ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200389

Learning Fairness-Aware Relational Structures

Yue Zhang and **Arti Ramesh** [1]

Abstract. The development of fair machine learning models that effectively avert bias and discrimination is an important problem that has garnered attention in recent years. The necessity of encoding complex relational dependencies among the features and variables for competent predictions require the development of fair, yet expressive relational models. In this work, we introduce *Fair-A3SL*, a fairness-aware structure learning algorithm for learning relational structures, which incorporates fairness measures while learning relational graphical model structures. Our approach is versatile in being able to encode a wide range of fairness metrics such as statistical parity difference, overestimation, equalized odds, and equal opportunity, including recently proposed relational fairness measures. While existing approaches employ the fairness measures on pre-determined model structures post prediction, Fair-A3SL directly learns the structure while optimizing for the fairness measures and hence is able to remove any structural bias in the model. We demonstrate the effectiveness of our learned model structures when compared with the state-of-the-art fairness models quantitatively and qualitatively on datasets representing three different modeling scenarios: i) a relational dataset, ii) a recidivism prediction dataset widely used in studying discrimination, and iii) a recommender systems dataset. Our results show that Fair-A3SL can learn fair, yet interpretable and expressive structures capable of making accurate predictions.

1 INTRODUCTION

The widespread growth and prevalence of machine learning models for crucial decision-making tasks has raised questions on the fairness of the underlying models. Machine learning models have been mostly employed as a black box with little or no transparency or they are too complex to comprehend for non-experts, which further exacerbates this problem. This has led to an increased interest in creating fair machine learning models. The goal of fairness-aware machine learning is to ensure that the decisions made by models do not discriminate against a certain group(s) of individuals [12, 13, 4].

Fairness has been well studied in the social science and policy-making domains [3] and is emerging as an important area of research in computer science and specifically, the machine learning community. Most existing work on fairness focus on developing metrics to remove biases after prediction and identifying and removing sensitive attributes [13, 15, 22] . There is limited existing work on fairness in relational domains. Farnadi et al.'s [10] work on developing fairness metrics for relational domains and fairness-aware MAP inference for hinge-loss Markov random fields (HL-MRFs) [2] is the first work in this direction. Farnadi et al. [10] note that in many social contexts, discrimination is the result of complex interactions and cannot be described solely in terms of attributes of an individual. While this process is helpful in removing the biases in the inference procedure, it ignores the structural biases in the model structure. This is especially relevant for relational models, where the model structure is instrumental in obtaining the predictions and the biases ingrained in the structure are harder to detect and eliminate.

Contributions In this work, we develop Fair-A3SL, a fairness-aware structure learning algorithm for hinge-loss Markov random fields (HL-MRFs). Fair-A3SL extends a recently developed deep reinforcement learning-based structure learning algorithm for HL-MRFs, A3SL [26], to automatically learn *fair* relational graphical model structures. Fair-A3SL has the ability to encode almost all different state-of-the-art widely-used fairness metrics: equalized odds [13], equal opportunity [13], statistical parity difference [16], recently developed relational fairness measures of risk difference, risk reward, and relative chance [10], and fairness measures for collaborative filtering, non-parity and overestimation [23]. Fair-A3SL possesses the ability to encode multiple model-based and post-processing fairness measures in a single algorithm and can jointly optimize for them to learn a fair model structure. It also offers flexibility in encoding and enforcing these measures through user-defined coefficients that capture the impact of these measures, therefore providing the much needed customizability to enable applicability across multiple domains. The added strength of Fair-A3SL arises from its ability to learn interpretable fair structures that do not compromise on performance, further alleviating the problem of opaqueness and lack of interpretability in machine learning models. *To the best of our knowledge, ours is the first approach that directly focuses on learning fair relational model structures from data.*

