A Declarative Approach to Fairness in Relational Domains

Golnoosh Farnadi[1,2], Behrouz Babaki[1], Lise Getoor[3]
[1]Polytechnique Montréal, [2] Mila, [3] UC Santa Cruz
farnadig@mila.quebec, behrouz.babaki@polymtl.ca, getoor@soe.ucsc.edu

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 financial qualifications. 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 violations 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 the 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 [14, 7, 16]. Fairness has been well studied in the social sciences and legal scholarship (for an in-depth review see [6]), and there is emerging work on fairness-aware ML within the AI and computer science communities. For example, fairness through awareness/Lipschitz property [11], individual fairness [27], statistical parity/group fairness [17], counterfactual fairness [19], demographic parity/disparate impact [14, 10], preference-based fairness [26], and equality of opportunity [16].

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

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 sensitive 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.[1] 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 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 is also widely used for learning in relational domains. Moreover, we extend an existing framework for statistical relational learning [15] called probabilistic soft logic (PSL)[2] [5]. PSL combines logic and probability for learning and reasoning over uncertain relational domains. One of the most common reasoning tasks 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 guarantees. An early version of this paper which this work builds upon and extends appeared in [13].

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

Discrimination in social contexts have been studied in the field of social psychology [9, 8, 22]. There is a large literature on various aspects of relational bias in social contexts such as *in-group-out-group bias*, *gender bias*, and *ethnicity-based favoritism* that can result in discrimination. As an example, consider gender bias in the workplace that reflects stereotypically masculine criteria and male-based favoritism. Such gender bias typically places women in lower positions and negatively impacts their opportunities. Further, lack of women in leadership positions may affect the promotion of women and results in a glass ceiling that keeps women from rising beyond a certain level in the hierarchy. This scenario shows that considering protected attributes such as gender is not always sufficient to detect the source of bias and avoid discrimination, one also has to consider the relational information, in this case the organization hierarchy. Note that this can be generalized to any ingroup/outgroup scenario where the sensitive attribute could be race, religion, age, marital-status, etc.

The existing work on designing fair algorithms in machine learning exclusively focuses 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 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

[1] 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.

[2] http://psl.linqs.org/

of the protected group are subject to a fair decision (we discuss different fairness measures in Section 4). We illustrate attribute-based fairness in 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 whether to grant a loan to an applicant using a predictive model. The bank needs to ensure that the obey fair lending practices and ensure that gender, race, sexual orientation of applicants has no influence on the decision. In this scenario, the protected group is the historically disadvantaged applicants.

The current fairness-aware machine learning techniques are not capable of modeling relations and hence cannot be used to make the decision making model fair. However, in many decision making scenarios, especially in social and organizational settings, the domain is relational, and the protected group itself might be best represented using a relational definition. We illustrate this setting in the following scenario:

Example 2 (Performance Review): Consider an organization where decisions about the promotion of employees is based on two criteria: 1) an objective performance measure, and 2) the opinion of their direct and indirect managers above them. The opinions are inferred from the performance reviews which are collected periodically. Not every manager can submit a review for all its subordinates, this is especially the case for top-level managers who have a large number of subordinates. Hence, the opinions of managers are collectively inferred from the opinions of their sub-ordinates. However, some employees may be biased, and judge other employees unfavorably, by favoring employees who are similar to themselves (same gender, race, religion, etc.) over employees who are dissimilar. The organization needs to ensure that promotion of employees do not have any relational bias caused by in-group-out-group favoritism.

Example 2 describes a prediction problem over a database that consists of relations between employees. Such prediction tasks are best handled by techniques from the relational learning domain. To ensure fair prediction in such settings, we need to extend the notion of *attribute-based fairness* to *relational fairness*. Throughout this paper, we use the performance review 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 an 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 performance review problem in Example 2 the atom $Manager(m, e)$ indicates whether or not employee m is a direct or indirect manager of employee e.

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 $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$, $\alpha \rightarrow \beta$ 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 $F[v] := \forall u,\ \textit{Manager}(u, v) \to \neg\textit{SameGroup}(u, v)$. The population specified by this formula is the set of individuals all of whose managers belong to a group different from theirs.

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 a *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 $\textit{DP}[v] := (F_1[v], F_2[v])$, where $F_1[v]$ and $F_2[v]$ are formulae.

