Fairness in Relational Domains

Golnoosh Farnadi^{1**}**, Behrouz Babaki**^{2**}**, Lise Getoor**^{1}
^{1}University of California, Santa Cruz, USA,
^{2}KU Leuven, Belgium

Abstract

AI and machine learning tools are being used with increasing frequency for decision making in domains that affect peoples' lives such as employment, education, policing and loan approval. These uses raise concerns about biases of algorithmic discrimination and have motivated the development of fairness-aware machine learning. However, existing fairness approaches are based solely on attributes of individuals. In many cases, discrimination is much more complex, and taking into account the social, organizational, and other connections between individuals is important. We introduce new notions of fairness that are able to capture the relational structure in a domain. We use first-order logic to provide a flexible and expressive language for specifying complex relational patterns of discrimination. Furthermore, we extend an existing statistical relational learning framework, probabilistic soft logic (PSL), to incorporate our definition of relational fairness. We refer to this fairness-aware framework FairPSL. FairPSL makes use of the logical definitions of fairnesss but also supports a probabilistic interpretation. In particular, we show how to perform maximum a posteriori(MAP) inference by exploiting probabilistic dependencies within the domain while avoiding violation of fairness guarantees. Preliminary empirical evaluation shows that we are able to make both accurate and fair decisions.

1 Introduction

Over the past few years, AI and machine learning have become essential components in operations that drive the modern society, e.g., in financial, administrative, and educational spheres. *Discrimination* happens when qualities of individuals which are not relevant to the decision making process influence the decision. Delegating decision making to an automated process raises questions about discriminating against individuals with certain traits based on biases in the data. This is especially important when the decisions have a potential to impact the lives of individuals, for example, the decisions on granting loans, assigning credit, and employment.

Fairness is defined as the absence of discrimination in a decision making process. The goal of *fairness-aware* machine learning is to ensure that the decisions made by an algorithm do not discriminate against a population of individuals (Feldman et al. 2015; Boyd, Levy, and Marwick 2014; Hardt et al. 2016). Fairness has been well studied in the social sciences and legal scholarship (for an in-depth review see (Barocas and Selbst 2016)), and there is emerging work on fairness-aware ML within the AI and computer science communities. For example, fairness through awareness/Lipschitz property (Dwork et al. 2012), individual fairness (Zemel et al. 2013), statistical parity/group fairness (Kamishima, Akaho, and Sakuma 2011), counterfactual fairness (Kusner et al. 2017), demographic parity/disparate impact (Feldman et al. 2015; Chouldechova 2017), preference-based fairness (Zafar et al. 2017), and equality of opportunity (Hardt et al. 2016).

The existing work in fairness-aware machine learning is based on a definition of discrimination where a decision is influenced by an *attribute* of an individual. An attribute value upon which discrimination is based (such as gender, race, or religion) is called a *sensitive attribute*. The sensitve attribute defines a population of vulnerable individuals known as the *protected group*. A fair decision-making process treats the protected group the same as the *unprotected group*.

However, in many social contexts, discrimination is the result of complex interactions and can not be described solely in terms of attributes of an individual. For example, consider an imaginary scenario in an organization in which younger female workers who have older male supervisors have lower chances of promotion than their male counterparts.[*] This discrimination pattern involves two attributes of the individual (gender and age), a relationship with another individual (supervisor), and two attributes of the second individual. Addressing such complex cases poses two challenges. First, the concepts of discrimination and fairness need to be extended to capture not only attributes of individuals but also the relationships between them. Second, a process is required that ensures that fair decisions are made about individuals who are affected by such patterns. In this paper we address both of these two challenges. We use first-order logic (FOL) to extend the notion of fairness to the relational setting. FOL is an expressive representation for relational problems which also widely used for learning in relational domains. Moreover, we extend an existing framework for statistical relational learning (Getoor and Taskar 2007) called probabilistic soft logic (PSL)[†] (Bach et al. 2017). PSL

[**] Equal contributors.

[*] Of course, many other patterns may be possible: female bosses may promote female subordinates and discriminate against male workers, or male bosses may promote female employees. Our goal is to provide a general framework which is able to describe arbitrarily complex discrimination patterns.

[†] http://psl.linqs.org/

combines logic and probability for learning and reasoning over uncertain relational domains. One of the most common reasoning task in PSL is called maximum a posteriori (MAP) inference, which is performed by finding the most probable truth values for unknowns over a set of given evidence. We develop a new MAP inference algorithm which is able to maximize the a posteriori values of unknown variables *subject to* fairness gaurantees.

Our contributions are as follows: 1) We propose fairness-aware machine learning for the relational setting; 2) We extend PSL into a fairness-aware framework called FairPSL which can represent the logical definition of fairness; 3) we develop a new MAP inference algorithm which is able to maximize the posteriori values of unknown variables *subject to* fairness guarantees; 4) We empirically evaluate our proposed framework on synthetic data.

