Operationalizing Individual Fairness with Pairwise Fair Representations

Preethi Lahoti
Max Planck Institute for Informatics
Saarland Informatics Campus
Saarbrücken, Germany
plahoti@mpi-inf.mpg.de

Krishna P. Gummadi
Max Planck Institute for Software Systems
Saarland Informatics Campus
Saarbrücken, Germany
gummadi@mpi-sws.org

Gerhard Weikum
Max Planck Institute for Informatics
Saarland Informatics Campus
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

We revisit the notion of individual fairness proposed by Dwork et al. A central challenge in operationalizing their approach is the difficulty in eliciting a human specification of a similarity metric. In this paper, we propose an operationalization of individual fairness that does not rely on a human specification of a distance metric. Instead, we propose novel approaches to elicit and leverage side-information on equally deserving individuals to counter subordination between social groups. We model this knowledge as a fairness graph, and learn a unified Pairwise Fair Representation (PFR) of the data that captures both data-driven similarity between individuals and the pairwise side-information in fairness graph. We elicit fairness judgments from a variety of sources, including humans judgments for two real-world datasets on recidivism prediction (COMPAS) and violent neighborhood prediction (Crime & Communities). Our experiments show that the PFR model for operationalizing individual fairness is practically viable.

1. INTRODUCTION

1.1 Motivation

Machine learning based prediction and ranking models are playing an increasing role in decision making scenarios that affect human lives. Examples include loan approval decisions in banking, candidate rankings in employment, welfare benefit determination in social services, and recidivism risk prediction in criminal justice. The societal impact of these algorithmic decisions have raised concerns about their fairness [3, 13], and recent research has started to investigate how to incorporate formalized notions of fairness into machine prediction models (e.g., [14, 20, 24, 22, 37]).

Individual vs Group Fairness: The fairness notions explored by the bulk of the works can be broadly categorized as targeting either *group fairness* [32, 16] or *individual fairness* [14]. Group fairness notions attempt to ensure that members of all protected groups in the population (e.g., based on demographic attributes like gender or race) receive their "fair share of beneficial outcomes" in a downstream task. To this end, one or more *protected attributes* and respective values are specified, and given special treatment in machine learning models. Numerous operationalizations of group fairness have been proposed and evaluated including demographic parity [16], equality of opportunity [20], equalized odds [20], and envy-free group fairness [36]. These operationalizations differ in the measures used to quantify a group's "fair share of beneficial outcomes" as well as the mechanisms used to optimize for the fairness measures.

While effective at countering group-based discrimination in decision outcomes, group fairness notions do not address unfairness in outcomes at the level of individual users. For instance, it is natural for individuals to compare their outcomes with those of others with similar qualifications (independently of their group membership) and perceive any differences in outcomes amongst individuals with similar standing as unfair.

Individual Fairness: In their seminal work [14], Dwork et al. introduced a powerful notion of fairness called individual fairness, which states that "similar individuals should be treated similarly". In the original form of individual fairness introduced in [14], the authors envisioned that a task-specific similarity metric would be provided by human experts that captures the similarity between individuals (e.g., "a student who studies at University W and has a GPA X is similar to another student who studies at University Y and has GPA Z"). The individual fairness notion stipulates that individuals who are deemed similar according to this *task-specific similarity metric* should receive similar outcomes. Operationalizing this strong notion of fairness can help in avoiding unfairness at an individual level.

However, eliciting such a quantitative measure of similarity from humans has been the most challenging aspect of the individual fairness framework, and little progress has been made on this open problem. Two noteworthy subsequent works on individual fairness are [39] and [29], wherein the authors operationalize a simplified notion of similarity metric. Concretely, they assume a distance metric (similarity metric) such as a *weighted* euclidean distance over a feature space of data atttributes, and aim to learn *fair feature weights* for this distance metric. This simplification of the individual fairness notion largely limits the scope of the original idea of [14]: "... a (near ground-truth) approximation agreed upon by the society of the extent to which two individuals are deemed similar with respect to the task ...".

In this work we revisit the original notion of individual fairness. There are two main challenges in its operationalization: First, it is very difficult, if not impossible for humans to come up with a precise quantitative similarity metric that can be used to measure "who is similar to whom". Second, even if we assume that humans are capable of giving a precise similarity metric, it is still challenging for experts to model subjective side-information such as "who should be

treated similar to whom" in the form of a similarity metric.

Examples: The challenge is illustrated by two scenarios:

- Consider the task of selecting researchers for academic jobs. Due to the difference in publication culture of various communities, the citation counts of *successful* researchers in programming language are known to be typically lower than that of *successful* machine learning researchers. An expert recruiter might have the background information for fair selection that "an ML researcher with high citations is similarly strong and thus equally deserving as a PL researcher with relatively lower citations". It is all but easy to specify this background knowledge in the form of a similarity metric.
- Consider the task of selecting students for Graduate School in the US. It is well known that SAT tests can be taken multiple times, and only the best score is reported for admissions. Further, each attempt to re-take the SAT test comes at a financial cost. Due to complex interplay of historical subordination and social circumstances, it is known that, on average, SAT scores for African-American students are lower than for white students [7]. Keeping anti-subordination in mind, a fairness expert might deem an African-American student with a relatively lower SAT score to be similar to and equally deserving as a white student with a slightly higher score. Once again, it is not easy to model this information as a quantitative similarity metric.

Research Questions: We address the following research questions in this paper.

- [RQ1] How to elicit and model various kinds of background information on individual fairness?
- [RQ2] How to encode this background information, such that downstream tasks can make use of it for data-driven predictions and decision making?

1.2 Approach

[RQ1] From Distance Metric to Fairness Graph.

Key Idea: It is difficult, if not impossible, for human experts to judge "the extent to which two individuals are similar", much less formulate a precise *similarity metric*. In this paper, we posit that it is much easier for experts to make pairwise judgments about who is equally deserving and should be treated similar to whom. An argument along these lines has been made by [21] in their work on subjective individual fairness.

We propose to capture these pairwise judgments as a *fairness graph*, G, with edges between pairs of individuals deemed similar with respect to the given task. In Section 3.2 we address some of the practical challenges that arise in eliciting pairwise judgments such as comparing individuals from diverse groups, and we present various methods to construct fairness graphs.

It is worth highlighting that we only need pairwise judgments for a small sample of individuals in the training data for the application task. Naturally, no human judgments are elicited for test data (unseen data). So once the prediction model for the application at hand has been learned, only the regular data attributes of individuals are needed.

[RQ2] Learning Pairwise Fair Representations.

Given a fairness graph G, the goal of an individually fair algorithm is to minimize the inconsistency (differences) in outcomes for pairs of individuals connected in graph G. Thus, every edge in graph G represents a fairness constraint that algorithms needs to satisfy. In Section 3, we propose a model called *PFR* (for Pairwise Fair Representations), that learns a new data representation with the aim of preserving the utility of the input feature space (i.e., retaining as much information of the input as possible), while incorporating the individual fairness constraints captured by the fairness graph G.