Example 5: The two formulae in the discrimination pattern $\textit{DP}[v] := \big((\forall u,\ \textit{Manager}(u, v) \to \neg\textit{SameGroup}(u, v)),\ \textit{Employee}(v)\big)$ specify two populations, namely all employees and those employees who belong to a group different from their managers.

Given the definition of the discriminative pattern, we have a rich language to define the scope of the protected and unprotected groups 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 $PG := \big\{v : \big(\forall u,\ \textit{Manager}(u, v) \to \neg\textit{SameGroup}(u, v)\big) \wedge \textit{Employee}(v)\big\}$ and the unprotected group is $UG := \big\{v : \big(\exists u,\ \textit{Manager}(u, v) \wedge \textit{SameGroup}(u, v)\big) \wedge \textit{Employee}(v)\big\}$. This means our protected group is the set of employees belonging to a group different from their managers, and our unprotected group consists of other employees.

Discrimination is defined in terms of a treatment or decision that distinguishes between the protected and unprotected groups. Here we define the *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 $ToPromote(v)$ indicates whether or not v receives a promotion.

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 *Applicant*(v). For example, discrimination pattern of our loan processing problem in Example 1 is of the form *DP* $:= (\textit{Female}(v), \textit{Applicant}(v))$ that denotes female applicants as the protected group (i.e., $PG := \{v : \textit{Female}(v)\}$) and male applicants as the unprotected group (i.e., $UG := \{v : \neg\textit{Female}(v)\}$).

4 Fairness Measures

Over the past few years, many fairness measures have been introduced [24]. An important class of these measures are *group fairness* measures which quantify the inequality between different subgroups. Some of the most popular measures in this class include *equal opportunity*, *equalized odds*, and *demographic parity* [16]. In this paper we restrict our focus to the latter. In an attribute-value setting, demographic parity means that the decision should be independent of the protected attributes. Assume that binary variables A and C denote the decision and protected attributes, and the preferred value of A is one. Demographic parity requires that:

$$P(A = 1|C = 0) = P(A = 1|C = 1)$$

We will now generalize this measure to the relational setting 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 *DP*$(F_1[v], F_2[v])$, and interpretation I, these counts are computed by the following equations:

$$a \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge F_1[v] \wedge F_2[v]\big) \tag{5}$$

$$c \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge \neg F_1[v] \wedge F_2[v]\big) \tag{6}$$

$$n_1 \equiv \sum_{v \in D_v} I\big(F_1[v] \wedge F_2[v]\big) \tag{7}$$

$$n_2 \equiv \sum_{v \in D_v} I\big(\neg F_1[v] \wedge F_2[v]\big) \tag{8}$$

The proportions of denying for protected and unprotected groups are $p_1 = \frac{a}{n_1}$ and $p_2 = \frac{c}{n_2}$, respectively. There are a number of data-driven measures [20] which quantify demographic disparity and can be defined in terms of p_1 and p_2:

- **Risk difference**: $RD = p_1 - p_2$, also known as absolute risk reduction.
- **Risk Ratio**: $RR = \frac{p_1}{p_2}$, also known as relative risk.
- **Relative Chance**: $RC = \frac{1-p_1}{1-p_2}$ also, known as selection rate.

These measures have been used in the legal systems of European Union, UK, and US [1, 2, 3]. 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 [15] 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, "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 [5]. Unlike other SRL frameworks whose atoms are Boolean, atoms in PSL can take continuous values in the interval $[0, 1]$. PSL is an expressive modeling language that can incorporate domain knowledge with first-order logical rules and has been used successfully in various domains, including bioinformatics [23], recommender systems [18], natural language processing [12], information retrieval [4], and social network analysis [25], 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:

$$\lambda_r : T_1 \wedge T_2 \wedge \ldots \wedge T_w \rightarrow H_1 \vee H_2 \vee \ldots \vee H_l \tag{9}$$