2 Motivation

The existing work on designing fair algorithms in machine learning exclusively focus on *attribute-based fairness*, which is based on the following assumptions: First, there is an assumption that the individuals (sometimes referred to as units or entities) are independent and described by simple attribute vectors. Second, the group for which one wishes to ensure fairness known as as the protected group is defined on the basis of some attribute values. Finally, there is a decision that is associated with each individual, and the goal is to ensure that members of the protected group are subject to a fair decision. We illustrate the attribute-based fairness by the following example.

Example 1 (Loan Processing)**.** *A bank bases its decisions about granting a loan on attributes of the applicant. The goal is to decide about granting the loan for each applicant using a predictive model. The administration needs to ensure that the gender of applicants has no influence on the decision. i.e., the female applicants receive the loan at a rate close to the male applicants. In this scenario, the protected group is female applicants.*

Contrary to the assumptions of the attribute-based setting, the data can be relational, and the protected group itself might be represented by a relational definition. The current fairness-aware machine learning techniques are not capable to model relations and hence cannot be used to make the decision making model fair. We illustrate this setting by a scenario, inspired by the motivating example of (Choi, Darwiche, and den Broeck 2017).

Example 2 (Paper Reviewing)**.** *Consider the reviewing process in a conference, where each submitted paper is reviewed by two reviewers. The area chair summarizes the reviews. In general, the reviewers are likely to write positive reviews about high-quality papers. Moreover, when a reviewer writes a positive review for a paper, it is less likely for her to evaluate other papers positively.*

The program chair has only time to read the review summaries from the area chair. She is faced with the problem of estimating the true quality of the papers from these summaries to decide whether paper should be accepted for presentation at the conference or not. The organizers know that the affiliation of student authors does not correlate with the quality of their papers. However, they notice that the students from undistinguished institutes have been under-represented (perhaps the style of their papers diverts the attention of reviewers from the technical quality). The organizers wish to accept high-quality papers, while ensuring that the discrimination against this group of researchers is eliminated. In this scenario, the protected group is student authors from undistinguished institutes.

Example 2 describes a prediction problem over a database that consists of relations between papers, authors, and reviewers. Such prediction tasks are best handled by techniques from the relational learning domain. To ensure fair prediction in such settings, we have to extend the notion of *attribute-based fairness* to *relational fairness* that we will address in the next section. Throughout this paper, we use the paper reviewing problem as a running example for relational fairness.

3 Fairness Formalism

A representation that can describe different types of entities and different relationships between them is called relational. In this section, we use first-order logic to define relational fairness. We employ first-order logic as an expressive representation formalism which can represent objects and complex relationships between them. We start by defining atom:

Definition 1 (Atom)**.** *An atom is an expression of the form $P(a_1, a_2, \ldots, a_n)$ where each argument $a_1, a_2, \ldots, a_n$ is either a constant or a variable. The finite set of all possible substitutions of a variable to a constant for a particular variable a is called its domain D_a. If all variables in $P(a_1, a_2, \ldots, a_n)$ are substituted by some constant from their respective domain, then we call the resulting atom a ground atom.*

Example 3. *In our loan processing problem (Example 1), we can represent applicants' attributes by atoms. For instance, atom $Female(v)$ indicates whether or not applicant v is female. Similarly, we can represent relations with atoms. In the paper reviewing problem in Example 2 the atom $Review(r, p)$ indicates whether or not reviewer r reviews paper p.*

The relational setting provides the flexibility to express complex definitions with formulae.

Definition 2 (Formula)**.** *A formula is defined by induction: every atom is a formula. If α and β are formulae, then $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$ are formulae. If x is a variable and α is a formula, then the quantified expressions of the form $\exists x\ \alpha$ and $\forall x\ \alpha$ are formulae.*

To characterize groups of individuals based on a formula, we define the notion of *population*.

Definition 3 (Population)**.** *We denote formula F which has only one free variable v (i.e. other variables in F are quantified) by $F[v]$. The population defined by $F[v]$ is the set of substitutions of v for which $F[v]$ holds.*

Example 4. *Consider the formula $F[v] := \forall u, \neg Affiliated(v, u) \vee \neg TopRank(u)$. The population specified by this formula is the set individuals who are not affiliated with a top-rank institute.*

The truth value of a formula is derived from the truth value of atoms that it comprises, according to the rules of logic. Each possible assignment of truth values to ground atoms is called an *interpretation*.