In our experiments, we demonstrate Fair-A3SL's versatility in being able to encode many different fairness measures and learn fair models for multiple domains. We evaluate the effectiveness of our learned structures in three datasets: i) paper review dataset, a relational dataset used in Farnadi et al. [10] that showcases the ability of our models to learn fair network and collective model structures, ii) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset, a popular dataset used in many existing fairness work allowing us to compare Fair-A3SL with many state-of-the-art fairness models, and iii) MovieLens dataset, a popular dataset used in recommender systems, that enables us to integrate fairness measures used in collaborative filtering in Fair-A3SL. Fair-A3SL is able to learn structures that eliminate bias at the structure level, requires minimal pre-processing (no other pre-processing other than what is needed for computing the fairness metrics), and can potentially be used easily in sensitive applications to learn interpretable, expressive, and fair model structures that possess good prediction performance for making accurate predictions.

2 RELATED WORK

The state-of-the-art bias mitigation algorithms can be grouped into three categories that include pre-processing, model-based, and post-

[1] SUNY Binghamton, USA, email: {yzhan202, artir}@binghamton.edu

processing methods. Pre-processing methods work by directly mitigating the bias in the training data itself. Examples of this approach include optimized preprocessing [6], which modifies training data features and labels, reweighting [14], which modifies the weights of different training examples, disparate impact remover [12], which edits feature values to improve group fairness, and learning fair representations [24], which learns fair representations by obfuscating information about protected attributes.

Model-based methods are used to mitigate bias in classifiers; for example, adversarial debiasing [25] uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions. Prejudice remover [16] adds a discrimination-aware regularization term to the learning objective. Meta Fair Classifier [7] takes the fairness measure as part of the input and returns a classifier optimized for that metric. Our approach falls in this category. Existing approaches only learn the parameter values or apply regularization to lessen the effect of sensitive attributes. The fairness measures are not used to directly induce the structure, hence leaving behind some possibility of bias. Our approach differs from existing approaches in that it directly learns the graphical model structure by optimizing for the fairness measures. Thus, our approach is capable of mitigating structural bias in the model, which helps in creating an overall fairer model.

The third class of algorithms focus on post-processing methods to mitigate bias in predictions. For example, reject option classification [15] changes predictions from a classifier to make them fairer. Equalized odds post-processing [13] modifies the predicted labels using an optimization scheme to make predictions fairer. Calibrated equalized odds post-processing [22] optimizes over calibrated classifier score outputs that lead to fair output labels.

3 BACKGROUND FOR FAIR-A3SL

Before delving into the details of Fair-A3SL, we provide necessary background on hinge-loss Markov random fields (HL-MRFs) [2], the probabilistic programming templating language for encoding them, Probabilistic Soft Logic (PSL) [2], and a recently developed structure learning algorithm for learning interpretable relational structures in HL-MRFs, asynchronous advantage actor-critic for structure learning (A3SL) [26].

3.1 Hinge-loss Markov Random Fields

HL-MRFs are a recently developed scalable class of continuous, conditional graphical models [2]. HL-MRFs can be specified using Probabilistic Soft Logic (PSL) [2], a first-order logic templating language. In PSL, random variables are represented as logical atoms and weighted rules define dependencies between them of the form: $\lambda : P(a) \wedge Q(a,b) \rightarrow R(b)$, where *P*, *Q*, and *R* are predicates, a and b are variables, and λ is the weight associated with the rule. The weight of the rule r indicates its importance in the HL-MRF model, which is defined as

$$P(Y|X) \propto \exp\Big(-\sum_{r=1}^{M} \lambda_r \phi_r(Y,X)\Big)$$
$$\phi_r(Y,X) = (\max\{l_r(Y,X), 0\})^{\rho_r} \tag{1}$$

where $P(Y|X)$ is the probability density function of a subset of logical atoms *Y* given observed logical atoms *X*, $\phi_r(Y,X)$ is a hinge-loss potential corresponding to an instantiation of a rule r, and is specified by a linear function l_r and optional exponent $\rho_r \in \{1,2\}$. HL-MRFs admit tractable MAP inference regardless of the graph structure of the graphical model, making it feasible to reason over complex user-specified dependencies. This is possible because HL-MRFs operate on continuous random variables and encode dependencies using potential functions that are convex, so MAP inference in these models is always a convex optimization problem. Farnadi et al. [10] extend the MAP inference algorithm to be able to maximize the a-posteriori values of unknown variables subject to fairness guarantees.