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:

$$d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\} \tag{10}$$

$R1$	λ_1	: $\mathit{IsQualified}(e) \rightarrow \mathit{HighPerformance}(e)$
$R2$	λ_1	: $\neg \mathit{IsQualified}(e) \rightarrow \neg \mathit{HighPerformance}(e)$
$R3$	∞	: $\mathit{PositiveReview}(e1, e2) \rightarrow \mathit{PositiveOpinion}(e1, e2)$
$R4$	∞	: $\neg \mathit{PositiveReview}(e1, e2) \rightarrow \neg \mathit{PositiveOpinion}(e1, e2)$
$R5$	λ_1	: $\mathit{PositiveOpinion}(e1, e2) \wedge \mathit{Manager}(m, e1) \rightarrow \mathit{PositiveOpinion}(m, e2)$
$R6$	λ_1	: $\neg \mathit{PositiveOpinion}(e1, e2) \wedge \mathit{Manager}(m, e1) \rightarrow \neg \mathit{PositiveOpinion}(m, e2)$
$R7$	λ_1	: $\mathit{PositiveOpinion}(m, e) \wedge \mathit{Manager}(m, e) \rightarrow \mathit{IsQualified}(e)$
$R8$	λ_1	: $\neg \mathit{PositiveOpinion}(m, e) \wedge \mathit{Manager}(m, e) \rightarrow \neg \mathit{IsQualified}(e)$
$R9$	λ_1	: $\neg \mathit{ToPromote}(e)$
$R10$	∞	: $\mathit{IsQualified}(e) \rightarrow \mathit{ToPromote}(e)$
$R11$	∞	: $\neg \mathit{IsQualified}(e) \rightarrow \neg \mathit{ToPromote}(e)$

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

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:

$$f(I) = \frac{1}{Z} \exp[- \sum_{r \in R} \lambda_r (d_r(I))^p] \tag{11}$$

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) [5]. 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

$$I_{MAP}(Y) = arg \max_{I(Y)} P(I(Y)|I(X)) \tag{12}$$

Maximizing the density function in Equation 11 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 performance reviewing problem in Example2 is given in Table 1. The goal of MAP inference for this problem is to infer employees to promote. We simplified the model by assigning the same weight to all soft rules (i.e., $\lambda_i = 1$ where $i = \{1, 2, 5, 6, 7, 8, 9\}$). Below we explain the meaning of each rule in the model.

Rule $R1$ indicates that qualified employees have high performance and similarly rule $R2$ expresses that a negative qualification of employees is derived from their low performance. Rules $R5$ and $R6$ presents the propagation of opinion from bottom to top of the organizational hierarchy, i.e., managers have similar opinions towards employees given the opinions of their sub-ordinate managers. And rules $R7$ and $R8$ indicate that the positive/negative opinion of direct/indirect managers derive from the qualification of an employee. Rule $R9$ indicates the prior that not all employees get promoted. We also have four hard constraints (i.e., rules $R3$, $R4$, $R10$ and $R11$) where the weight of the rules are ∞. Rules $R3$ and $R4$ indicate that submitted positive/negative reviews should reflect positive/negative opinions. And two rules $R10$ and $R11$ show that a highly qualified employee should get promoted.

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 $\textit{RD} \equiv \frac{\mathbf{a}}{n_1} - \frac{\mathbf{c}}{n_2}$. The δ-fairness constraint $-\delta \leq \textit{RD} \leq \delta$ can be encoded as the following constraints:

$$n_2\mathbf{a} - n_1\mathbf{c} - n_1 n_2 \delta \leq 0 \tag{13}$$

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

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

$$n_2\mathbf{a} - (1+\delta) n_1\mathbf{c} \leq 0 \tag{15}$$

$$n_2\mathbf{a} - (1-\delta) n_1\mathbf{c} \geq 0 \tag{16}$$

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

$$- n_2\mathbf{a} + (1+\delta) n_1\mathbf{c} - \delta n_1 n_2 \leq 0 \tag{17}$$

$$- n_2\mathbf{a} + (1-\delta) n_1\mathbf{c} + \delta n_1 n_2 \geq 0 \tag{18}$$

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 5 and 6 as:

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

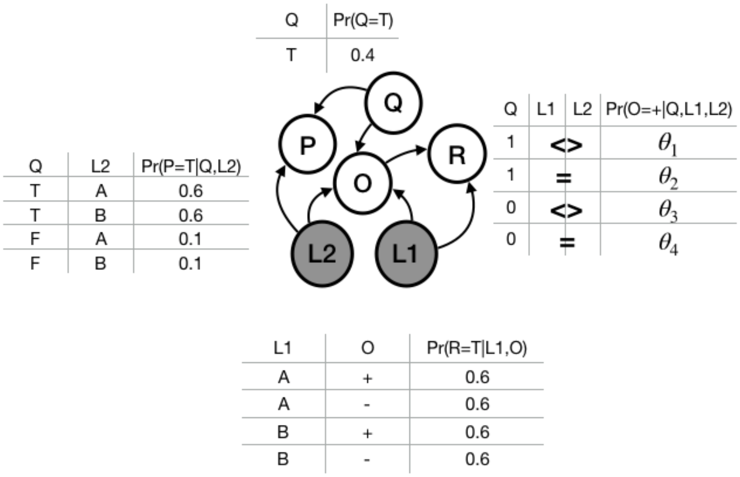
$$\mathbf{c} = \sum_{v \in D_v^c} I(\neg d(v)) = |D_v^c| - \sum_{v \in D_v^c} I(d(v))$$

which indicates that $\mathbf{a}$ and $\mathbf{c}$ can be expressed as linear combinations of variables in the optimization problem. This means that constraints 13-18 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 $RD/RC/RR$?



Q	Pr(Q=T)
T	0.4

Q	L2	Pr(P=T|Q,L2)
T	A	0.6
T	B	0.6
F	A	0.1
F	B	0.1

Q	L1	L2	Pr(O=+|Q,L1,L2)
1	<>		θ_1
1	=		θ_2
0	<>		θ_3
0	=		θ_4

L1	O	Pr(R=T|L1,O)
A	+	0.6
A	-	0.6
B	+	0.6
B	-	0.6

Figure 1: The model used for generating the datasets. There are four binary random variables, P, Q, O, and R. **P**: indicates whether or not the employee has high performance; **Q**: indicates whether or not an employee has high qualification; **O**: indicates whether or not the colleague submits the positive opinion towards the employee; **R**: indicates whether or not the colleague has a positive opinion towards the employee; **L1, L2**: indicates the label of the review provider and review receiver (observed).

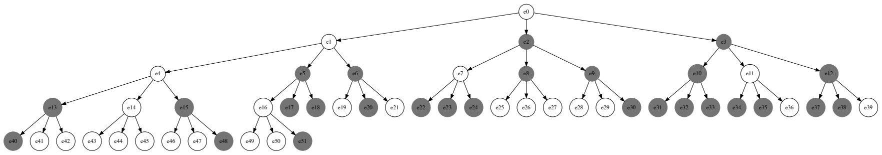


Figure 2: An example of an organizational hierarchy with five levels and 50 employees with k=3. Each employee either has label A (shown with grey) or B (shown with white).

Q2 How is decision quality affected by imposing δ-fairness constraints?

Note that although we present the result for specific parameters of the framework in this section, we ran extensive analysis and the results we present are representative. We implemented the MAP inference routines of PSL and FairPSL in Python, using Gurobi-8.1[3] as the backend solver. The FairPSL code, code for the data generator and data is publicly available[4].

7.1 Data generation

We evaluate the FairPSL inference algorithm on synthetic datasets representing the performance review scenario (introduced in Example 2). The organization hierarchy is generated synthetically. The organization hierarchy generator is parameterized by two numbers: the number of employees in the organization (n) and the number of employees managed by each manager (k). Each employee is randomly assigned with a label *A* or *B*. An examples organization hierarchy with n=50 and k=3 is shown in Figure 2.

For each employee, we use the generative model of Figure 1 to draw assignments for all the random variables. We assume that only 40% of employees are qualified for promotion and regardless of their labels, employees submit only 60% of their opinions. In addition, due to various personal and environmental factors, only 60% of high quality employees perform well while 10% of low quality employees also perform well regardless of their labels. Note that these numbers are not specific and just chosen for the framework to serve as a representative setting and a proof of concept. The conditional probability table for the opinion variable O is parameterized by four values $(\theta_1, \theta_2, \theta_3, \theta_4)$ which together determine the degree of discrimination against the protected group. Since other parameters in the Bayesian network did not have a direct effect on the degree of discrimination, we fixed them to arbitrary values.

