

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 human 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.

PVLDB Reference Format:

Preethi Lahoti, Krishna P. Gummadi, Gerhard Weikum. Operationalizing Individual Fairness with Pairwise Fair Representations. *PVLDB*, 13(4): 506 - 518, 2019.
DOI: <https://doi.org/10.14778/3372716.3372723>

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 has raised concerns about their

fairness [3, 12], and recent research has started to investigate how to incorporate formalized notions of fairness into machine prediction models (e.g., [13, 19, 22, 34]).

Individual vs Group Fairness: The fairness notions explored by the bulk of the works can be broadly categorized as targeting either *group fairness* [29, 15] or *individual fairness* [13]. 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 [15], equality of opportunity [19], equalized odds [19], and envy-free group fairness [33]. 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 [13], 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 [13], the authors envisioned that a task-specific similarity metric would be provided by human experts which 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 [37] and [26], wherein the authors operationalize a simplified notion of similarity

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 13, No. 4

ISSN 2150-8097.

DOI: <https://doi.org/10.14778/3372716.3372723>

metric. Concretely, they assume a distance metric (similarity metric) such as a *weighted* Euclidean distance over a feature space of data attributes, 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 [13]: “. . . 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” as a quantitative 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 as 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 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.

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. We view this as valuable side information, but we consider it to be subjective and noisy. Aggregation over many users can mitigate this, but we cannot expect G to be perfectly fair. Further, for generality, we do not assume that these are always complete. In many applications, only partial and sometimes sparse fairness judgments would be available. In our experiments, we study the sensitivity to the amount of data in G in Subsection 4.5. 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 the algorithm needs to satisfy. In Section 3, we propose a model called *PFR* (for Pairwise Fair Representations), which 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 fairness constraints captured in the fairness graph.

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 subsample of individuals in the 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: Two popular notions of group fairness are demographic parity, which requires equality of beneficial outcome prediction rates between different socially salient groups, [8, 21, 29], and equalized odds that aims to achieve equality of prediction error rates between groups [19]. Approaches to achieve group fairness include de-biasing the input data via data perturbation, re-sampling, modifying the value of protected attribute/class labels [30, 21, 29, 15] as well as incorporating group fairness as an additional constraint in the objective function of machine learning models [23, 8, 35]. Similar approaches to achieve group fairness have been proposed for fair ranking [4, 14, 36], fair set selection and clustering [9, 32] 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 [25, 10, 16, 11].

Bridging Individual and Group Fairness: Approaches to enforcing group fairness have mostly ignored individual fairness and vice versa. In [37] and [26], authors operationalize individual fairness by learning a restricted form of distance metric from the data. Some recent works use the objective of the learning algorithm itself to implicitly define the similarity metric [31, 5, 24]. 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 address 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 the 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. Other works like [17] [20] make similar arguments. Further, we show that by appropriately constraining outcomes for pairs of individuals belonging to different groups, we are able to achieve both individual and group fairness to a large degree.

Learning Pairwise Fair Representations: In terms of our technical machinery, the closest prior work is [37, 26] 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 mapping equally deserving individuals as closely as possible. Like [37, 26] our method can be used to find representations for new individuals not seen in the training data.

Finally, the core optimization problem we formulate relates to graph embedding and representation learning [18]. The aim of graph embedding approaches is to learn a representation for the nodes in the graph encoding the edges between nodes as well as the attributes of the nodes [27, 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 :

$$X = [x_1, x_2, x_3, \dots, x_N] \in R^{M \times N}$$

- Z is a low-rank representation of X in a D -dimensional space where $D \ll M$.

$$Z = [z_1, z_2, z_3, \dots, z_N] \in R^{D \times N}$$

- 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 which 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:

$$W_{ij}^X = \begin{cases} \exp\left(\frac{-\|x_i - x_j\|^2}{t}\right), & \text{if } x_i \in N_p(x_j) \text{ or } x_j \in N_p(x_i) \\ 0, & \text{otherwise} \end{cases}$$

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 irrespective of their group membership.

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 [24] 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 programming language researcher and machine learning 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 .*

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

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:*

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

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 judgments 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: 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 $x_1 \dots x_N$ and M numeric or categorical attributes, *PFR* computes a lower-dimensional latent matrix Z of N records each with $D < M$ values.