Definition 4 (Interpretation)**.** *An interpretation I is a mapping that associates a truth value $I(P)$ to each ground atom P. For Boolean truth values, I associates true to 1 and false to 0 truth values. For soft logic (see Definition 10) I maps each ground atom P to a truth value in interval $[0, 1]$.*

In attribute-based fairness, it is assumed that there is a certain attribute of individuals, i.e. the sensitive attribute, that we do not want to affect a decision. Gender, race, religion and marital status are examples of sensitive attributes. Discrimination has been defined in social science studies as a treatment in favor or against a group of individuals given their sensitive attribute. This group of individuals is the protected group.

In a relational setting, both the sensitive attributes of an individual and their participation in various relations may have an undesired effect on the final decision. We characterize the protected group in a relational setting by means of a population. In practice, we are often interested in maintaining fairness for a specific population such as applicants, students, employees, etc. This population is then partitioned into the protected and unprotected groups. We define *discriminative pattern* which is a pair of formulae to capture these groups: 1) $F_1[v]$: to specify the difference between the protected and unprotected groups and 2) $F_2[v]$: to specify the population over which we want to maintain fairness.

Definition 5 (Discriminative pattern)**.** *A discriminative pattern is a pair* $\mathit{DP}[v] := (F_1[v], F_2[v])$ *, where* $F_1[v]$ *and* $F_2[v]$ *are formulae.*

Example 5. *The two formulae in the discrimination pattern* $\mathit{DP}[v] := \big((\forall u, \neg \mathit{Affiliated}(v,u) \vee \neg \mathit{TopRank}(u)), \mathit{Student}(v)\big)$ *specify two populations, namely the student authors and the individuals who are not affiliated with top-rank institutes.*

Given the definition of the discriminative pattern, we have a rich language to define the scope of the protected and unprotected group in a relational setting.

Definition 6 (Protected group)**.** *Given an interpretation* I*, the protected group is a population of the form:*

$$PG := \{v : F_1[v] \wedge F_2[v]\}$$

which is defined as the set of all instances hold for variable v *for which* $F_1[v] \wedge F_2[v]$ *is true under interpretation* I*, that is,* $I(F_1[v] \wedge F_2[v]) = 1$*. Similarly, the* unprotected group *is a population of the form:*

$$UG := \{v : \neg F_1[v] \wedge F_2[v]\}$$

which is defined as the set of all instances hold for variable v *for which* $I(\neg F_1[v] \wedge F_2[v]) = 1$*.*

Example 6. *The protected group of the discrimination pattern specified in Example 5 is* $PG := \big\{v : \big(\forall u, \neg \mathit{Affiliated}(v,u) \vee \neg \mathit{TopRank}(u)\big) \wedge \mathit{Student}(v)\big\}$ *and the unprotected group is* $UG := \big\{v : \big(\exists u, \mathit{Affiliated}(v,u) \wedge \mathit{TopRank}(u)\big) \wedge \mathit{Student}(v)\big\}$*. This means our protected group is students who are not affiliated with top-rank institutes and our unprotected group is students affiliated with top-rank institutes.*

Discrimination is defined in terms of a treatment or decision that distinguishes between the protected and unprotected groups. Here we define *decision* atom.

Definition 7 (Decision atom)**.** *A decision atom* $d(v)$ *is an atom containing exactly one variable* v *that specifies a decision affecting the protected group which is defined either by law or end-user.*

Example 7. *The decision atom* $\mathit{Presents}(v)$ *indicates whether or not* v *presents their paper.*

Note that the fairness formulation in this section is designed for the relational setting, however relational fairness subsumes the attribute-based fairness such that: a sensitive attribute is defined by an atom with one argument and $F_2[v]$ in discrimination pattern is $\top$. For example, discrimination pattern of our loan processing problem in Example 1 is of the form $\mathit{DP} := (\mathit{Female}(v), \top)$ that denotes female applicants as the protected group (i.e., $PG := \{v : \mathit{Female}(v)\}$) and male applicants as the unprotected group (i.e., $UG := \{v : \neg \mathit{Female}(v)\}$).

4 Fairness Measures

In this section, we formulate common fairness measures using the notations defined in Section 3. Let a and c denote the counts of denial (i.e., negative decisions) for protected and unprotected groups, and n_1 and n_2 denote their sizes, respectively. Given the decision atom $d(v)$, discriminative pattern $\mathit{DP}(F_1[v], F_2[v])$, and interpretation I, these counts are computed by the following equations:

$$a \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge F_1[v] \wedge F_2[v]\big) \tag{1}$$

$$c \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge \neg F_1[v] \wedge F_2[v]\big) \tag{2}$$

$$n_1 \equiv \sum_{v \in D_v} I\big(F_1[v] \wedge F_2[v]\big) \tag{3}$$

$$n_2 \equiv \sum_{v \in D_v} I\big(\neg F_1[v] \wedge F_2[v]\big) \tag{4}$$

The proportions of denying for protected and unprotected groups are $p_1 = \frac{a}{n_1}$ and $p_2 = \frac{c}{n_2}$, respectively. Existing data-driven fairness measures (Pedreschi, Ruggieri, and Turini 2012) can be defined in terms of p_1 and p_2 based on our relational notions as follows:

1. **Risk difference**: $RD = p_1 - p_2$, also known as absolute risk reduction. The UK uses RD as its legal definition of fairness measure (UKl).
2. **Risk Ratio**: $RR = \frac{p_1}{p_2}$, also known as relative risk. The EU court of justice typically use the RR as a measure of fairness (EUl).
3. **Relative Chance**: $RC = \frac{1-p_1}{1-p_2}$ also, known as selection rate. The US laws and courts mainly use the RC as a measure of fairness (USl).

Notice that RR is the ratio of the proportion of benefit denial between the protected and unprotected groups, while RC is the ratio of the proportion of benefit granted. Finally, we introduce the notion of δ-fairness.

Definition 8 (δ-fairness)**.** *If a fairness measure for a decision making process falls within some* δ*-window, then the process is* δ-fair*. Given* $0 \leq \delta \leq 1$*, the* δ*-windows for measures RD/RR/RC are defined as:*

$$\begin{aligned} -\delta \leq & RD \leq \delta \\ 1-\delta \leq & RR \leq 1+\delta \\ 1-\delta \leq & RC \leq 1+\delta \end{aligned}$$

To overcome the limitations of attribute-based fairness, we introduce a new statistical relational learning (SRL) framework (Getoor and Taskar 2007) suitable for modelling fairness in relational domain. In the next Section, we review probabilistic soft logic (PSL). We then extend PSL with the definition of relational fairness introduced above in Section 6. Our fairness-aware framework which we refer to as "FairPSL" is the first SRL framework that performs fair inference.

5 Background: Probabilistic Soft Logic

In this section, we review the syntax and semantics of PSL, and in the next section we extend MAP inference in PSL with fairness constraints to define MAP inference in FairPSL.

PSL is a probabilistic programming language for defining hinge-loss Markov random fields (Bach et al. 2017). Unlike other SRL frameworks whose atoms are Boolean, atoms in PSL can take continuous values in the interval $[0, 1]$. PSL is as an expressive modeling language that can incorporate domain knowledge with first-order logical rules and has been used successfully in various domains, including bioinformatics (Sridhar, Fakhraei, and Getoor 2016), recommender systems (Kouki et al. 2015), natural language processing (Ebrahimi, Dou, and Lowd 2016), information retrieval (Alshukaili, Fernandes, and Paton 2016), and social network analysis (West et al. 2014), among many others.

A PSL model is defined by a set of first-order logical rules called *PSL rules*.

Definition 9 (PSL rule). *a PSL rule r is an expression of the form:*

$$\lambda_r : T_1 \wedge T_2 \wedge \ldots \wedge T_w \rightarrow H_1 \vee H_2 \vee \ldots \vee H_l \tag{5}$$

where $T_1, T_2, \ldots, T_w, H_1, H_2, \ldots, H_l$ are atoms or negated atoms and $\lambda_r \in \mathbb{R}^+ \cup \infty$ is the weight of the rule r. We call $T_1 \wedge T_2 \wedge \ldots \wedge T_w$ the body of r (r_{body}), and $H_1 \vee H_2 \vee \ldots \vee H_l$ the head of r (r_{head}).

Since atoms in PSL take on continuous values in the unit interval $[0, 1]$, next we define soft logic to calculate the value of the PSL rules under an interpretation I.

Definition 10 (Soft logic). *The ($\tilde{\wedge}$) and ($\tilde{\vee}$) and negation ($\tilde{\neg}$) are defined as follows. For $m, n \in [0, 1]$ we have: $m \tilde{\wedge} n = \max(m + n - 1, 0)$, $m \tilde{\vee} n = \min(m + n, 1)$ and $\tilde{\neg} m = 1 - m$. The $\tilde{}$ indicates the relaxation over Boolean values.*

The probability of truth value assignments in PSL is determined by the rules' *distance to satisfaction*.