Specifically, *PFR* aims to learn a latent data representation that preserves the local neighborhoods in the input data space, while ensuring that individuals connected in the fairness graph are mapped to nearby points in the learned representation. Since local neighborhoods in the learned representation capture individual fairness, once a fair representation is learned, any out-of-the-box downstream predictor can be directly applied. *PFR* takes as input

- data records for individuals in the form of a feature matrix X for training a predictor, and
- a (sparse) fairness graph G that captures pairwise similarity for a small sample of individuals in training data.

The output of *PFR* is a mapping from the input feature space to the new representation space that can be applied to data records of novel unseen individuals.

1.3 Contribution

The key contributions of this paper are:

- A practically viable operationalization of the individual fairness paradigm that overcomes the challenge of human specification of a distance metric, by eliciting easier and more intuitive forms of human judgments.
- Novel methods for transforming such human judgments into pairwise constraints in a fairness graph G.
- A mathematical optimization model and representation learning method, called *PFR*, that combines the input data X and the fairness graph G into a unified representation by learning a latent model with graph embedding.
- Demonstrating the effectiveness of our approach at achieving both individual and group fairness using comprehensive experiments with synthetic as well as real-life data on recidivism prediction (Compas) and violent neighborhoods prediction (Crime and Communities).

2. RELATED WORK

Operationalizing fairness notions: Prior works on algorithmic fairness explore two broad families of fairness notions: group fairness and individual fairness.

Group fairness: A majority of the literature on fair learning has focused on group fairness. For instance, the group fairness notion of disparate impact or demographic parity in its various forms [8, 23, 32, 14] requires equality of beneficial outcome prediction rates between different socially salient groups. Approaches to achieve group fairness include debiasing the input data via data perturbation, re-sampling, modifying the value of protected attribute/class labels [33, 23, 32, 16] as well as incorporating demographic parity as an additional constraint in the objective function of machine learning models [25, 8, 38]. Another popular notion of group fairness is disparate mistreatment or equalized odds that aims to achieve equality of prediction error rates between groups [20, 37]. Similar approaches to achieve group fairness have been proposed for other tasks such as fair ranking

[4, 15, 9], fair set selection and clustering [10, 35] Recently, several researchers have highlighted the inherent incompatibility between different notions of group fairness and the inherent trade-offs when attempting to achieve them simultaneously [28, 11, 17, 12].

Individual fairness: Despite its appeal, few works have investigated individual fairness. The central challenge in operationalizing individual fairness has been to specify a similarity metric that captures which individuals deserve similar treatment. Some recent works use the objective of the learning algorithm itself to implicitly define the similarity metric [34, 5, 26]. For instance, when learning a classifier, these works would use the class labels in the training data or predicted class labels to measure similarity. However, fairness notions are meant to target addressing societal inequities that are not captured in the training data (with potentially biased labels and missing features). In such scenarios, the fairness objectives are in conflict with learning objectives.

In this work, we assume that human experts with background knowledge of past societal unfairness and future societal goals could provide coarse-grained judgments on whether pairs of individuals deserve similar outcomes. [18] similarly assumes "a regulator who knows fairness when he sees it, but cannot enunciate a quantitative fairness metric over individuals", but it considers individual fairness in a restricted online learning setting.

Bridging individual and group fairness: Approaches to enforcing group fairness have mostly ignored individual fairness and vice versa. Intuitively, the concept of individual fairness appears sufficiently strong and broad to subsume group fairness. In this work, we show that by appropriately constraining outcomes for pairs of individuals belonging to different groups, we are able to achieve group fairness to a large degree. Our approach is loosely inspired by the idea of "fair affirmative action" in [14], that attempts to achieve both statistical parity and individual fairness. However, unlike [14] that assumes a hypothetical distance metric, our approach is based on a fairness graph that can be constructed in practice.

Learning pairwise fair representations: In terms of our technical machinery, the closest prior work is [39, 29] that aim to learn new representations for individuals that "retain as much information in the input feature space as possible, while losing any information that can identify individuals' protected group membership". Our approach aims to learn new representations for individuals that retain the input data to the best possible extent, while clustering equally deserving individuals as closely as possible. Like [39, 29] our method can be used to find representations for new individuals not seen in the training data. This ability crucially distinguishes our problem from the classical metric labeling problem [27], where the goal is to classify a given set of individuals with pairwise relationships (i.e., costs for being assigned different labels).

Finally, the core optimization problem we formulate relates to graph embedding and representation learning [19]. The aim of graph embedding approaches is to a learn a representation for the nodes in the graph encoding the edges between nodes as well as the attributes of the nodes [30, 1]. Similarly, we wish to learn a representation encoding both the features of individuals as well as their interconnecting edges in the fairness graph.

3. MODEL

3.1 Notation

- X is an input data matrix of n data records and m numerical or categorical attributes. We use X to denote both the matrix and the population of individuals x_i:
$$X = [x_1, x_2, x_3, \cdots x_n] \in R^{m \times n}$$
- Z is a low-rank representation of X in a d-dimensional space where $d \ll m$.
$$Z = [z_1, z_2, z_3, \cdots z_n] \in R^{d \times n}$$
- S is a random variable representing the values that the protected-group attribute can take. We assume a single attribute in this role; if there are multiple attributes that require fair-share protection, we simply combine them into one. We allow more than two values for this attribute, going beyond the usual binary model (e.g., gender = male or female, race = white or others). $X_s \subset X$ denotes the subset of individuals in X who are members of group $s \in S$.
- W^X is the adjacency matrix of a k-nearest-neighbor graph over the input space X given by:
$$W^X_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x_i}-\mathbf{x_j}\|^2}{t}\right) & \text{, if } x_i \in N_p(x_j) \text{ or } x_j \in N_p(x_i) \\ 0 & \text{, otherwise} \end{cases}$$
where $N_p(x_i)$ denotes the set of p nearest neighbors of x_i in euclidean space (excluding the protected attributes), and t is a scalar hyper-parameter.
- W^F is the adjacency matrix of the fairness graph G whose nodes are individuals and whose edges are connections between individuals that are equally deserving and must be treated similarly.

3.2 From Distance Metric to Fairness Graph

In this section we address the question of how to elicit side-information on individual fairness and model it as a fairness graph G and its corresponding adjacency matrix as W^F. The key idea of our approach is rooted in the following observations:

- Humans have a strong intuition about whether two individuals are similar or not. However, it is difficult for humans to specify a quantitative *similarity metric*.
- In contrast, it is more natural to make other forms of judgments such as (i)"Is A similar to B with respect to the given task?", or (ii)"How suitable is A for the given task (e.g., on a Likert scale)".
- However, these kinds of judgments are difficult to elicit when the pairs of individuals belong to diverse, incomparable groups. In such cases, it is easier for humans to compare individuals within the same group, as opposed to comparing individuals between groups. Pairwise judgements can be beneficial even if they are available only sparsely, that is, for samples of pairs.