We model utility into the notion of preserving local neighborhoods of user records in the attribute space X in 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_{ij}^X . 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* , denoted by $Loss_X$.

$$Loss_X = \sum_{i,j=1}^N \|z_i - z_j\|^2 W_{ij}^X \quad (3)$$

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* , denoted by $Loss_F$.

$$Loss_F = \sum_{i,j=1}^N \|z_i - z_j\|^2 W_{ij}^F \quad (4)$$

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 orthonormality constraint on Z is imposed to avoid trivial results. The trivial result being that all the datapoints are mapped to same point.

$$\begin{aligned} &\text{Minimize } (1 - \gamma) \sum_{i,j=1}^N \|z_i - z_j\|^2 W_{ij}^X + \gamma \sum_{i,j=1}^N \|z_i - z_j\|^2 W_{ij}^F \\ &\text{subject to } Z^T Z = I \end{aligned} \quad (5)$$

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:

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

where $Tr(\cdot)$ 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:

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

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

$$X((1 - \gamma)L^X + \gamma L^F)X^T \mathbf{v}_i = \lambda \mathbf{v}_i \quad (7)$$

It follows that the columns of optimal V are the eigenvectors corresponding to D smallest eigenvalues denoted by $V = [\mathbf{v}_1 \mathbf{v}_2 \mathbf{v}_3 \dots \mathbf{v}_D]$, 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 in `scipy.linalg.lapack` python library [2].

3.3.4 Inference

Given an input vector x_i for a previously unseen individual, the *PFR* method computes its fair representation as $z_i = V^T x_i$ where z_i is the projection of the datapoint x_i on the learned basis V . It is important to note that the fairness graph W^F is only required during the training phase to learn the basis V . Once the $M \times D$ matrix V is learned, we do not need any fairness labels for newly seen data.

3.3.5 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). If we assume that

$$Z = V^T \Phi(X) \quad \text{where } V = \sum_{i=1}^n \alpha_i \Phi(x_i)$$

with a Mercer kernel matrix K where $K_{i,j} = k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. 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.

$$K((1 - \gamma)L^X + \gamma L^F)K \alpha_i = \lambda \alpha_i$$

Analogously to the solution of Equation 7, the solution to the *kernel PFR* is given by $A = [\alpha_1 \alpha_2 \dots \alpha_D]$ where $\alpha_1 \dots \alpha_D$ 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 methods on a binary classification task as a downstream application. We address the following key questions in our main results in Subsection 4.2, 4.3.2 and 4.3.3:

- [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?

In addition, to understand the robustness of our model to the main hyper-parameter γ , as well as the sensitivity of the model to the number of labels in the fairness graph, we report additional results in Subsection 4.4, and 4.5.

4.1 Experimental Setup

Baselines: We compare the performance of the following methods

- *Original representation*: a naive representation of the input dataset wherein the protected attributes are masked.
- *iFair [26]*: a representation learning method, which optimizes for two objectives: (i) individual fairness in W^X , and (ii) obfuscating protected attributes.
- *LFR [37]*: a representation learning method, which optimizes for three objectives: (i) accuracy (ii) individual fairness in W^X and (iii) demographic parity.
- *Hardt [19]*: 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: For 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 fairness labels that are used to construct W^F as additional numerical features in the training data. Since we only have judgments for a sample of training data, we treat the rest as missing values and set them to -1. Note that this enhancement is only for training data as fairness labels are not available for unseen 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 X are available. Concrete details for each of the datasets follow in their respective subsections.

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 a

independent test set to measure performance. All reported results are averages over 10 runs on independent test sets.

Datasets and Task: We compare all methods on downstream classification using three datasets: (i) a synthetic dataset for US university admission with 203 numerical features, and two real-world datasets: (ii) crime and communities dataset for violent neighbourhood predictions with 96 numerical features and 46 one-hot encoded features (for categorical attributes), and (iii) compas dataset for recidivism prediction with 9 numerical and 420 one-hot encoded features. In order to check the “true” dimensionality of the datasets we computed the smallest rank k for SVD that achieves a relative error of at most 0.01 for the Frobenius norm difference between the SVD reconstruction and the original data. For the three datasets, these dimensionalities are 156, 69, and 117 respectively. Table 1 summarizes the statistics for each dataset, including base-rate (fraction of samples belonging to the positive class, for both the protected group and its complement). In all experiments, the representation learning methods are followed by an out-of-the-box logistic regression classifier trained on the corresponding representations.