Definition 11 (The distance to satisfaction). *The distance to satisfaction $d_r(I)$ of a rule r under an interpretation I is defined as:*

$$d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\} \tag{6}$$

By using Definition 10, one can show that the closer the interpretation of a grounded rule r is to 1, the smaller its distance to satisfaction. A PSL model induces a distribution over interpretations I. Let R be the set of all grounded rules, then the probability density function is:

$$f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \tag{7}$$

where λ_r is the weight of rule r, $Z = \int_I \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p]$ is a normalization constant, and $p \in \{1, 2\}$ provides a choice of two different loss functions, $p = 1$ (i.e., linear), and $p = 2$ (i.e, quadratic). These probabilistic models are instances of hinge-loss Markov random fields (HL-MRF) (Bach et al. 2017). The goal of maximum a posteriori (MAP) inference is to find the most probable truth assignments I_{MPE} of unknown ground atoms given the evidence which is defined by the interpretation I. Let X be all the evidence, i.e., X is the set of ground atoms such that $\forall x \in X, I(x)$ is known, and let Y be the set of ground atoms such that $\forall y \in Y, I(y)$ is unknown. Then we have

$$I_{MAP}(Y) = arg \max_{I(Y)} P(I(Y)|I(X)) \tag{8}$$

Maximizing the density function in Equation 7 is equivalent to minimizing the weighted sum of the distances to satisfaction of all rules in PSL.

Example 8. *The simplified PSL model for the paper reviewing problem in Example2 is given in Table 1. The goal of MAP inference for this problem is to infer authors whom present at the conference (i.e., authors with an accepted paper) given the paper summaries. We simplified the model by assigning the same weight to rules $R1$ to $R6$ (i.e., $\lambda_i = 1$ where $i = \{1, 2, 3, 4, 5, 6\}$). Bellow we explain the meaning of each rule in the model.*

Rule $R1$ indicates that having a positive summary is resulted from two positive reviews and similarly rule $R2$ expresses that a negative summary of a paper is derived from negative reviews. Rule $R3$ indicates that submitting a positive review reduces the chances of another positive review by the same reviewer. And rules $R4$ and $R5$ indicate that high quality papers receive positive reviews. Rule $R6$ indicates the prior that not all submitted papers are high quality papers. We also have two hard constraints (i.e., rules $R7$ and $R8$) where weight of the rules are ∞. These two rules show that a high quality paper should get accepted and the author presents the accepted paper at the conference.

6 Fairness-aware PSL (FairPSL)

The standard MAP inference in PSL aims at finding values that maximize the conditional probability of unknowns. Once a decision is made according to these values, one can use the fairness measure to quantify the degree of discrimination. A simple way to incorporate fairness in MAP inference is to add the δ-fairness constraints to the corresponding optimization problem.

Consider risk difference, *RD*, where $\mathit{RD} \equiv \frac{\mathbf{a}}{n_1} - \frac{\mathbf{c}}{n_2}$. The δ-fairness constraint $-\delta \leq \mathit{RD} \leq \delta$ can be encoded as the following constraints:

$$n_2\mathbf{a} - n_1\mathbf{c} - n_1 n_2 \delta \leq 0 \tag{9}$$

$$n_2\mathbf{a} - n_1\mathbf{c} + n_1 n_2 \delta \geq 0 \tag{10}$$

Similarly, from $\mathit{RR} \equiv \frac{\mathbf{a}/n_1}{\mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq \mathit{RR} \leq 1 + \delta$ we obtain:

$$n_2\mathbf{a} - (1 + \delta) n_1\mathbf{c} \leq 0 \tag{11}$$

$$n_2\mathbf{a} - (1 - \delta) n_1\mathbf{c} \geq 0 \tag{12}$$

And finally, $\mathit{RC} \equiv \frac{1 - \mathbf{a}/n_1}{1 - \mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq \mathit{RC} \leq 1 + \delta$ gives:

$$- n_2\mathbf{a} + (1 + \delta) n_1\mathbf{c} - \delta n_1 n_2 \leq 0 \tag{13}$$

$$- n_2\mathbf{a} + (1 - \delta) n_1\mathbf{c} + \delta n_1 n_2 \geq 0 \tag{14}$$

$R1$:	λ_1:	$\textit{PositiveSummary}(p) \wedge \textit{PositiveReview}(r2, p) \wedge (r1 \neq r2) \rightarrow \textit{PositiveReview}(r1, p)$
$R2$:	λ_2:	$\neg\textit{PositiveSummary}(p) \wedge \textit{Reviews}(r, p) \rightarrow \neg\textit{PositiveReview}(r, p)$
$R3$:	λ_3:	$\textit{PositiveReview}(r, p1) \wedge \textit{Reviews}(r, p2) \rightarrow \neg\textit{PositiveReview}(r, p2)$
$R4$:	λ_4:	$\textit{HighQuality}(p) \wedge \textit{Reviews}(r, p) \rightarrow \textit{PositiveReview}(r, p)$
$R5$:	λ_5:	$\neg\textit{HighQuality}(p) \wedge \textit{Reviews}(r, p) \rightarrow \neg\textit{PositiveReview}(r, p)$
$R6$:	λ_6:	$\neg\textit{HighQuality}(p)$
$R7$:	∞:	$\textit{HighQuality}(p) \wedge \textit{Submits}(a, p) \rightarrow \textit{Presents}(a)$
$R8$:	∞:	$\neg\textit{HighQuality}(p) \wedge \textit{Submits}(a, p) \rightarrow \neg\textit{Presents}(a)$

Table 1: A simplified PSL model for the *Paper Reviewing* problem

A primary advantage of PSL over similar frameworks is that its MAP inference task reduces to a convex optimization problem which can be solved in polynomial time. To preserve this advantage, we need to ensure that the problem will remain convex after the addition of δ-fairness constraints.