Next, we present two models for constructing fairness graphs, which overcome the outlined difficulties via

(i) eliciting (binary) pairwise judgments of individuals who should be treated similarly, or grouping individuals into equivalence classes (see Subsection 3.2.1) and

(ii) eliciting within-group rankings of individuals and connecting individuals across groups who fall within the same quantiles of the per-group distributions (see Subsection 3.2.2).

3.2.1 Fairness Graph for Comparable Individuals

The most direct way to create a fairness graph is to elicit (binary) pairwise similarity judgments about a small sample of individuals in the input data, and to create a graph W^F such that there is an edge between two individuals if they are deemed similarly qualified for a certain task (e.g., being invited for job interviews).

Another alternative is to elicit judgments that map individuals into discrete equivalence classes. Given a number of such judgments for a sample of individuals in the input dataset, we can construct a fairness graph W_F by creating an edge between two individuals if they belong to the same equivalence class.

Definition 1. (Equivalence Class Graph) *Let $[x_i]$ denote the equivalence class of an element $x_i \in X$. We construct an undirected graph W^F associated to X, where the nodes of the graph are the elements of X, and two nodes x_i and x_j are connected if and only if $[x_i] = [x_j]$.*

The fairness graph built from such equivalence classes identifies equally deserving individuals – a valuable asset for learning a fair data representation. Note that the graph may be sparse, if information on equivalence can be obtained merely for sampled representatives.

3.2.2 Fairness Graph for Incomparable Individuals

However, at times, our individuals are from diverse and incomparable groups. In such cases, it is difficult if not infeasible to ask humans for pairwise judgments about individuals *across groups*. Even with the best intentions of being fair, human evaluators may be misguided by wide-spread bias. If we can elicit a ranked ordering of individuals per-group, and pool them into quantiles (e.g., the top-10-percent), then one could assume that individuals from different groups who belong to the same quantile in their respective rankings, are similar to each other. Arguments along these lines have been made also by [26] in their notion of meritocratic fairness.

Specifically, our idea is to first obtain within-group rankings of individuals (e.g., rank men and women separately) based on their suitability for the decision task at hand, and then construct a between-group fairness graph by linking all individuals ranked in the same k^{th} quantile across the different groups (e.g., link PL researcher and ML researcher who are similarly ranked in their own groups). The relative rankings of individuals within a group, whether they are obtained from human judgments or from secondary data sources, are less prone to be influenced by discriminatory (group-based) biases.

Formally, given (X_s, Y_s) for all $s \in S$, where Y_s is a random variable depicting the ranked position of individuals in X_s. We construct a *between-group quantile graph* using Definitions 2 and 3.

Definition 2. (k-th quantile) *Given a random variable Y, the k-th quantile Q_k is that value of y in the range of Y, denoted y_k, for which the probability of having a value less than or equal to y is k.*

$$Q(k) = \{y : Pr(Y \leq y) = k\} \quad \text{where} \quad 0 < k < 1 \qquad (1)$$

For the non-continuous behavior of discrete variables, we would add appropriate ceil functions to the definition, but we skip this technicality.

Definition 3. (Between-group quantile graph) *Let $X_s^k \subset X$ denote the subset of individuals who belong to group $s \in S$ and whose scores lie in the k-th quantile. We can construct a multipartite graph W^F whose edges are given by:*

$$W_{ij}^F = \begin{cases} 1 & , \text{ if } x_i \in X_s^k \quad \text{and} \quad x_j \in X_{s'}^k \quad , \quad s \neq s' \\ 0 & , \text{ otherwise} \end{cases} \qquad (2)$$

That is, there exists an edge between a pair of individuals $\{x_i, x_j\} \in X$ if x_i and x_j have different group memberships and their scores $\{y_i, y_j\}$ lie in the same quantile. For the case of two groups (e.g., gender is male or female), the graph is a bipartite graph.

This model of creating between-group quantile graphs is general enough to consider any kind of per-group ranked judgment. Therefore, this model is not necessarily limited to legally protected groups (e.g., gender, race), it can be used for any socially salient groups that are incomparable for the given task (e.g., machine learning vs. programming language researchers). Note again that the pairwise judgements may be sparse, if such information is obtained only for sampled representatives.

3.3 Learning Pairwise Fair Representations

In this section we address the question [RQ2]: How to encode the background information such that downstream tasks can make use of it for the decision making?

3.3.1 Objective Function

In fair machine learning, such as fair classification models, the objective usually is to maximize the classifier accuracy (or some other quality metric) while satisfying constraints on group fairness statistics such as parity. For learning fair data representations that can be used in any downstream application – classifiers or regression models with varying target variables unknown at learning time – the objective needs to be generalized accordingly. To this end, the *PFR* model aims to combine the utility of the learned representation and, at the same time, preserve the information from the pairwise fairness graph. Starting with matrix X of n data records $x_1 \ldots x_n$ and m numeric or categorial attributes, *PFR* computes a lower-dimensional latent matrix Z of n records each with $d < m$ values.

Utility is cast into a notion of data loss. In matrix factorization, this usually means to minimize the error when using Z to reconstruct an approximation of X. In our approach, we do not adopt this standard error, but instead cast the data loss into a measure for how well the neighborhoods of data records are preserved when mapping the attribute space X into the latent representation Z.

Reflecting the fairness graph in the learner's optimization for Z is a demanding and a priori open problem. Our solution *PFR* casts this issue into a graph embedding that is incorporated into the overall objective function. The following discusses the technical details of *PFR* 's optimization.

Preserving the input data: For each data record x_i in the input space, we consider the set $N_p(x_i)$ of its p nearest neighbors with regard to the distance defined by the kernel function given by W^X_{ij}. For all points x_j within $N_p(x_i)$, we want the corresponding latent representations z_j to be close to the representation z_i, in terms of their L2-norm distance. This is formalized by the *Loss in* W^X - $Loss_X$.

$$Loss_X = \sum_{i,j=1}^{N} \|z_i - z_j\|^2 W^X_{ij} \tag{3}$$

Note that this objective requires only local neighborhoods in X to be preserved in the transformed space. We disregard data points outside of p-neighborhoods. This relaxation increases the feasible solution space for the dimensionality reduction.

Learning a fair graph embedding: Given a fairness graph W^F, the goal for Z is to preserve neighborhood properties in W^F. In contrast to $Loss_X$, however, we do not need any distance metric here, but can directly leverage the fairness graph. If two data points x_i, x_j are connected in W^F, we aim to map them to representations z_i and z_j close to each other. This is formalized by the *Loss in* W^F - $Loss_F$.

$$Loss_F = \sum_{i,j=1}^{N} \|z_i - z_j\|^2 W^F_{ij} \tag{4}$$

Intuitively, for data points connected in W^F, we add a penalty when their representations are far apart in Z.