Table 1: Experimental settings and dataset statistics

Dataset	No of records	No. of features	True Rank	Base-rate (s = 0)	Base-rate (s = 1)	Protected attribute
Synthetic	1000	203	156	0.51	0.48	Race
Crime	1993	142	69	0.35	0.86	Race
Compas	8803	429	117	0.41	0.55	Race

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. Equal 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 $s \in S$:

$$P(\hat{Y} = 1 | s = 0) = P(\hat{Y} = 1 | s = 1) \quad (8)$$

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

4.2 Experiments on Synthetic Data

We simulate the US graduate admissions scenario of Section 1.1 where our task is to predict the ability of a candidate to complete graduate school (binary classification). To this end, we imagine that the features in a college admission task can be grouped into two categories. First set

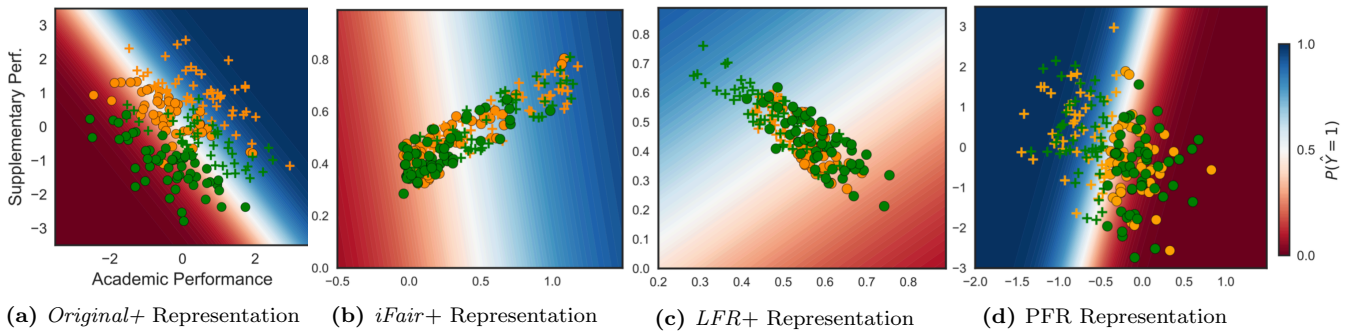


Figure 1: Comparison of (a) *Original+* (b) *iFair+* (c) *LFR+* and (d) *PFR* representations on a synthetic dataset. Colors depict membership to protected group (S): orange (non-protected) and green (protected). Markers denote *true* class labels: $Y = 1$ (marker +) and $Y = 0$ (marker o). Contour plots visualize decision boundary of a classifier trained on the representations. Blue color corresponds to predicted positive classification, red to predicted negative class. The more intensive the color, the higher or lower the score of the classifier.

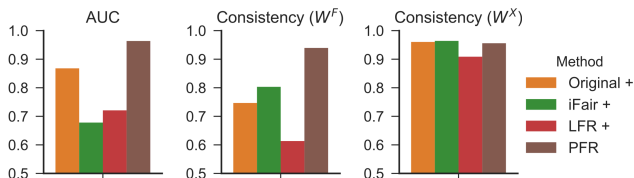


Figure 2: Results for Synthetic low dimension dataset: Comparison of utility vs individual fairness trade-off across methods. Higher values are better.

of features which are related to their *academic performance* such as overall GPA, grades in each of the high schools subjects like Mathematics, Science, Languages, etc. Second set of features are related to their *supplementary performance* which constitute their overall application package such as SAT scores, admission essay, extracurricular activities, etc.

We assume that the scores for the second set of features can be inflated for individuals who have higher privilege in the society, for instance by re-taking SAT exam, and receiving professional coaching. Suppose we live in a society where our population consists of two groups $s = 0$ or 1 , and the group membership has a high correlation with individual’s privilege. This would result in a scenario where the two groups have different feature distributions. Further, if we assume that the inflation in the scores does not increase the ability of the candidate to complete college, the relevance functions for the two groups would also be different.

Creating Synthetic Datasets: We simulate this scenario by generating synthetic data for two population groups X_0 and X_1 . Our dataset consists of three main features: group, *academic performance*, and *supplementary performance*. The correlation between *academic performance* and *supplementary performance* is set to 0.3. We have additional 100 numerical features with high correlation to *academic performance*, and 100 numerical features with high correlation to *supplementary performance*. We set the value of correlation between related features by drawing uniformly from $[0.75, 1.0]$. We use the correlation between features to construct the covariance matrix for a multivariate Gaussian distribution of dimensionality 203. To reflect the point that one groups has inflated scores for the features related to *supplementary performance*, we set the mean for these features for the non-protected group one standard deviation

higher than the mean for the protected group.

In total we generate 600 samples for training, and 400 samples as a withheld test set. We run our experiments on two versions of the synthetic dataset: (i) a *low-dimensional* dataset, which is a subset of the high-dimensional data consisting of only three features: Group, Academic Performance and Supplementary performance, and (ii) a *high-dimensional* dataset with all 203 features. Experiments on the low-dimensional dataset are performed in order to be able to visually compare the original and learned representations. Dataset statistics are shown in Table 1.