Theorem 1. *The following condition is sufficient for preserving the convexity of MAP inference problem after addition of δ-fairness constraints: The formulae $F_1[v]$ and $F_2[v]$ do not contain an atom $y \in Y$ and all atoms in $F_1[v]$ and $F_2[v]$ have values zero or one.*

Proof. Since $I(F_1[v])$ and $I(F_2[v])$ do not depend on $I(Y)$, the values n_1 and n_2 are constants that can be computed in advance. Let us define the sets $D_v^a = \{v \in D_v : F_1[v] \wedge F_2[v] \text{ is true}\}$ and $D_v^c = \{v \in D_v : \neg F_1[v] \wedge F_2[v] \text{ is true}\}$. Since $F_1[v]$ and $F_2[v]$ can be only zero or one, we can rewrite the equations 1 and 2 as:

$$\mathbf{a} = \sum_{v \in D_v^a} I(\neg d(v)) = |D_v^a| - \sum_{v \in D_v^a} I(d(v))$$

$$\mathbf{c} = \sum_{v \in D_v^c} I(\neg d(v)) = |D_v^c| - \sum_{v \in D_v^c} I(d(v))$$

which indicates that $\mathbf{a}$ and $\mathbf{c}$ can be expressed as linear combinations of variables in the optimization problem. This means that constraints 9-14 are linear. Hence, addition of these constraints preserves the convexity of the optimization problem. □

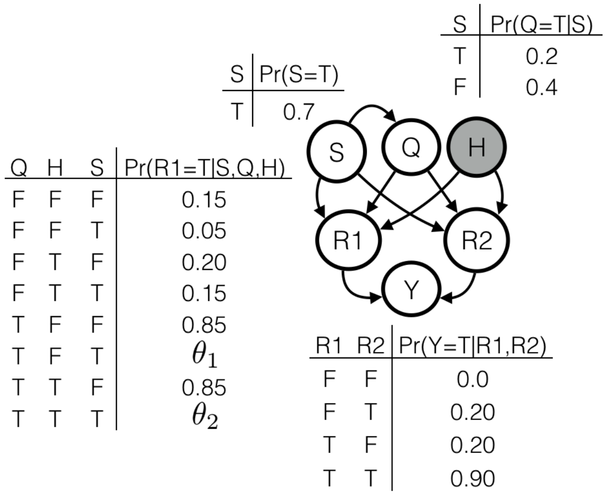
7 Experiments

We show the effectiveness of FairPSL by performing an empirical evaluation. We investigate two research questions in our experiments:

Q1 What is the effect of the fairness threshold δ on the fairness measures $RD/RC/RR$?

Q2 How are the accuracy of predictions affected by imposing the δ-fairness constraints?

We implemented the MAP inference routines of PSL and FairPSL in python using the convex optimization library CVXPY[‡] (Diamond and Boyd 2016). The FairPSL code, code for the data generator and data will be made publicly available upon publication of the paper.



S	Pr(S=T)
T	0.7

S	Pr(Q=T|S)
T	0.2
F	0.4

Q	H	S	Pr(R1=T|S,Q,H)
F	F	F	0.15
F	F	T	0.05
F	T	F	0.20
F	T	T	0.15
T	F	F	0.85
T	F	T	θ_1
T	T	F	0.85
T	T	T	θ_2

R1	R2	Pr(Y=T|R1,R2)
F	F	0.0
F	T	0.20
T	F	0.20
T	T	0.90

Figure 1: The model used for generating the datasets. There are five binary random variables, S, Q, H, R1, R1 and Y. **S**: indicates whether or not the author is a student; **Q**: indicates whether or not the paper is high quality; **H**: indicates whether or not the author is affiliated with a top-rank institute (observed); **R1, R2**: indicates whether or not the first/second reviewer gives the paper a positive review; **Y**: indicates whether or not the area chair writes a positive summary.

Data generation

We evaluate the FairPSL inference algorithm on synthetic datasets representing the paper reviewing scenario (introduced in Example 2). The parameters for the synthetic data generator are set as follows: we assume that there are 20 institutes. Each institute is equally likely to be top-rank or not. The number of papers submitted from each institute follows a *binomial*$(10, 0.5)$ distribution. The number of reviewers is assumed to be 30. For each paper, two reviewers are randomly assigned (If the two reviewers happen to be the same, we repeat this procedure).