Combined objective: Based on the above considerations, a fair representation Z is computed by minimizing the combined objectives of Equations 3 and 4. The parameter γ weighs the importance tradeoff between W^X and W^F. As γ increases influence of the fairness graph W^F increases. An additional ortho-normality constraint on Z is imposed to avoid trivial results. The trivial result being that all the datapoints are mapped to same point.

$$\text{Minimize} \quad (1-\gamma) \sum_{i,j=1}^{N} \|z_i - z_j\|^2 W^X_{ij} + \gamma \sum_{i,j=1}^{N} \|z_i - z_j\|^2 W^F_{ij}$$
$$\text{subject to} \quad Z^T Z = I \tag{5}$$

3.3.2 Equivalence to Trace Optimization Problem

Next, we show that the optimization problem in Equation 5 can be transformed and solved as an equivalent eigenvector problem. To do so, we assume that the learnt representation Z is a linear transformation of X given by $Z = V^T X$.

We start by showing that minimizing $\|z_i - z_j\|^2 W_{ij}$ is equivalent to minimizing the trace $Tr(V^T X L X^T V)$. Here we use W to denote W^X or W^F, as the following mathematical derivation holds for both of them analogously:

$$\begin{aligned} Loss &= \sum_{i,j=1}^{n} \|z_i - z_j\|^2 W_{ij} \\ &= \sum_{i,j=1}^{n} Tr((z_i - z_j)^T (z_i - z_j)) W_{ij} \\ &= 2 \cdot Tr(\sum_{i,j=1}^{n} z_i^T z_i D_{ii} - \sum_{i,j=1}^{n} z_i^T z_j W_{ij}) \\ &= 2 \cdot Tr(V^T X L X^T V) \end{aligned}$$

where $Tr(.)$ denotes the trace of a matrix, D is a diagonal matrix whose entries are column sums of W, and $L = D - W$ is the graph Laplacian constructed from matrix W. Analogous to L, we use L^X to denote graph laplacian of W^X, and L^F to denote graph laplacian of W^F.

3.3.3 Optimization Problem

Considering the results of Subsection 3.3.2, we can transform the above combined objective in Equation 5 into a trace optimization problem as follows:

$$\begin{aligned} &\text{Minimize J(V)} = \quad Tr\{V^T X((1-\gamma)L^X + \gamma L^F) X^T V\} \\ &\text{subject to} \quad V^T V = I \end{aligned} \tag{6}$$

We aim to learn an $m \times d$ matrix V such that for each input vector $x_i \in X$, we have the low-dimensional representation $z_i = V^T x_i$, where $z_i \in Z$ is the mapping of the data point x_i on to the learned basis V. The objective function is subjected to the constraint $V^T V = I$ to eliminate trivial solutions.

Applying Lagrangian multipliers, we can formulate the trace optimization problem in Equation 6 as an eigenvector problem

$$X((1-\gamma)L^X + \gamma L^F) X^T \mathbf{v_i} = \lambda \mathbf{v_i} \tag{7}$$

It follows that the columns of optimal V are the eigenvectors corresponding to d smallest eigenvalues denoted by $V = [\mathbf{v_1 v_2 v_3} \cdots \mathbf{v_d}]$, and γ is a regularization hyper-parameter. Finally, the d-dimensional representation of input X is given by $Z = V^T X$.

Implementation: The above standard eigenvalue problem for symmetric matrices can be solved in $O(n^3)$ using iterative algorithms. In our implementation we use the standard eigenvalue solver implementation from scipy.linalg.lapack python library [2].

3.3.4 Kernelized Variants of PFR

In this paper, we restrict ourselves to assume that the representation Z is a linear transformation of X given by $Z = V^T X$. However, *PFR* can be generalized to a non-linear setting by replacing X with a non-linear mapping $\phi(X)$ and then performing *PFR* on the outputs of ϕ (potentially in a higher-dimensional space).

For this purpose, assume that $Z = V^T \Phi(X)$ and $V = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$ with a Mercer kernel matrix K where $K_{i,j} = k(x_i, x_i) = \Phi(x_i)^T \Phi(x_j)$. We can show that the trace optimization problem in Equation 7 can be generalized to this non-linear kernel setting, and it can be conveniently solved by working with Mercer kernels without having to compute $\Phi(X)$. We arrive at the following generalized optimization problem.

$$K((1-\gamma)L^X + \gamma L^F) K \alpha_\mathbf{i} = \lambda \alpha_\mathbf{i} \tag{8}$$

Analogously to the solution of Equation 7, the solution to the *kernel PFR* is given by $A = [\alpha_\mathbf{1} \alpha_\mathbf{2} \alpha_\mathbf{3} \cdots \alpha_\mathbf{d}]$ where $\alpha_1 \cdots \alpha_d$ are the d smallest eigenvectors. Finally, the learned representation of X is given by $Z = V^T \Phi(X) = A^T K$.

In this paper we present results only for *linear PFR*, leaving the investigation of *kernel PFR* for future work.

4. EXPERIMENTS

This section reports on experiments with synthetic and real-life datasets. We compare a variety of fairness-enhancing methods on a binary classification task as a downstream application. The following key questions are addressed:

- [Q1] What do the learned representations look like?
- [Q2] What is the effect on individual fairness?
- [Q3] What is the influence on the trade-off between fairness and utility?
- [Q4] What is the influence on group fairness?
- [Q5] What is the influence of the PFR hyper-parameter γ on individual fairness and utility?

4.1 Experimental Setup

Baselines: We compare the performance of *PFR* with the following methods

- *Original representation:* a naive representation of the input dataset wherein the protected attributes are masked.
- *iFair [29]*: an unsupervised representation learning method, which optimizes for two objectives: (i) individual fairness in W^X, and (ii) obfuscating protected attributes.
- *LFR [39]*: a supervised representation learning method, which optimizes for three objectives: (i) accuracy (ii) individual fairness in W^X and (iii) demographic parity.
- *Hardt [20]*: a post-processing method that aims to minimize the difference in error rates between groups by optimizing for the group-fairness measure *EqOdd* (Equality of Odds).
- *PFR*: Our unsupervised representation learning method that optimizes for two objectives (i) individual fairness as per W^F and (ii) individual fairness as per W^X.

Augmenting baselines: In order to ensure fair comparison we compare *PFR* with augmented versions of all methods (named with *suffix* +). In the augmented version, we give each method an advantage by enhancing it with the information in the fairness graph W^F. Since none of the methods can be naturally extended to incorporate the fairness graph as it is, we make our best attempt at modeling the side-information that is used to construct W^F as a numerical feature, and include this as an additional attribute in the respective training data.

Hyper-parameter tuning: We use the same experimental setup and hyper-parameter tuning techniques for all methods. Each dataset is split into separate training and test sets. On the training set, we perform 5-fold cross-validation (i.e., splitting into 4 folds for training and 1 for validation) to find the best hyper-parameters for each model via *grid search*. Once hyper-parameters are tuned, we use the independent test set to measure performance.