Ground Truth Labels: Despite average score on *supplementary performance* features 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 *supplementary performance* scores. To implement this scenario, we set the *true* class label for group $X_{s=1}$ to positive (1) if *academic + supplementary* score ≥ 0 and for group $X_{s=0}$ as positive (1) if *academic + supplementary* score ≥ 1 . 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 : In this experiment we simulate the scenario for eliciting human input on fairness, wherein we have access to a fairness oracle who can make the judgments of the form “Is A similar to B?” described in Subsection 3.2.1. To this end, we randomly sample $N \log_2 N := 5538$ pairs (out of the possible $N^2 := 600 \times 600$). We then constructed ours fairness graphs W^F by querying a fairness oracle for Yes/No answers to similarity pairs. If the two points are similar we add an edge between the two nodes.

Fairness oracle for this task is a machine learning model consisting of two separate logistic regression models, one for each group, $X_{S=0}$ and $X_{S=1}$ respectively. Given a pair of points, if their prediction probabilities fall in the same quantile, they are deemed similar by the fairness oracle.

Augmenting Baselines: We cast each row of the matrix W^F (of the fairness graph) into n additional binary features for the respective individual. That is, for every user record, n additional 0/1 features indicate pairwise equivalence. All baselines have access to this information via the augmented input matrix X .

4.2.1 Results on Synthetic Low Dimension 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+* [26], *LFR+* [37], 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. We make the following observations:

- Individual fairness regarding W^F : *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.
- Individual fairness regarding W^X : *PFR*’s performance for *consistency* (W^X) is as good as other approaches, however *PFR* manages to achieve high performance for significantly better performance on *AUC* and *consistency* (W^F).

[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* achieves by far the best *AUC*, even outperforming the original representation. While this may surprise on first glance, it is indeed an expected outcome. The fairness edges in W_F help *PFR* overcome the challenge of different groups having different feature distributions (observe Figure 1a). In contrast, *PFR* is able to learn a unified representation that maps deserving candidates of one group close to deserving candidates of the other group (observe Figure 1d), which helps in improving *AUC*.

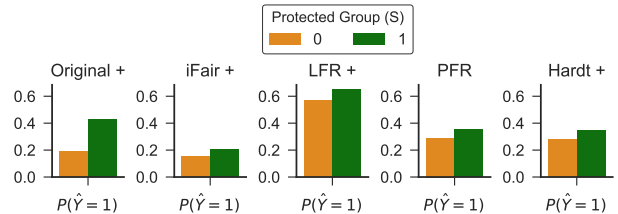
[Q4] Influence on Group Fairness: In addition to *Original+*, *iFair+*, *LFR+* and *PFR*, we include the *Hardt* model in the comparison here, as it is widely viewed as the state-of-the-art method for group fairness.

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

- Disparate Mistreatment (Figure 3a): We observe that *Original+* model has high difference in error rates (aka. Equality of Odds). *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* succeeding in mapping equally deserving individuals from both groups to close proximity in its latent space.
- Disparate Impact (Figure 3b): The *Original+* approach exhibits a substantial difference in the per-group positive prediction rates of the two groups. In contrast, *iFair+*, *LFR+*, and *PFR* representation have the *orange* and *green* data points well-mixed, and this way achieve nearly equal rates of positive predictions for both groups. Likewise *Hardt+* has the same desired effect.



(a) Per-group error rates (FPR and FNR)



(b) Per-group in rates of positive prediction

Figure 3: Results for Synthetic low dimension dataset: Difference in (a) error rates and (b) rate of positive predictions between protected and non-protected groups. Lower values are better.

4.2.2 Results on Synthetic High Dimension Dataset

The results for the high-dimensional synthetic data are largely consistent with the results for the low-dimensional case of Subsection 4.2.1. Therefore, we discuss them only

briefly. Figure 4 shows results for AUC , $consistency(W^F)$, and $consistency(W^X)$. Figure 5 shows results on group fairness measures.

Utility vs. Individual fairness regarding W^F : On first glance, $LFR+$ seems to perform best on consistency with regard to W^F . However, this is trivially achieved by giving the same prediction to almost all datapoints: the classifier using the learned $LFR+$ representation accepts virtually all individuals, hence its very poor AUC of around 0.55. In essence, $LFR+$ fails to learn how to cope with the utility-fairness trade-off. Therefore, we consider this method as degenerated (for this dataset) and dismiss it as a real baseline.

Among the other methods, PFR significantly outperforms all competitors by achieving the best performance on $consistency(W^F)$, similar performance as other approaches on $consistency(W^X)$, but for a significantly better performance on AUC , as shown in Figure 4.

Group Fairness: Once again, PFR clearly outperforms all other methods on group fairness. It achieves near-equal error rates across groups, and near-equal rates of positive predictions as shown in Figures 5a and 5b. 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. $LFR+$ seems to achieve good results as well, but this is again due to accepting virtually all individuals (see above).