Our approach to learning fair structures focuses on learning logical constructs that particularly bring out the modeling capabilities in HL-MRFs. Below, we provide examples from two datasets we use in our experiments, a relational paper review dataset and a correctional center recidivism prediction dataset:

1. Relational Dependencies and Collective Rules: *highQuality(P)* $\wedge$ *positiveReviews(*R_1*,P)* $\rightarrow$ *positiveReviews(*R_2*,P)*, which captures if paper *P* is of high quality and reviewer R_1 gives the paper a positive review, then reviewer R_2 also gives the paper a positive review. Note that *positiveReviews* is a target predicate and this rule collectively predicts it for both the reviewers.
2. Feature Dependencies: *priorFelony(U, I)* $\wedge$ *africanAmerican(U)* $\rightarrow$ *recidivism(U)*, which captures (unfairly) that if user *U* has committed a prior felony *I* and the race of the user is African American, the user has a higher chance of recidivism. These two features come together to predict recidivism.

3.2 Asynchronous advantage actor-critic structure learning (A3SL) for HL-MRFs

Asynchronous advantage actor-critic structure learning algorithm (A3SL) [26], a recently developed structure learning algorithm for HL-MRFs, adapts a neural policy gradient algorithm asynchronous advantage actor-critic (A3C) [20] for the structure learning problem. A3SL learns interpretable and expressive structures for HL-MRFs by finding the clause set C and corresponding weight vector Λ that maximizes the objective: $J_{\text{A3SL}} = L(Y,X) +$ *Interpretability Priors*, where $L(Y,X)$ is the HL-MRF probability density, $logP(Y|X)$, given by Equation 1. *Interpretability Priors* consist of a combination of priors on the total number of clauses, the maximum possible length of a clause, and domain-specific semantic constraints. The inclusion of semantic constraints and a performance-based utility function allows the algorithm to learn structures that are interpretable and data-driven, thus optimizing for both while being able to rectify any domain-specific intuitions that are not true in the data. The objective function J_{A3SL} is defined as,

$$J_{\text{A3SL}} = \Big(L(Y,X) - \alpha_{\text{len}} * \frac{1}{|C|}\sum_{c\in C}\text{length}(c)$$
$$- \alpha_{\text{num}} * |C| - \alpha_{\text{sem}} * \sum_{c\in C}(\text{Dist}(c) * \lambda_c)\Big) \tag{2}$$

where α_{len}, α_{num}, and α_{sem} parameters denote the strength of the different constraints, λ_c denotes the weight for PSL clause c, and Dist(c) denotes the deviation of clause c from semantic constraints (discussed more in Section 4.4). We refer the reader to [26] for additional details.

4 FAIR-A3SL: FAIRNESS-AWARE STRUCTURE LEARNING FOR HL-MRFS

In this section, we develop Fair-A3SL by incorporating the different fairness measures in the A3SL problem formulation and objective. We first introduce the Fair-A3SL algorithm and then describe all the fairness-related components in the algorithm in detail in the following sections.

4.1 Fair-A3SL algorithm

Algorithm 1 gives the Fair-A3SL algorithm. The algorithm follows an actor-critic reinforcement learning setup to learn the clause list C at each step. Our environment consists of predicates for features (denoted by X), target variables (Y), and data corresponding to X and ground truth data for Y. And each intermediate state s_t at time t comprises of either a partially constructed or a complete set of first order logic clauses, denoted by C. Our action space is defined by all the predicates X, Y, and their negative counterparts, and a special token END. At time t, action a_t adds a new predicate to the current clause or chooses to return the clause by adding an END.