Downstream-task: We compare all the methods on downstream classification tasks for a synthetic dataset (US university admission) and two real-world datasets (recidivism prediction and violent neighbourhood prediction). Table 1 gives details of experimental settings and statistics for each dataset, including base-rate (fraction of samples belonging to the positive class, for both the protected group and its complement). Specific details of each dataset are discussed in later subsections. In all the experiments, the representation learning approaches are followed by a out-of-the-box logistic regression classifier trained on the corresponding representations.

Dataset	$\|X\|$	$\|X_{s=0}\|$	$\|X_{s=1}\|$	Base-rate (s = 0)	Base-rate (s = 1)	Classification task	Protected attribute
Synthetic	600	300	300	0.51	0.48	Is successful	Race
Crime	1993	1423	570	0.35	0.86	Is violent	Race
Compas	8803	4218	4585	0.41	0.55	Is rearrested	Race

Table 1: Experimental setting and statistics of the datasets.

Evaluation Measures:

- **Utility** is measured as AUC (area under the ROC curve).
- **Individual Fairness** is measured as the *consistency* of outcomes between individuals who are similar to each other. We report consistency values as per both the similarity graphs, W^X and W^F.

$$Consistency = 1 - \frac{\sum_i \sum_j |\hat{y}_i - \hat{y}_j| \cdot W_{ij}}{\sum_i \sum_j W_{ij}} \quad \forall \quad i \neq j$$

- **Group Fairness**
  - **Disparate Mistreatment (aka. Equality of Odds):** A binary classifier avoids disparate mistreatment if the group-wise error rates are the same across all groups. In our experiments, we report per-group false positive rate (FPR) and false negative rate (FNR).
  - **Disparate Impact (aka. Demographic Parity):** A binary classifier avoids disparate impact if the rate of positive predictions is the same across all groups.

$$P(\hat{Y} = 1|s = 0) = P(\hat{Y} = 1|s = 1) \tag{9}$$

In our experiments, we report per-group rate of positive predictions.

4.2 Analysis on Synthetic Data

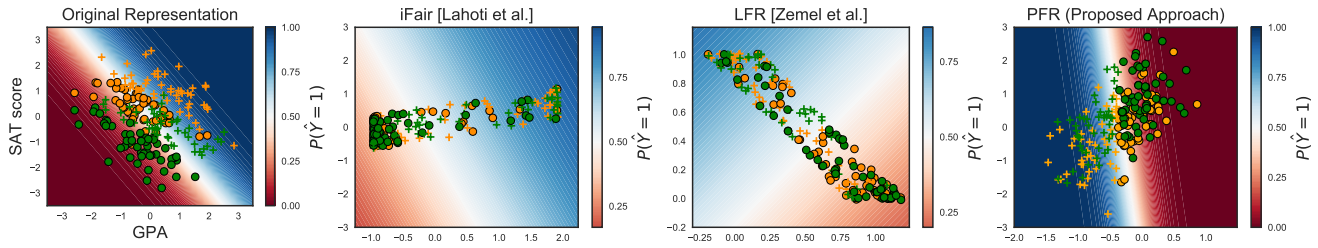
We simulate the US graduate admissions scenario of Section 1.1 where the population consists of two groups $s = 0$ or 1. For each candidate we know their score on the SAT entrance exam – *SAT score* – and average grades – *GPA*. Our task is it predict the ability of a candidate to complete graduate school (binary classification).

It is known that the SAT test can be taken multiple times, and only the best score is reported for admissions. Further, each attempt to re-take the SAT comes at a financial cost. Suppose we live in a society where group membership has a high correlation with individuals with affluent and educated parents. This would imply that one group has access to expensive tutoring for the SAT and can take the test multiple times, which leads to increased SAT scores for one group.

4.2.1 Synthetic dataset

We simulate this scenario by generating data for two populations X_0 and X_1 such that the two groups have similar distributions for GPA, but one group has slightly higher values for SAT score than the other. We generate synthetic data where features value for GPA and SAT score for group $X_{s=0}$ were drawn from $\mathcal{N}([100, 110], [25, -5; -5, 25])$ and for group $X_{s=1}$ from $\mathcal{N}([100, 100], [25, -5; -5, 25])$.

Despite average SAT scores for group $X_{s=0}$ being higher than for the protected group $X_{s=1}$, we assume that the ability to complete graduate school is the same for both groups; that is, members of $X_{s=0}$ and $X_{s=1}$ are equally deserving if we adjust their SAT scores. To implement this scenario, we set the *true* class label for group $X_{s=0}$ to positive (1) if GPA



(a) Original Representation (b) iFair Representation (c) LFR Representation (d) PFR Representation

Figure 1: Comparison of (a) Original representation (b) iFair (c)LFR and (d)PFR representations on a synthetic dataset. Original representation is standardized to zero mean and unit variance.

+ SAT ≥ 210 and for group $X_{s=1}$ as positive (1) if GPA + SAT ≥ 200. Figure 1a visualizes the generated dataset. The colors depict the membership to groups (S): S = 0 (orange) and S = 1 (green). The markers denote *true* class labels Y = 1 (marker +) and Y = 0 (marker o).

Fairness Graph W^F: We simulate human judgments of fairness by connecting similarly deserving candidates of one group to another. That is, we add edges between "orange plus" and "green plus" and between "orange o" and "green o", respectively. Concretely, we generate within-group rankings of candidates for the two groups separately using the prediction probability of a standard logistic regression model, and use these rankings to construct a between-group quantile graph as per Definitions 2 and 3.

4.2.2 *Results on Synthetic Dataset*

[Q1] What do the learned representations look like? In this subsection we inspect the original representations and contrast them with learned representations via *iFair* [29], *LFR* [39], and our proposed model *PFR*. Figure 1 visualizes the original dataset and the learned representations for each of the models with the number of latent dimensions set to $d = 2$ during the learning. The contour plots in (b), (c) and (d) denote the decision boundaries of logistic regression classifiers trained on the respective learned representations. Blue color corresponds to positive classification, red to negative; the more intensive the color, the higher or lower the score of the classifier. We observe several interesting points:

- First, in the original data, the two groups are separated from each other: *green* and *orange* datapoints are relatively far apart. Further, the deserving candidates of one group are relatively far away from the deserving candidates of the other group. That is, "green plus" are far from "orange plus", illustrating the inherent unfairness in the original data.
- In contrast, for all three representation learning techniques – *iFair*, *LFR* and *PFR* – the *green* and *orange* data points are well-mixed. This shows that these representations are able to make protected and non-protected group members indistinguishable from each other – a key property towards fairness.
- The major difference between the learned representations is that, *PFR* succeeds in mapping the deserving candidates of one group close to the deserving candidates of the other group (i.e., "green plus" are close to "orange plus"). Neither iFair nor LFR can achieve this desired effect, to the same extent.

[Q2] Effect on Individual Fairness: Figure 2 shows the best achievable trade-off between utility and the two notions of individual fairness.