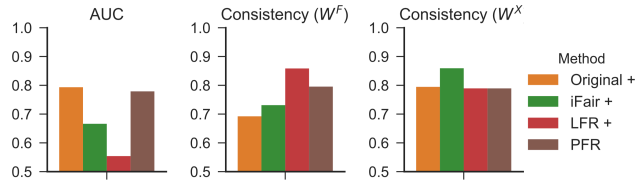


Figure 4: Results for Synthetic high dimension dataset: Comparison of utility vs individual fairness trade-off across methods. Higher values are better.

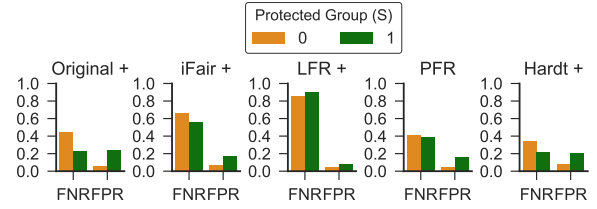
4.3 Experiments on Real-World Datasets

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

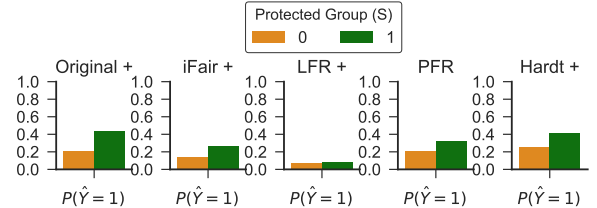
- *Crime & Communities* [28] is a dataset consisting of socio-economic (e.g., income), demographic (e.g., race), and law/policing data (e.g., patrolling) records for neighborhoods in the US. We set *isViolent* as target variable for a binary classification task. We consider the communities with majority 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



(a) Per-group error rates (FPR and FNR)



(b) Per-group rates of positive prediction

Figure 5: Results for Synthetic high dimension dataset: Difference in (a) error rates between protected and non-protected groups and (b) rate of positive predictions.

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 neighborhoods, 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. 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 sort these decile scores for each group separately to

simulate per-group ranking fairness judgments. We then use these per-group rankings as the fairness judgment to construct the fairness graphs for incomparable individuals as discussed in Subsection 3.2.2. Specifically, we compute k quantiles over the ranking as per Definition 2 and then, construct W^F as described in Definition 3. 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: We give our baselines access to the elicited fairness labels by adding them as numerical features to the rows of the input matrix X . For the Crime and Communities data, we added the elicited ratings (1 to 5 stars) as numerical features, with missing values set to -1. For the Compas data, where the fairness labels are per-group rankings, we added the ranking position of each individual within its respective group as a numerical feature.

4.3.2 Results on Crime & Communities Dataset

[Q2] Effect on Individual Fairness: Results on individual fairness and utility are given in Figure 6. We observe that even though all the methods have access to the same fairness information, only *PFR* shows an improvement in consistency W^F over the baseline. *PFR* outperforms all other methods on individual fairness (consistency W^F). However, this gain for W^F comes at the cost of losing in consistency as per 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 higher performance of *PFR* on individual fairness regarding W^F comes with a drop in utility as shown by the AUC bars in Figure 6. This is because, unlike the case of the synthetic data in Subsection 4.2, the side-information for the fairness graph W^F is not strongly aligned with the ground-truth for the classifier. In terms of relative comparison, we observe that only *PFR* shows an improvement in consistency W^F over the baseline, the other approaches show no improvement. The performance of *iFair+* and *LFR+* on consistency on W^F and consistency on W^X is same as that of *Original+*, however for a lower AUC. *Hardt+* loses on all the three measures.

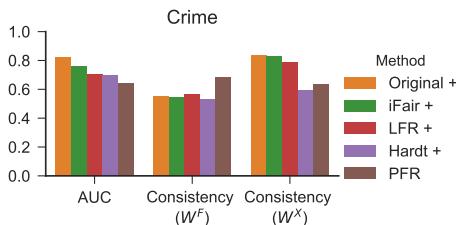
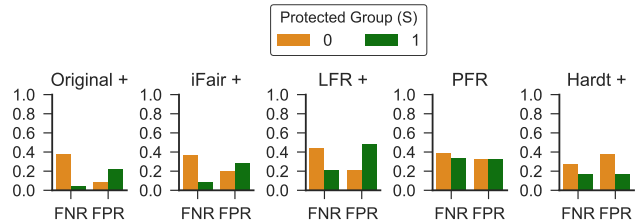


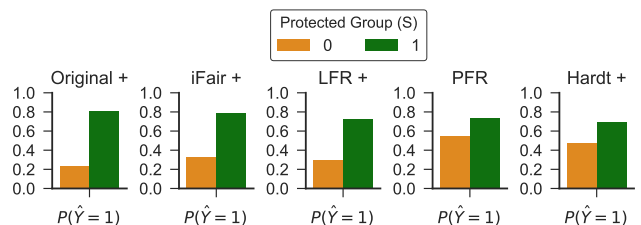
Figure 6: Crime & Communities data: utility vs. individual fairness.