The results presented in this section are based on an organization hierarchy with 100 employees where $k = 5$. However, the results of the framework are not sensitive to the settings as we test the framework

[3] www.gurobi.com

[4] https://github.com/gfarnadi/FairPSL

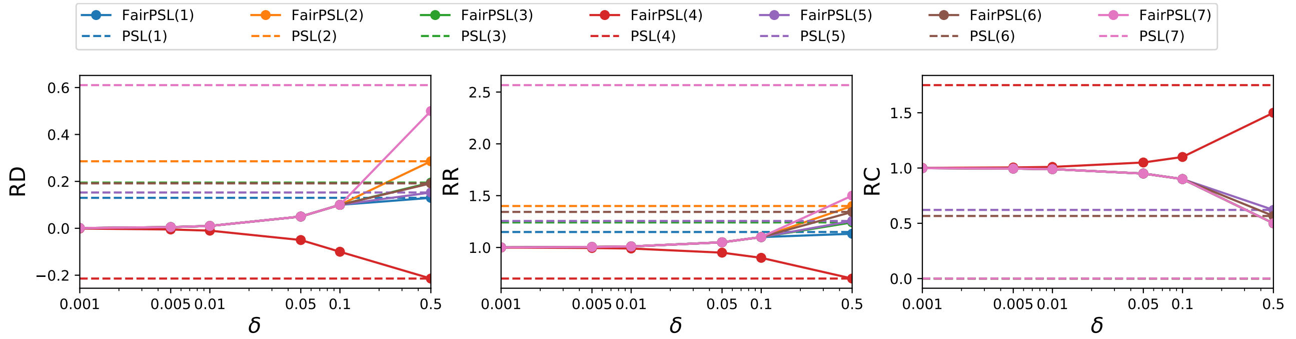


Figure 3: Fairness score of predictions obtained by MAP inference of PSL and FairPSL, according to the fairness measures *RD*, *RR*, and *RC*. The labels of datasets are mentioned with parenthesis next to the inference method. The FairPSL values of each measure are obtained by adding the δ-fairness constraint of that measure.

with various organization sizes ranging from 50 to 500 employees and various degree for k ranging from 3 to 10. We generated seven datasets given the organization hierarchy using different values for the θ parameters: $(0.0, 1.0, 0.0, 0.0)$, $(0.33, 1.0, 0.0, 0.0)$, $(0.66, 1.0, 0.0, 0.0)$, $(1.0, 1.0, 0.0, 0.0)$, $(1.0, 1.0, 0.0, 0.33)$, $(1.0, 1.0, 0.0, 0.66)$, $(1.0, 1.0, 0.0, 1.0)$.

In the first three settings the discrimination originates from negative opinions towards qualified outgroup employees. The first setup is an extreme case where the opinion towards outgroup employees is always negative. The discrimination in the last three settings originates from positive opinions towards unqualified ingroup employees. The last setup is an extreme case where the opinion towards ingroup employees is always positive. The fourth setup represent unbiased opinions where employees are treated similarly based on their qualification.

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 observed atoms are *Manager(m,e)*, *PositiveReview(e1,e2)* and labels of all employees. The truth values for all other atoms are obtained via MAP inference. We use the truth values obtained for the decision atoms *ToPromote(e)* to compute the fairness measures. We defined the discriminative pattern, and the protected and unprotected groups of this problem in Section 3.

7.2 Evaluation results

To answer **Q1**, we run the MAP inference algorithm of PSL and FairPSL on seven 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 3 shows the fairness 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 0, as it computes a difference. Since RR and RC compute ratios, the best possible score according to these measures is 1. 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 the 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. Note that decision quality is the same as the accuracy of predictions. To answer this question, we compare the inferred values for the decision atoms *ToPromote(e)* against their actual values. These values are extracted from the known values of *IsQualified(e)* according to rules 11 and 12 in Table 1. Figure 4 shows the area under the curve of the receiver operating characteristic (AUC) of predicting the decision variable in three groups, namely the protected group, the unprotected group (i.e., promotion of the employees