- Individual fairness regarding W^X: We observe that *iFair*, *LFR* have similar performance as *PFR* for *consistency* (W^X), but with a much lower *AUC*. This is in line with our observations on the learned representations. Since the protected and non-protected groups are made indistinguishable in the learned representations, the classifier yields similar outcomes to similar individuals irrespective of their group membership, hence leading to high consistency. In contrast, the high value of *consistency* for the *Original* model is because of the trivial effect of giving the same (but incorrect) prediction to all nearby individuals.
- Individual fairness regarding W^F: We observe that *PFR* significantly outperforms all competitors in terms of *consistency* (W^F). This follows from the observation that, unlike *Original*, *iFair* and *LFR* representations, *PFR* maps similarly deserving individuals close to each other in its latent space.

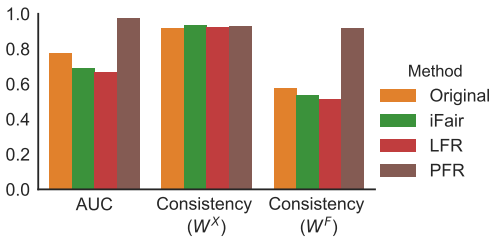
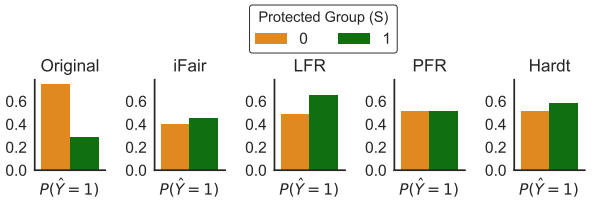


Figure 2: Comparison of Utility vs Individual Fairness trade-off across methods. Higher values are better.

[Q3] Trade-off between Utility and Fairness: The AUC bars in Figure 2 show the results on classifier utility for the different methods under comparison.

- Utility (AUC): PFR achives by far the best AUC. While this may surprise on first glance, it is indeed an expected outcome. The fairness edges in W_F reflect the true deservingness for both groups, which helps the classifier's accuracy. The other learned representations exhibit a small loss in utility compared to the Original data, as they trade off fairness for utility.

[Q4] Influence on Group Fairness: In addition to Original, iFair, LFR and PFR, we include the *Hardt* model in the comparison here, as it is widely viewed as the state-of-the-art method for group fairness.



(a) Difference in rates of positive prediction

(b) Difference in error rates (FPR and FNR)

Figure 3: Difference in (a) rate of positive predictions and (b) error rates between protected and non-protected groups

Figure 3a shows the per-group positive predictions rates, and Figure 3b shows the per-group error rates. The smaller the difference in the values of the two groups, the higher the group fairness. We make the following interesting observations:

- Disparate Impact (Figure 3a): The *Original* data exhibits a substantial difference in the per-group positive predictions rates. A classifier trained for AUC favors the *orange* group. In contrast, *iFair*, *LFR* and *PFR* have the *orange* and *green* data points well-mixed, and this way achieve nearly equal rates for both groups. Likewise *Hardt* has the same desired effect.
- Disparate Mistreatment (Figure 3b): For this measure (aka. Equality of Odds), we also observe the strong bias of the *Original* data, and the degrees of countering it by the learned representations. The latter exhibit notable differences, though. *iFair* and *LFR* balance the error rates across groups fairly well, but still have fairly high error rates, indicating their loss on utility. *PFR* and *Hardt* have well balanced error rates and generally lower error. For *Hardt*, this is the expected effect, as it is optimized for the very goal of Equality of Odds. *PFR* achieves the best balance and lowest error rates, which is remarkable as its objective function does not directly consider group fairness. Again, the effect is explained by *PFR* succeding in mapping equally deserving individuals from both groups to close proximity in its latent space.

[Q5] Influence of Hyper-Parameter γ: *PFR* aims to preserve proximity for both W^F and W^X, where the hyper-parameter γ controls the relative influence of W^F and W^X.

- Individual Fairness: Figures 4a and 4b show the influence of γ on individual fairness as per W^F and W^X, respectively. As expected, as γ increases, the consistency as per W^F increases and the consistency as per W^X decreases. It is worth highlighting that the extent of this trade-off depends on the degree of conflict between the two graphs. If W^F contains judgements of equal deserv-

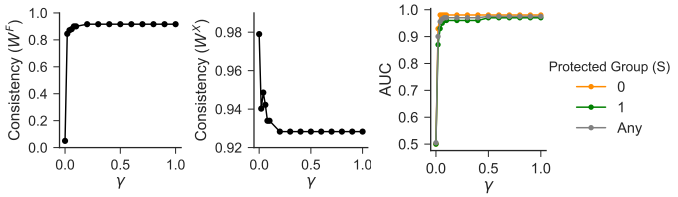


Figure 4: Influence of γ on (a)Individual fairness w.r.t W^F (b) Individual fairness w.r.t W^X and (c)Utility

ingness for data points that are far apart in the feature space, there is a natural conflict between the two notions of individual fairness. However, one may argue that this is the right approach from a human perspective to counter data-centric bias and unfairness (like in the student admission scenario).

- Utility: Figure 4c shows the influence of γ on the utility measure (AUC). As γ increases, the AUC of *PFR* increases. The improvement holds for both protected and non-protected groups. This gain in AUC is because the constraints in the fairness graph W^F are in line with ground-truth. So higher influence of W^F helps utility. If W^F were in tension with ground-truth labels, like when being motivated by *affirmative action* fairness, a drop in utility would be unavoidable. Again, one may argue that human judgement should overrule the data-only-centric decision making.

4.3 Experiments on Real-World Datasets

We evaluate the performance of *PFR* on the following two real world datasets

- *Crime & Communities* [31] is a dataset consisting of socio-economic (e.g., income), demographic (e.g., race), and law/policing data (e.g., patrolling) records for neighborhoods in the US. We set *isViolent* as target variable for a binary classification task. We consider the communities with majority population white as non-protected group and the rest as protected group.
- *Compas* data collected by ProPublica [3] contains criminal records comprising offenders' criminal histories and demographic features (gender, race, age etc.). We use the information on whether the offender was re-arrested as the target variable for binary classification. As protected attribute $s \in \{0,1\}$ we use race: African-American (1) vs. others (0).

4.3.1 Constructing the Fairness Graph W^F

Crime & Communities: We need to elicit pairwise judgments of similarity that model whether two neighborhoods are similar in terms of crime and safety. To this end, we collected human reviews on crime and safety for neighborhoods in the US from http://niche.com. The judgments are given in the form of 1-star to 5-star ratings by current and past residents of these neighborhoods. We aggregate the judgments and compute mean ratings for all neighborhoods. We were able to collect reviews for about 1500 (out of 2000) communities. W^F is then constructed by the technique of Subsection 3.2.1.

Although this kind of human input is subjective, the aggregation over many reviews lifts it to a level of inter-subjective

side-information reflecting social consensus by first-hand experience of people. Nevertheless, the fairness graph may be biased in favor of the African-american neighbourhoods, since residents tend to have positive perception of their neighborhood's safety.