[Q4] Influence on Group Fairness Figure 7a shows the per-group error rates, and 7b shows the per-group positive prediction rates. Smaller differences in the values between the two groups are preferable. The following observations are notable:

- 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.
- Disparate Impact (aka. Demographic parity): *PFR* outperforms all the methods by achieving near perfect balance (i.e., near-equal rates of positive predictions).



(a) Per-group error rates (FPR and FNR)



(b) Per-group rates of positive prediction

Figure 7: Crime & Communities data: (a) error rates (lower values are better) and (b) positive prediction rates.

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 on W^F , as shown in Figure 8.

Group Fairness: However, *PFR* clearly outperforms all other methods on group fairness. It achieves near-equal rates of positive predictions as shown in Figure 9b, and near-equal error rates across groups as shown in Figure 9a. 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.

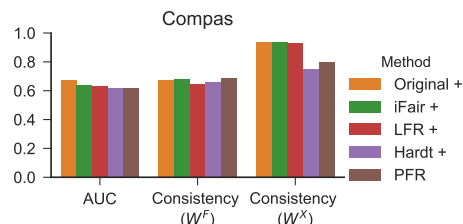


Figure 8: Compas data: utility vs. individual fairness.

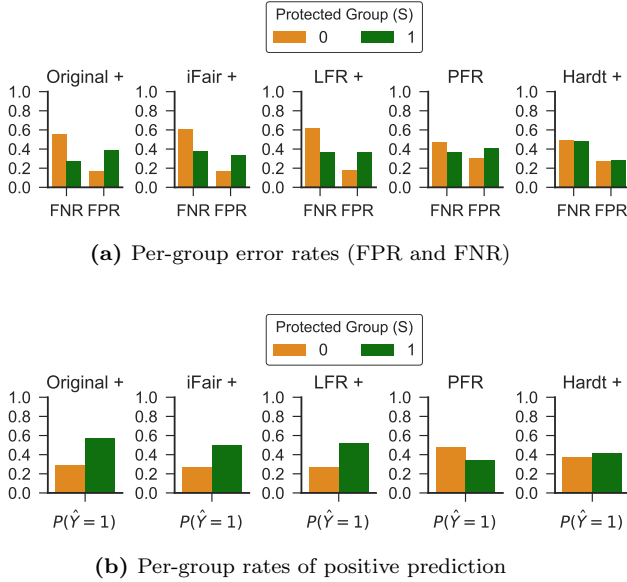


Figure 9: Compas data: (a) error rates (lower values are better) and (b) positive predictions rates.

4.4 Influence of PFR Hyper-Parameter γ

In this subsection we analyze the influence of γ on the trade-off between individual fairness (consistency W^F) and utility (AUC) of the downstream classifiers. To this end, we keep all other hyper-parameters set to their values for the best result in the main experiments, and systematically vary the value of hyper-parameter γ in $[0,1]$.

Recall that *PFR* aims to preserve local neighborhoods in the input space X (given by W^X), as well as the similarity given by the fairness graph W^F , where the hyper-parameter γ controls the relative influence of W^X and W^F . Figure 10 shows the influence of γ on individual fairness and utility for (a) low-dimensional synthetic, (b) Crime and (c) Compas data, respectively. We make the following key observations.

Individual Fairness: We observe that with increasing γ the consistency with regard to W^F increases. This is in line with our expectation: as γ increases the influence of W^F on the objective function, the performance of the model on individual fairness (consistency W^F) improves. This trend holds for all the datasets. It is worth highlighting that the improvement in individual fairness is for newly seen test samples that were unknown at the time when the fairness graph W^F was constructed and the *PFR* model was learned. This demonstrates the ability of *PFR* to generalize individual fairness to unseen data.

Utility: The influence of γ on the utility is more nuanced. We observe that the extent of the trade-off between individual fairness in W^F and utility depends on the degree of conflict between the pairwise W^F , and the classifier’s ground-truth labels.