Algorithm 1 Fair-A3SL algorithm

```
Input: A collection of predicates, $X = \{x_j; j = 1, ..., m\}$, $Y = \{y_j; j = 1, ..., n\}$, , Ground truth labels $Y_g$ for $Y$
Let $C = \{c_0, c_1, .., c_M\}$ denote set of first-order logic clauses, and corresponding weights $\Lambda$
Let $C_{list}$ denote list of $C$ obtained with reward $> 0$.
Output: Optimal $C$ denoted by $C^*$
1: function $C^*$ = Fair-A3SL($Y, X$)
2:     for each thread asynchronously do
3:         Construct clause list $C$ under A3SL agent policy
4:         Initialize weights $\Lambda$ for $C$
5:         Perform weight learning and update $\Lambda$.
6:         Perform fairness-aware inference and get $\hat{Y}$ /* MAP inference with fairness constraints */
7:         Obtain reward Utility($Y$, $\hat{Y}$) = $\log P(Y, X)$ - $\alpha *$ fairness priors
8:         Add $C$ to $C_{list}$
9:         Accumulate gradients and update policy and value function parameters according to new state C
10:    $C^*$ = optimal $C$ from $C_{list}$
11:    return $C^*$
```

In the Fair-A3SL algorithm, we present two main ways of encoding the fairness measures: i) as MAP inference constraints, and ii) as priors in the objective function. The fairness measures encoded as constraints are integrated as linear inequality constraints in the MAP inference for HL-MRFs; we present more details in Section 4.2. Step 6 in Algorithm 1 captures this step, where fairness-aware inference subject to the fairness MAP inference constraints is performed.

To include fairness measures as priors, we turn to the reward/utility function in Step 7 of Algorithm 1. The immediate reward r_t is equal to the value of objective function at step t if the clause set construction is complete; r_t equals 0 otherwise. The cumulative reward $R_t = \sum_{k=0}^{\infty} \gamma r_{t+k}$ is equal to the value of the objective function, where γ is the discount factor, and we set it to 1 in all our experiments. The fairness measures encoded as priors are integrated in the reward utility function, the new utility after incorporating the priors becomes *Utility*$(Y, \hat{Y}) = \log P(Y, X) - \alpha *$ *fairness priors*, where $P(Y|X)$ is the HL-MRF objective given by Equation 1 and α denotes the strength of the fairness prior(s). The algorithm returns the clause list with the best accumulated reward calculated using the utility function as the optimal clause list C^*.

4.2 Fairness aeasures as MAP inference constraints

Here, we discuss how to integrate different fairness measures as MAP inference constraints. First, we start with the assumption that we are given a dataset consisting of n samples $\{(A_i, X_i, Y_i)\}_{i=1}^{n}$. Here, A denotes one or more sensitive attributes such as gender and race, X denotes other non-sensitive features, and Y denotes the ground-truth labels. We group instances or users based on their sensitive attributes into two groups, protected and unprotected. We then define, $a = \sum_{x \in \text{protected group}} \neg\hat{Y}(x)$, $c = \sum_{x \in \text{unprotected group}} \neg\hat{Y}(x)$, $g_1 = |\text{protected group}|$, $g_2 = |\text{unprotected group}|$. $\hat{Y}$ refers to a positive prediction (e.g., acceptance) and $\neg\hat{Y}$ refers to a negative prediction (e.g., denial) from the trained model. The proportions of denial for protected and unprotected groups are $p_1 = \frac{a}{g_1}$ and $p_2 = \frac{c}{g_2}$, respectively, where g_1 and g_2 are constants [10, 21].