For each paper, we use the generative model of Figure 1 to draw assignments for all the random variables. Note that the conditional probability table for the review variable R_i is parameterized by two values (θ_1, θ_2) which together determine the degree of discrimination against the protected group. We generated three datasets using these values $(0.3, 0.7)$, $(0.6, 0.6)$, and $(0.5, 0.9)$ for parameters (θ_1, θ_2). Each dataset has around 100 papers (i.e., dataset

[‡] https://cvxgrp.github.io/cvxpy/

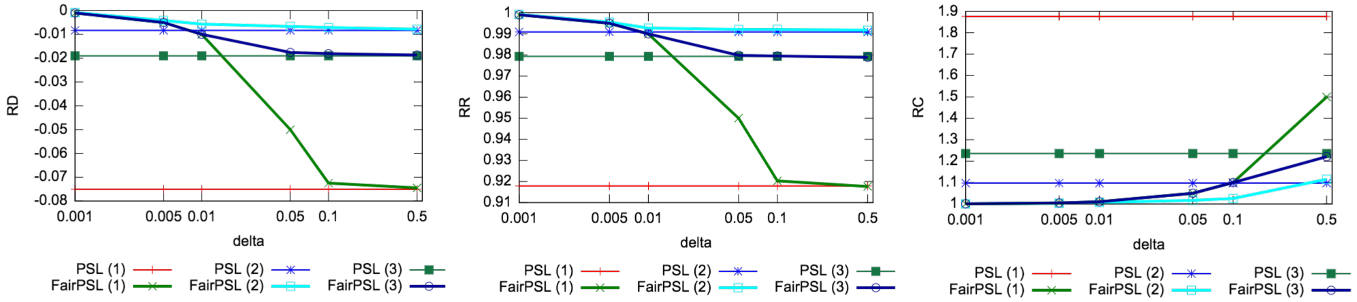


Figure 2: Score of predictions obtained by MAP inference of PSL and FairPSL, according to the fairness measures *RD*, *RR*, and *RC*. The labels of datasets are mentioned with parenthesis next to the inference method. The FairPSL values of each measure are obtained by adding the δ-fairness constraint of that measure.

#1 has 102 papers, dataset #2 has 109 papers and dataset #3 has 101 papers).

MAP Inference We use the model presented in Table 1 for MAP inference in PSL and FairPSL (Recall that in FairPSL, the δ-fairness constraints corresponding to one of the fairness measures are also added to the model). The only observed atoms are *PositiveSummary(P)*. The truth values for all other atoms is obtained via MAP inference. We use the truth values obtained for the decision atoms *Presents(A)* to compute the fairness measures. We defined the discriminative pattern, the protected and unprotected groups of this problem in Section 3.

Evaluation results

To answer **Q1**, we run the MAP inference algorithm of PSL and FairPSL on three synthetic datasets. We run the MAP inference of FairPSL multiple times on each dataset: For each of the three fairness measures, we add the corresponding δ-fairness constraint with five thresholds $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

Figure 2 shows the score of predictions in terms of the three fairness measures. As expected, tighter δ-fairness constraints lead to better scores. Note that the best possible score according to RD is zero, as it computes a difference. Since RR and RC compute ratios, the best possible score according to these measures is one. In our experiments, with any of these measures, taking $\delta = 0.001$ pushes the score of predictions to its limit.

The δ-fairness constraints modify the optimization problem of MAP inference by reducing its feasible region to solutions that conform with fairness guarantees. Research question **Q2** is concerned with the effect of this reduction on the accuracy of predictions. To answer this question, we compare the inferred values for the decision atoms *Presents(A)* against their actual values. These values are extracted from the known values of *HighQuality(p)* according to rules 7 and 8 in Table 1. The results which are presented in Table 2, include the area under the curve of the receiver operating characteristic (ROC) of predicting the decision variable for 4 groups. These groups are the protected group (i.e., acceptance of the students who are not affiliated with top-rank institute), the unprotected group (i.e., acceptance of the students who are affiliated with top-rank institute), the rest of the authors (i.e., acceptance of the non-students), and all authors. By doing so, we not only calculate the accuracy of the prediction for the decision atom of the protected and unprotected groups, but also we calculate it for other authors to make sure that our fairness constraints do not propagate bias towards other authors. The results of FairPSL with δ-fairness constraints RR and RC are either equal or very close to the results of FairPSL with δ-fairness constraints RD, therefore due to the space limitation, we omit them from this table. According to Table 2, the results of both PSL and FairPSL in all three datasets are very close to each other. Although the predictions of FairPSL for dataset#1 and dataset#3 are better than PSL, it is not definitive that FairPSL improves the predictions. We observe that prediction of MAP inference for both FairPSL and PSL are similar, thus FairPSL guarantees fairness without hurting accuracy. Further investigation is required on the effect of the various discrimination ranges on the prediction results of FairPSL (i.e., θ_1 and θ_2), which remains for our future work.