- If W^F indicates equal deservingness for data points that have different ground-truth labels, there is a natural conflict between individual fairness and utility. We observe this case for the real-world datasets Crime and Compas where W^F is in tension with ground-truth labels – presumably due to implicit anti-subordination embedded in graph or equivalently, due to historic discrimination in

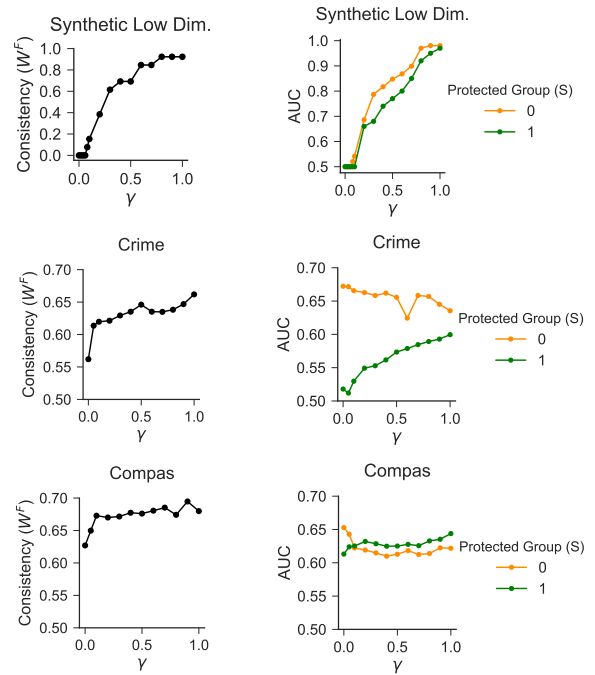


Figure 10: Influence of γ on individual fairness and utility.

the classification ground-truth. With increasing γ , there is a slight drop in the utility *AUC* for the non-protected group. However, there is an improvement in *AUC* for the protected group. The overall *AUC* drops by a few percentage points, but stays at a high level even for very high γ . So we trade off a substantial gain in individual fairness for a small loss in utility. This is a clear case of how incorporating side-information on pairwise judgments can help to improve algorithmic decision making for historically disadvantaged groups.

- In contrast, if W^F pairs of equal deservingness are compatible with the classifier’s ground-truth labels, there is no trade-off between utility and individual fairness. In such cases, W^F may even help to improve the utility by better learning a similarity manifold in the input space. We observe this case for the synthetic data where W^F is consistent with the ground-truth labels. As γ increases, the *AUC* of a classifier trained on *PFR* is enhanced. The improvement in *AUC* holds for both protected and non-protected groups.

4.5 Sensitivity to Sparseness of Fairness Graph

In this section we study the sensitivity of *PFR* to the sparseness of the labeled pairs in the fairness graph W^F . We fix all hyper-parameters to their best values in the main experiments, and systematically vary the fraction of data-points for which we use pairwise fairness labels. The results are shown in Figure 11. All results reported are on out-of-sample withheld test set of fairness graph W^F . Recall that *PFR* accesses fairness labels only for training data. For test data, it solely has the data attributes available in X .

Setup: For the synthetic data, we uniformly at random sampled fractions of $[\log_2 N, \frac{N}{5}, \dots, N, N \log_2 N, N^2]$ pairs from the training data, which for this data translates into $[9, 120, \dots, 600, 5537, 360000]$ pairs. For the Crime data, we varied the percentage of training samples for which use

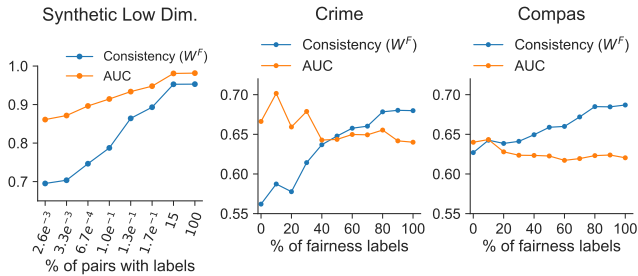


Figure 11: Influence of fairness-graph sparseness.

equivalence labels, in steps of 10% from 10% to 100%. For the Compas data, we varied the percentage of training data points for which we elicit per-group rankings, in steps of 10% from 10% to 100%.

Results: We observe the following trends.

- Increasing the fraction of fairness labels improve the results on individual fairness (consistency for W^F), while hurting utility (AUC) only mildly (or even improving it in certain cases).
- For the synthetic data, even with as little as 0.17% of the fairness labels, the results are already fairly close to the best possible: consistency for W^F is already 90%, and AUC reaches 95%.
- For the Crime data, we need about 30 to 40% to get close to the best results for the full fairness graph. However, even with sparseness as low as 10%, PFR degrades smoothly: consistency W^F is 59% compared to 68% for the full graph, and AUC is affected only mildly by the sparseness.
- For the Compas data, we observe similar trends: even with very sparse W^F we stay within a few percent of the best possible consistency, and AUC varies only mildly with changing sparseness of the fairness graph.

These observations indicate that the PFR model yields benefits already with a small amount of human judgements of equally deserving individuals.