Following Farnadi et al.'s the definition of δ-fairness, the fairness measures can be defined in terms of p_1 and p_2 as follows, where $0 \leq \delta \leq 1$,

$$\text{Risk difference: RD} = p_1 - p_2; \qquad -\delta \leq \text{RD} \leq \delta$$

$$\text{Risk Ratio: RR} = \frac{p_1}{p_2}; \qquad 1 - \delta \leq \text{RR} \leq 1 + \delta$$

$$\text{Relative Chance: RC} = \frac{1 - p_1}{1 - p_2}; \qquad 1 - \delta \leq \text{RC} \leq 1 + \delta$$

The δ-fairness constraints above translate to *six* linear inequality constraints in the HL-MRF framework. For example, the linear inequality constraints $l_1(Y, X)$ and $l_2(Y, X)$ defined for satisfying the inequality $-\delta \leq \text{RD} \leq \delta$ have the forms shown below, where x_1,..., x_{g_1} are instances in the protected group, and $x_{g_1+1}, ..., x_{g_1+g_2}$ are instances in the unprotected group, and the total number of instances $n = g_1 + g_2$.

$$l_1 \Rightarrow \text{RD} \leq \delta \Rightarrow \left(g_2 ... g_2, -g_1, ..., -g_1\right) * \begin{pmatrix} \hat{Y}(x_1) \\ \hat{Y}(x_2) \\ \vdots \\ \hat{Y}(x_n) \end{pmatrix} \geq -g_1 g_2 \delta$$

$$l_2 \Rightarrow \text{RD} \geq -\delta \Rightarrow \left(g_2, ..., g_2, -g_1, ..., -g_1\right) * \begin{pmatrix} \hat{Y}(x_1) \\ \hat{Y}(x_2) \\ \vdots \\ \hat{Y}(x_n) \end{pmatrix} \leq g_1 g_2 \delta$$

Next, we consider a fairness metric for collaborative filtering [23]: non-parity unfairness. Non-parity unfairness is defined as the absolute difference between the overall predicted average ratings of protected users and those of unprotected users:

$$U_{par} = |E_{\text{protected}}[\hat{Y}] - E_{\text{unprotected}}[\hat{Y}]|$$

$$E_{\text{protected}}[\hat{Y}] = \frac{1}{g_1} \sum_{\{(i,j)|i \in \text{protected group}\}} \hat{Y}_{i,j}$$

$$E_{\text{unprotected}}[\hat{Y}] = \frac{1}{g_2} \sum_{\{(i,j)|i \in \text{unprotected group}\}} \hat{Y}_{i,j}$$

where $\hat{Y}$ is the prediction, g_1 is the total rating by protected users and g_2 the total rating by unprotected users. Below, we demonstrate how to capture non-parity unfairness in Fair-A3SL as a MAP inference constraint. We get the corresponding δ-fairness linear inequality constraints l_3 and l_4 below, where n represents number of users u, m represents number of items v.

$$l_3 \Rightarrow U_{par} \geq -\delta$$

$$\Rightarrow \left(g_2 ... g_2, -g_1, ..., -g_1\right) * \begin{pmatrix} \hat{Y}(u_1, v_1) \\ \hat{Y}(u_1, v_2) \\ \vdots \\ \hat{Y}(u_n, v_m) \end{pmatrix} \geq -g_1 g_2 \delta$$

$$l_4 \Rightarrow U_{par} \leq \delta$$

$$\Rightarrow \big(g_2 ... g_2, -g_1, ..., -g_1\big) * \begin{pmatrix} \hat{Y}(u_1, v_1) \\ \hat{Y}(u_1, v_2) \\ \vdots \\ \hat{Y}(u_n, v_m) \end{pmatrix} \leq g_1 g_2 \delta$$

The linear form of the constraints is consistent with MAP inference in HL-MRF model; they can be seamlessly solved using a consensus-optimization algorithm based on the alternating direction method of multipliers (ADMM) [5]. To accomplish this, we extend the consensus optimization algorithm by Bach et al. [2] for MAP inference in HL-MRFs to include above defined fairness linear inequality constraints.

