic learning, which is computationally hard in the worst case. FlipTest differs from these works in that we do not require the subgroups of interest to be specified before the fairness testing.

7 CONCLUSION

FlipTest is a low-cost fairness testing framework that is sensitive to discrimination beyond group fairness metrics, is proficient at displaying unfair treatment in models with biased data, and can be used to as a first step towards detecting a model's method of discrimination. As future work, extending the framework beyond binary classifiers, which are the most commonly studied case in the fairness literature, and exploring the application of FlipTest to uncovering additional types of discrimination are both promising directions.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their thoughtful feedback. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1704845.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, 2015.

[2] Aniya Agarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. Automated test generation to detect individual discrimination in AI models. *arXiv preprint arXiv:1809.03260*, 2018.

[3] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. URL http://arxiv.org/abs/1810.01943.

[8] François Chollet et al. Keras. https://keras.io, 2015.

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

[10] City of Chicago. Strategic Subject List. https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np, 2017.

[11] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

[12] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, pages 598–617, 2016.

[13] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

[14] Eustasio del Barrio, Fabrice Gamboa, Paula Gordaliza, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. *arXiv preprint arXiv:1806.03195*, 2018.

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

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, pages 214–226, 2012.

[17] Equivant. Practitioner's guide to COMPAS core. http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[18] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[19] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Foundations of Software Engineering*, pages 498–510, 2017.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[21] Gurobi Optimization, LLC. Gurobi optimizer reference manual. https://www.gurobi.com/documentation/8.1/refman.pdf, 2019.

[22] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[23] Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1944–1953, 2018.

[24] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[25] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

[28] H. W. Kuhn and Bryn Yaw. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[29] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[30] Jacob Leygonie, Jennifer She, Amjad Almahairi, Sai Rajeswar, and Aaron Courville. Adversarial computation of optimal transport maps. *arXiv preprint arXiv:1906.09691*, 2019.

[31] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

[32] Parliament of the United Kingdom. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/contents, 2010.

[33] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

[34] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5–6):355–607, 2019.

[35] Kent Quanrud. Approximating optimal transport with linear programs. *arXiv preprint arXiv:1810.05957*, 2018.

[36] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

[37] Supreme Court of the United States. *Griggs v. Duke Power Co.* 401 U.S. 424, 1971.

[38] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *IEEE European Symposium on Security and Privacy*, pages 401–416, 2017.

[39] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[40] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation. https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/02/DevelopingPredictiveRiskModels-package_011618.pdf, 2017.

[41] Cédric Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

[42] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.

[43] Linda F Wightman and Henry Ramsey. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[44] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

[45] Karren D. Yang and Caroline Uhler. Scalable unbalanced optimal transport using generative adversarial networks. *arXiv preprint arXiv:1810.11447*, 2018.

[46] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[48] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[49] Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.