Compas: We need to elicit pairwise judgments of similarity that model whether two individuals are similar in terms of deserving to be granted parole and not becoming re-arrested later. However, it is virtually impossible for a human judge to fairly compare people from the groups of *African-Americans* vs. *Others*, without imparting the historic bias (with much higher historic recidivism of the former group). So this is a case, where we need to elicit pairwise judgments between diverse and incomparable groups.

We posit that it is fair, though, to elicit *within-group* rankings of risk assessment for each of the two groups, to create edges between individuals who belong to the same risk quantile of their respective group. To this end, we use Northpointe's Compas decile scores [6] as background information about within-in group ranking. These *decile scores* are computed by an undisclosed commercial algorithm which takes as input official criminal history and interview/questionnaire answers to a variety of behavioral, social and economic questions (e.g., substance abuse, school history, family background etc.). The decile scores assigned by this algorithm are *within-group* scores and are not meant to be compared across groups. We take the *decile scores* and compute k^{th} quantiles for each group separately, to construct W^F by the technique of Subsection 3.2.2.

Note that this fairness graph has an implicit anti-subordination assumption. That is, it assumes that individuals in k-th risk quantile of one group are similar to the individuals in k-th quantile of other group - irrespective of their true risk.

Augmenting Baselines: For fair comparison with *PFR*, we augment all other methods (named with *suffix* +) by giving them access to the information in the fairness graph W^F. as additional numerical features in the respective training data. Note that this enhancement is only for training, as this side-information is not available for the test data. This is in line with how PFR uses the pairwise comparisons: its representation is learned from the training data, but at test time, only data attributes W^X are available.

4.3.2 Results on Crime & Communities Dataset

[Q2] Effect on Individual Fairness: Results on individual fairness and utility (AUC) are given in Figure 5. We observe that *PFR* outperforms all other methods on individual fairness regarding W^F. However, this gain for W^F comes at the cost of losing in individual fairness regarding W^X. So in this case, the pairwise input from human judges exhibits pronounced tension with the data-attributes input. Deciding which of these sources should take priority is a matter of application design.

[Q3] Trade-off between Utility and Fairness: The improvement in individual fairness regarding W^F comes with a drop in utility as shown by the AUC bars in Figure 5. This is because, unlike the case of the synthetic data in Subsection 4.2, the side-information fir the fairness graph W^F is not strongly aligned with the ground-truth for the classifier. The other methods benefit from the side-information in their augmented versions, but still exhibit the same fundamental trade-off between individual fairness and utility.

In terms of relative comparison, *LFR+* performs best: AUC close to that of *PFR* while achieving some of the best values for the two notions of individual fairness. However, *LFR+* exhibits weaknesses (and notably inferior performance) on group fairness, as discussed next.

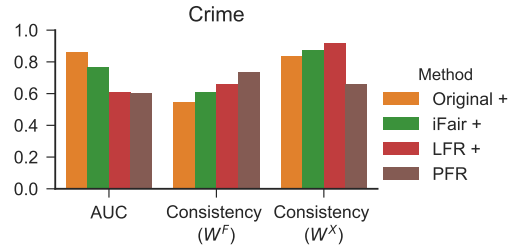
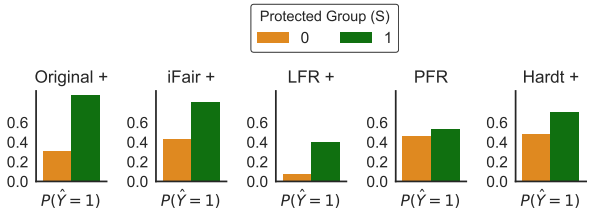
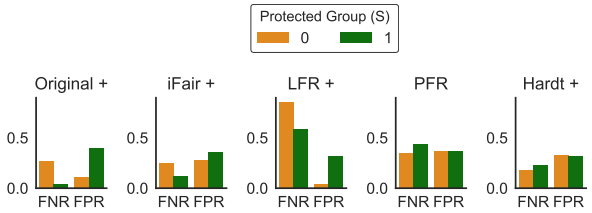


Figure 5: Crime & Communities Data: Utility vs. Individual Fairness (higher is better).



(a) Difference in rates of positive prediction



(b) Difference in error rates (FPR and FNR)

Figure 6: Crime & Communities Data: Difference between groups in (a) Rate of Positive Predictions and (b) Error Rates.

[Q4] Influence on Group Fairness Figure 6a shows the per-group positive prediction rates, and Figure 6b shows the per-group error rates. Smaller differences in the values between the two groups are preferable. The following observations are notable:

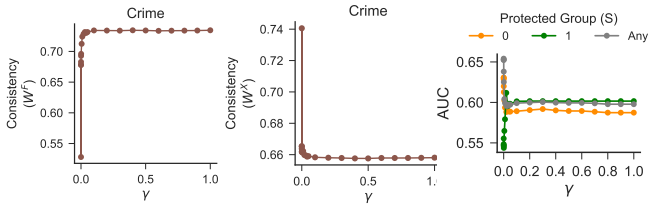
- Disparate Impact (aka. Demographic parity): *PFR* clearly outperforms all the methods by achieving near perfect balance (i.e., near-equal rates of positive predictions).
- Disparate Mistreatment (aka. Equality of Odds): *PFR* significantly outperforms all other methods on balancing the error rates of the two groups. Furthermore, it achieves nearly equal error rates comparable to the *Hardt* model, whose sole goal is to achieve equal error rates between groups via post-processing.

[Q5] Influence of Hyper-Parameter γ: Key points from the experiments are the following.

- Individual Fairness: Figure 7a and 7b show the influence of γ on individual fairness as per W^F and W^X, respectively. As expected, we observe that with increasing γ

the consistency with regard to W^F increases. Conversely, the consistency with regard to W^X decreases.

- Utility: Figure 7c shows the influence of γ on the AUC. With increasing γ, the influence of the pairwise constraints in W^F increases and the overall utility AUC for both groups together (S = Any) decreases. However, there is an improvement in AUC for the protected group, and the gap in AUC between the groups decreases. This results constitutes a clear case of how incorporating side-information on pairwise judgments can help in improving algorithmic decision making for historically disadvantaged groups.



(a) (b) (c)

Figure 7: Influence of γ on (a) Individual Fairness w.r.t. W^F, (b) Individual Fairness w.r.t. W^X and (c) Utility

4.3.3 *Results on Compas Dataset*

The results for the Compas dataset are mostly in line with the results for the synthetic data (in Subsection 4.2) and Crime & Communities datasets (in Subsection 4.3.2). Therefore, we report only briefly on them.

Utility vs. Individual Fairness: *PFR* performs similarly as the other representation learning methods in terms of utility and individual fairness, as shown in Figure 8.