Similarly, other fairness measures can also be incorporated in the Fair-A3SL framework as constraints. *Statistical Parity Difference* measures the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group [16] and *Disparate Impact* measures the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group [12]. Both these measures are similar to the relative chance (RC) relational measure and can be encoded similarly.

4.3 Fairness measures as objective priors

While certain fairness measures can be modeled as MAP inference constraints in the framework, the post-processing fairness measures can only be modeled as priors in our objective due to the absence of ground truth for target Y at test time as discussed below.

Equalized Odds Difference [13] measures the difference of false positive rate and true positive rate between unprivileged and privileged groups, which can be defined as $\sum_{y \in \{0,1\}} |Pr(\hat{Y} = 1 | A = 0, Y = y) - Pr(\hat{Y} = 1 | A = 1, Y = y)|$, where $\hat{Y}$ is the predicted value and Y is the ground truth. We cannot directly incorporate this measure as a MAP inference constraint since at test time the true value of Y is not available. This measure and other similar post-processing measures that rely on true ground-truth labels can be encoded as priors in the Fair-A3SL algorithm. We integrate the priors in the objective function, which then is used in computing the agent's rewards in the Fair-A3SL algorithm as discussed in Section 4.1.

Overestimation unfairness measures inconsistency in how much the predictions overestimate the true ratings [23]. This fairness measure is used in the collaborative filtering setting. Following equations give the formula for U_{over} and the expectation for the protected group $E_{\text{protected}}$. The average for $E_{\text{unprotected}}$ is computed analogously.

$$U_{\text{over}} = \frac{1}{m} \sum_{j=1}^{m} |\max(0, E_{\text{protected}}[\hat{Y}]_j - E_{\text{protected}}[Y]_j) - \max(0, E_{\text{unprotected}}[\hat{Y}]_j - E_{\text{unprotected}}[Y]_j)|$$

$$E_{\text{protected}}[\hat{Y}]_j = \frac{1}{|\{(i,j) | i \in \text{protected}\}|} \sum_{i \in \text{protected}} \hat{Y}_{i,j}$$

Equal Opportunity Difference measures the difference of true positive rates between the unprivileged and the privileged groups [13]. Average Odds Difference [1] measures the average difference of false positive rate and true positive rate between unprivileged and privileged groups. These measures are comparable to the Equalized Odds Difference measure and can be similarly encoded as priors.

4.4 Domain-specific semantic constraints

An interpretable model lays the foundation for fairness and transparency. In addition to inducing fairness-aware relational structures, we also include semantically meaningful domain constraints that do not contain any structural bias and encourage the algorithm to learn interpretable structures. This is helpful in making the resulting model more appealing to end users. Here, we show how to group predicates and their negative counterparts into two categories, *positive signals* and *negative signals* using the semantic interpretation of the predicate. If the user is unsure about the semantics of any predicate, they can be incorporated in both the categories to avoid any unintentional bias.

Table 1: Right reasons identified from domain semantics

positive signals $\Rightarrow$	any positive signal not already included
negative signals $\Rightarrow$	negative signal not already included
positive signals $\wedge \neg$negative signal $\Rightarrow$	positive signal not already included
negative signals $\wedge \neg$ positive signal $\Rightarrow$	negative signal not already included