4.6 Discussion and Lessons

The experimental results suggest several key findings.

- *Individual Fairness - Utility Trade-off:* The extent of this trade-off depends on the degree of conflict between the fairness graph and the classifier’s ground-truth labels. When edges in the fairness graph connect data points (for equally deserving individuals) that have different ground-truth labels, there is an inherent tension between individual fairness and utility.

For datasets where some compromise is unavoidable, PFR turns out to perform best in balancing the different goals. It is consistently best with regard to individual fairness, by a substantial margin over the other methods. On utility, its AUC is competitive and always close to the best performing method on this metric, typically within 2 percentage points of the best AUC result.

- *Balancing Individual Fairness and Group Fairness:* The human judgements cast into the fairness graph help PFR to perform well also on group fairness criteria. On these measures, PFR is almost as good as the method by Hardt et al., which is specifically geared for group fairness (but disregards individual fairness). To a large extent, this

is because the pairwise fairness judgments address historical subordination of groups. Eliciting human judgements is a crucial asset for fair machine learning in a wider sense.

- *Data Representation:* The graph-embedding approach of PFR appears to be the best way of incorporating the pairwise human judgements. Alternative representations of the same raw information such as additional features in the input dataset, as leveraged by the augmented baselines ($LFR+$, $iFair+$), perform considerably worse than PFR on consistency (W^F).

The W^F input is needed solely for the training data; previously unseen test data (at deployment of the learned representation and downstream classifier) does not have any pairwise judgments at all. This underlines the practical viability of PFR .

- *Graph Sparseness:* Even a small amount of pairwise fairness judgments helps PFR in improving fairness. At some point of extreme sparseness, PFR loses this advantage, but its performance degrades quite gracefully.
- *Robustness:* PFR is fairly robust to the dimensionality of the dataset. As the dimensionality of the input data increases, the performance of PFR drops a bit, but still outperforms other approaches in terms of balancing fairness and utility. Furthermore, PFR is quite insensitive to the choice of hyper-parameters. Its performance remains stable across a wide range of values.
- *Limitations:* When the data exhibits a strong conflict between fairness and utility goals, even PFR will fail to counter such tension and will have to prioritize either one of the two criteria while degrading on the other. The human judgements serve to mitigate exactly such cases of historical subordination and discrimination, but if they are too sparse or too noisy, their influence will be marginal. For the datasets in our experiments, we assumed that the information on equally deserving individuals would reflect high consensus among human judges. When this assumption is invalid for certain datasets, PFR will lose its advantages and perform as poorly as (but no worse than) other methods.

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 a representation learning model to learn Pairwise Fair Representations (PFR), as a fairness-enhanced input to downstream machine-learning tasks. Comprehensive experiments, with synthetic and real-life datasets, indicate that the pairwise judgements are beneficial for members of the protected group, resulting in high individual fairness and high group fairness (near-equal error rates across groups) with reasonably low loss in utility.

6. ACKNOWLEDGMENT

This research was supported by the ERC Synergy Grant “imPACT” (No. 610150) and ERC Advanced Grant “Foundations for Fair Social Computing” (No. 789373).

7. 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] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *NIPS*, 2017.
- [10] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.
- [11] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.
- [12] K. Crawford. Artificial intelligence’s white guy problem. *The New York Times 2016*, 2016.
- [13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, 2012.
- [14] S. Elbassuoni, S. Amer-Yahia, C. E. Atie, A. Ghizzawi, and B. Oualha. Exploring fairness of ranking in online job marketplaces. In *EDBT*, 2019.
- [15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.
- [16] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.
- [17] S. Gillen, C. Jung, M. Kearns, and A. Roth. Online learning with an unknown fairness metric. In *NeurIPS*, 2018.
- [18] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 2017.
- [19] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *In NIPS 2016*.
- [20] 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.
- [21] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, 2010.
- [22] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Considerations on fairness-aware data mining. In *ICDM*, 2012.
- [23] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, 2011.
- [24] M. Kearns, A. Roth, and Z. S. Wu. Meritocratic fairness for cross-population selection. In *ICML*, 2017.
- [25] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.
- [26] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, 2019.
- [27] Y. Lin, T. Liu, and H. Chen. Semantic manifold learning for image retrieval. In *ACM Multimedia*, 2005.
- [28] R. M. Communities and crime dataset, uci machine learning repository, 2009.
- [29] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, 2008.
- [30] B. Salimi, L. Rodriguez, B. Howe, and D. Suci. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, 2019.
- [31] 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.
- [32] J. Stoyanovich, K. Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *EDBT*, 2018.
- [33] M. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *NIPS*, 2017.
- [34] 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.
- [35] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.
- [36] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. A. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *CIKM*, 2017.
- [37] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.