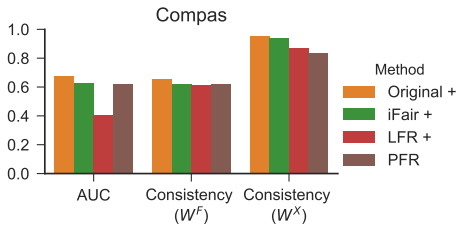
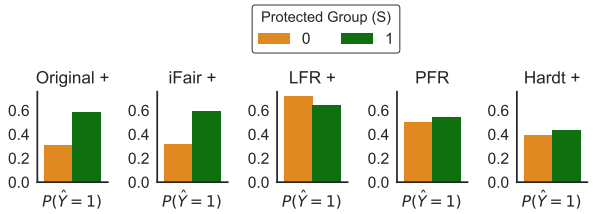


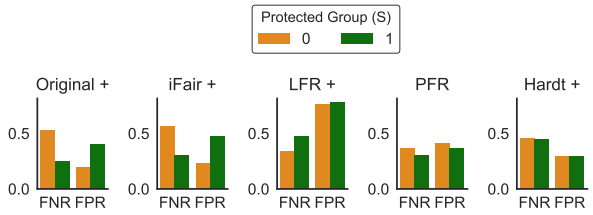
Figure 8: Compas dataset: Utility vs Individual-fairness. Higher values are better.

Group Fairness: However, *PFR* clearly outperforms all other methods on group fairness. It achieves near-equal rates of positive predictions as shown in Figure 9a, and near-equal error rates across groups as shown in Figure 9b. Again, *PFR*'s performance on group fairness is as good as that of *Hardt* which is solely designed for equalizing error rates by post-processing the classifier's outcomes.

Influence of Hyper-Parameter γ: Figures 10a and 10b show the same effects as observed for the other datasets: increasing γ helps consistency w.r.t. W^F and degrades consistency w.r.t. W^X. Likewise, Figure 10c confirms that higher γ hurts AUC over both groups together. However, as before, AUC for the protected group (S =1) improves and the gap in AUC between the two groups decreases when



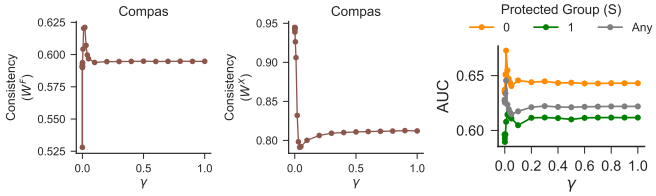
(a) Difference in rates of positive prediction



(b) Difference in error rates (FPR and FNR)

Figure 9: Compas Data: Difference between Groups in (a) Rate of Positive Predictions and (b) Error Rates.

γ is set higher. So the PFR way of incorporating pairwise judgements helps the protected group.



(a) (b) (c)

Figure 10: Influence of γ on (a) Individual Fairness w.r.t. W^F (b) Individual Fairness w.r.t. W^X and (c) Utility

4.4 Discussion and Lessons

PFR outperforms all other methods on individual fairness regarding W^F for an acceptable performance in AUC, even when these baselines are given the same side-information for their augmented version (suffixed +). The improvement in individual fairness in W^F comes at the expense of reducing individual fairness for W^X, an unavoidable trade-off if the two views of fairness – data attributes (W^X) and pairwise judgements (W^F) – exhibit inherent tension. As for group fairness, *PFR* clearly outperforms all other representation learning methods, with group-fairness metrics as good as those of *Hardt* whose sole optimization goal is to equalize the error rates. This strong behavior of PFR on group fairness measures is remarkable as PFR is not explicitly designed for this goal. It underlines, however, the point that pairwise judgements is highly beneficial side-information, especially when comparing individuals from a-priori incomparable groups via quantiles from within-group rankings. The flexibility to incorporate such information is a salient advantage of PFR, missing in prior works for fair representation learning.

5. CONCLUSIONS

This paper proposes a new departure for the hot topic of how to incorporate fairness in algorithmic decision making. Building on the paradigm of individual fairness, we devised a new method, called *PFR*, for operationalizing this line of models, by eliciting and leveraging side-information on pairs of individuals who are equally deserving and, thus, should be treated similarly for a given task. We developed an optimization model to learn Pairwise Fair Representations (*PFR*), using the fairness graphs of pairwise judgements as inputs. We carried out comprehensive experiments with the Compas recidivism data and decile scores derived from questionnaires, and with the Crime & Communities data on socio-economic properties and ratings of neighborhoods by former and current residents. In both cases, the side-information on fairness turned out to be beneficial for giving members of the protected group their deserved share, resulting in high individual fairness and high group fairness (near-equal error rates across groups), with reasonably low loss in utility.

6. REFERENCES

[1] E. Amid and A. Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *ICML*, 2015.

[2] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammarling, J. Demmel, C. Bischof, and D. Sorensen. Lapack: A portable linear algebra library for high-performance computers. In *ICS*, 1990.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *In ProPublica 2016*.

[4] A. Asudeh, H. V. Jagadish, J. Stoyanovich, and G. Das. Designing fair ranking schemes. In *SIGMOD*, 2019.

[5] J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*, 2018.

[6] T. Brennan, W. Dieterich, and B. Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *CJB*, 2009.

[7] R. L. Brooks. *Rethinking the American race problem*. Univ of California Press, 1992.

[8] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *ICDM*, 2009.

[9] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *ICALP*, 2018.

[10] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *NIPS*, 2017.

[11] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[12] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.

[13] K. Crawford. Artificial intelligenceâĂŹs white guy problem. *The New York Times 2016*, 2016.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, 2012.

[15] S. Elbassuoni, S. Amer-Yahia, C. E. Atie, A. Ghizzawi, and B. Oualha. Exploring fairness of ranking in online job marketplaces. In *EDBT*, 2019.

[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.

[17] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[18] S. Gillen, C. Jung, M. Kearns, and A. Roth. Online learning with an unknown fairness metric. In *NeurIPS*, 2018.

[19] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 2017.

[20] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *In NIPS 2016*.

[21] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

[22] M. J. K., J. R. L., C. R., and R. S. Counterfactual fairness. In *NIPS*, 2017.

[23] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, 2010.

[24] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Considerations on fairness-aware data mining. In *ICDM*, 2012.

[25] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, 2011.

[26] M. Kearns, A. Roth, and Z. S. Wu. Meritocratic fairness for cross-population selection. In *ICML*, 2017.

[27] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *JACM*, 2002.

[28] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

[29] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, 2019.

[30] Y. Lin, T. Liu, and H. Chen. Semantic manifold learning for image retrieval. In *ACM Multimedia*, 2005.

[31] R. M. Communities and crime dataset, uci machine learning repository, 2009.

[32] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, 2008.

[33] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, 2019.

[34] T. Speicher, H. Heidari, H. Grgic-Hlaca, K. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *KDD*, 2018.

[35] J. Stoyanovich, K. Yang, and H. V. Jagadish. Online

set selection with fairness and diversity constraints. In *EDBT*, 2018.

[36] M. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *NIPS*, 2017.

[37] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017.

[38] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[39] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.