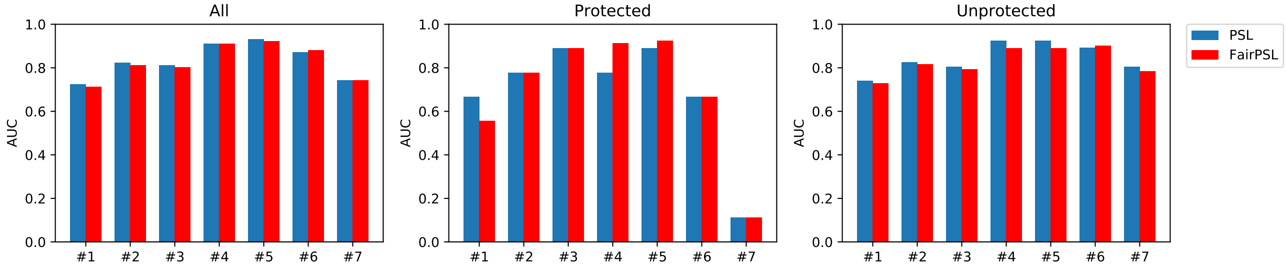


Figure 4: AUC score of predictions for truth values of unknown atoms *ToPromote(e)* using MAP inference of PSL and FairPSL with δ-fairness constraints RD with $\delta = 0.001$.

who have in-group managers), and all employees. By doing so, we make sure that our fairness constraints do not propagate bias towards either of the populations. Since the results of FairPSL with δ-fairness constraints RR and RC are very similar to the results of RD, we only report the latter here.

According to Figure 4, the results of both PSL and FairPSL in all seven datasets are close to each other. Note that although fairness may impose a cost in terms of overall accuracy, FairPSL often improves the accuracy of the protected class. Sometimes the overall predictions of FairPSL are even slightly better than PSL (e.g., dataset 6 and 7). As expected, the accuracy of the fourth setting where the opinions are unbiased are similar in both PSL and FairPSL. We observe that prediction of MAP inference for both FairPSL and PSL are similar, thus, in these settings at least, FairPSL guarantees fairness without hurting accuracy. Further investigation is required on the effect of the various ranges of discrimination (i.e., θ_1, θ_2, θ_3, θ_4) on the prediction results of FairPSL.

We also generate various types of organizations in which labels are not uniformly distributed, e.g., one population only occurs at the bottom levels of an organization. While we did not observe any differences in the behavior of our method with respect to accuracy and fairness measure, we found that the degree of discrimination is higher in such organizations. Further investigations on the structure of an organization on discrimination is an interesting direction for future research.

8 Conclusion and Future Directions

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 fairness 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 [21]. An interesting next step is discrimination pattern discovery in relational data.

Acknowledgements

This work is supported by the National Science Foundation under Grant Numbers CCF-1740850 and IIS-1703331. Golnoosh Farnadi and Behrouz Babaki are supported by postdoctoral scholarships from IVADO through the Canada First Research Excellence Fund (CFREF) grant.

References

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

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

[3] 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.

[4] Duhai Alshukaili, Alvaro A. A. Fernandes, and Norman W. Paton. Structuring linked data search results using probabilistic soft logic. In *International Semantic Web Conference (1)*, volume 9981 of *Lecture Notes in Computer Science*, pages 3–19, 2016.

[5] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18:109:1–109:67, 2017.

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

[7] Danah Boyd, Karen Levy, and Alice Marwick. The networked nature of algorithmic discrimination. In *Data and discrimination: Collected essays*, pages 53–57. 2014.

[8] Marilynn B Brewer. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307, 1979.

[9] Marilynn B Brewer. The social psychology of intergroup relations: Social categorization, ingroup bias, and outgroup prejudice. *Social Psychology: Handbook of Basic Principles*, 2007.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1703.00056, 2017.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[12] Javid Ebrahimi, Dejing Dou, and Daniel Lowd. Weakly supervised tweet stance classification by relational bootstrapping. In *EMNLP*, pages 1012–1017. The Association for Computational Linguistics, 2016.

[13] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 108–114. ACM, 2018.

[14] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.

[15] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT press Cambridge, 2007.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[17] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, pages 643–650. IEEE Computer Society, 2011.

[18] Pigi Kouki, Shobeir Fakhraei, James R. Foulds, Magdalini Eirinaki, and Lise Getoor. Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *RecSys*, pages 99–106. ACM, 2015.

[19] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pages 4069–4079, 2017.

[20] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *SAC*, pages 126–131. ACM, 2012.

[21] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. The discovery of discrimination. In *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 91–108. Springer, 2013.

[22] Cecilia L Ridgeway and Shelley J Correll. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & society*, 18(4):510–531, 2004.

[23] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics*, 32(20):3175–3182, 2016.

[24] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[25] Robert West, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *TACL*, 2:297–310, 2014.

[26] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *NIPS*, pages 228–238, 2017.

[27] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.