Dataset	Approach	Protected	Unprotected	Non-students	All
#1	*PSL*	0.935	0.722	0.882	0.834
	FairPSL(RD)	0.924	0.726	0.887	0.887
#2	*PSL*	0.575	0.944	0.877	0.810
	FairPSL(RD)	0.541	0.838	0.829	0.776
#3	*PSL*	0.926	0.727	0.701	0.762
	FairPSL(RD)	0.931	0.737	0.769	0.793

Table 2: ROC of predictions for truth values of unknown atoms *Presents(A)* using MAP inference of PSL and FairPSL with δ-fairness constraints RD with $\delta = 0.001$.

8 Conclusion and Future Direction

Many applications of AI and machine learning affect peoples' lives in important ways. While there is a growing body of work on fairness in AI and ML, it assumes an individualistic notion of fairness. In this paper, we have proposed a general framework for relational fairness which includes both a rich language for defining discrimination patterns and an efficient algorithm for performing inference subject to fairness constraints. We show our approach enforces fariness guarantees while preserving the accuracy of the predictions.

There are many avenues for expanding on this work. For example, here we assumed that the discriminative pattern is given, however an automatic mechanism to extract discriminatory situations hidden in a large amount of decision records is an important and required task. Discrimination discovery has been studied for attribute-based fairness (Pedreschi, Ruggieri, and Turini 2013). An interesting next step is discrimination pattern discovery in relational data.

References

[Alshukaili, Fernandes, and Paton 2016] Alshukaili, D.; Fernandes, A. A. A.; and Paton, N. W. 2016. Structuring linked data search results using probabilistic soft logic. In *International Semantic Web Conference (ISWC)*.

[Bach et al. 2017] Bach, S. H.; Broecheler, M.; Huang, B.; and Getoor, L. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*.

[Barocas and Selbst 2016] Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact.

[Boyd, Levy, and Marwick 2014] Boyd, D.; Levy, K.; and Marwick, A. 2014. The networked nature of algorithmic discrimination. In Gangadharan, S. P., ed., *Data and discrimination: Collected essays*. Washington, DC: New America. 53–57.

[Choi, Darwiche, and den Broeck 2017] Choi, Y.; Darwiche, A.; and den Broeck, G. V. 2017. Optimal feature selection for decision robustness in bayesian networks. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 1554–1560. ijcai.org.

[Chouldechova 2017] Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

[Diamond and Boyd 2016] Diamond, S., and Boyd, S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17(83):1–5.

[Dwork et al. 2012] Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

[Ebrahimi, Dou, and Lowd 2016] Ebrahimi, J.; Dou, D.; and Lowd, D. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[EUl] European union legislation. (a) racial equality directive, 2000; (b) employment equality directive, 2000; (c) gender employment directive, 2006; (d) equal treatment directive (proposal), 2008.

[Feldman et al. 2015] Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

[Getoor and Taskar 2007] Getoor, L., and Taskar, B. 2007. *Introduction to Statistical Relational Learning*. MIT press.

[Hardt et al. 2016] Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

[Kamishima, Akaho, and Sakuma 2011] Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 643–650. IEEE.

[Kouki et al. 2015] Kouki, P.; Fakhraei, S.; Foulds, J.; Eirinaki, M.; and Getoor, L. 2015. HyPER: A flexible and extensible probabilistic framework for hybrid recommender systems. In *ACM Conference on Recommender Systems (RecSys)*.

[Kusner et al. 2017] Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

[Pedreschi, Ruggieri, and Turini 2012] Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, 126–131. New York, NY, USA: ACM.

[Pedreschi, Ruggieri, and Turini 2013] Pedreschi, D.; Ruggieri, S.; and Turini, F. 2013. The discovery of discrimination. *Discrimination and privacy in the information society* 91–108.

[Sridhar, Fakhraei, and Getoor 2016] Sridhar, D.; Fakhraei, S.; and Getoor, L. 2016. A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics* 32(20):3175–3182.

[UKl] Uk legislation. (a) sex discrimination act, 1975, (b) race relation act, 1976.

[USl] United nations legislation. (a) universal declaration of human rights, 1948, (c) convention on the elimination of all forms of racial discrimination, 1966, (d) convention on the elimination of all forms of discrimination against women, 1979.

[West et al. 2014] West, R.; Paskov, H. S.; Leskovec, J.; and Potts, C. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics (TACL)* 2:297–310.

[Zafar et al. 2017] Zafar, M. B.; Valera, I.; Rodriguez, M. G.; Gummadi, K. P.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. *arXiv preprint arXiv:1707.00010*.

[Zemel et al. 2013] Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 325–333.