We illustrate this using the COMPAS dataset, one of the datasets widely used in fairness studies and also in our experiments. We capture *positive signals P*={*priorFelonHistory, priorMisdemeanorHistory, priorOtherHistory, juvFelonHistory, juvMisdemeanorHistory, juvOtherHistory, priors, felony, recidivism, ¬oldAge, longJailDay, ¬longJailDay*} that capture tendency toward recidivism and *negative signals N*={*¬felony, ¬recidivism, oldAge, longJailDay, ¬longJailDay*} that capture tendency against recidivism. Since at first we are not sure about the effect of *longJailDay* and its negative counterpart on *recidivism* prediction from domain knowledge, we place it in both categories. The domain-specific semantic constraints have the general structure in Table 1, where positive signals $\subseteq P$, negative signals $\subseteq N$, and any positive signal $\in P$, negative signal $\in N$. We use a distance function, *Dist(c)* to capture if the learned clause structure complies with or deviates from the right reasons identified by the expert: *Dist(c) = 0*, if the clause complies with the right reasons and *Dist(c) = 1*, otherwise. This distance function is then integrated in the objective functions discussed in Section 4.5. If the domain-specific guidance is not readily available for the specific domain, the model is able to work without them as well as they are added only to enhance interpretability when appropriate.

4.5 Fair-A3SL objective functions

We present two different objective functions that we use across our three predictive modeling scenarios that demonstrates how a combination of fairness constraints, fairness priors, and semantic constraints can be represented in an objective function. This objective can be easily modified to include/exclude specific fairness/semantic constraints or fairness priors.

4.5.1 *Fair-A3SL objective for relational models*

In the first objective, we use a combination of fairness measures both encoded as constraints and as priors. Here, we encode the relational fairness measures RR, RC, and RD as MAP inference constraints and the equalized odds difference measure as a prior in the objective along with interpretability priors for the specific domain in question. Equation 3 gives the Fair-A3SL objective function corresponding to this combination. We use this objective function in our experiments in Section 5.1 on the relational dataset and in Section 5.2 on the recidivism prediction dataset.

$$\begin{aligned} J_{\text{Fair-A3SL}} = &\log P(Y, X) + \textit{Interpretability Priors} \\ &+ \alpha_{\text{odds}} * U_{\text{odds}} \\ \text{s.t.} &- \delta \leq RD \leq \delta \\ &1 - \delta \leq RR \leq 1 + \delta \\ &1 - \delta \leq RC \leq 1 + \delta \end{aligned} \tag{3}$$

where U_{odds} refers to the equalized odds difference fairness measure and α_{odds} captures its degree of enforcement.

4.5.2 Fair-A3SL objective for recommender systems

For the recommender systems problem, we turn to the corresponding fairness measures of overestimation and non-parity. Equation 4 gives the Fair-A3SL objective for recommender systems. As is evident from the equation, here again we include a combination of constraints and priors in the objective; we incorporate the non-parity fairness measure as a MAP inference constraint (U_{par}) and overestimation as an objective prior (U_{over}). We use this objective for the experimental results in Section 5.3.

$$\begin{aligned} J_{\text{Fair-A3SL}} &= \log P(Y,X) + \alpha_{\text{over}} * U_{\text{over}} \\ s.t. & -\delta \leq U_{\text{par}} \leq \delta \end{aligned} \tag{4}$$

4.6 Highlights of Fair-A3SL

Our approach to fairness is versatile in its ability to encode many different fairness measures toward directly learning the graphical model structure. Fair-A3SL provides the capability of encoding fairness measures as constraints and/or as priors and has minimal pre-processing requirements (only those imposed by the underlying fairness measures). While many existing work indicate the importance of combining fairness measures for practitioners, they also note that there is often a trade-off between various fairness measures and it is challenging to construct a single fairness objective that performs well across different measures [23, 11]. While this remains true for conflicting measures, Fair-A3SL is a step in the right direction, where we present a platform that can incorporate a combination of fairness metrics while simultaneously optimizing for them. In Equations 3 and 4, we show some possible combinations and our results indicate Fair-A3SL can indeed optimize for multiple fairness metrics at the same time. These desirable qualities in Fair-A3SL can potentially help downstream users such as policy makers and decision making organizations (e.g., bank loans, student admissions) to successfully adopt the framework.

5 EXPERIMENTAL EVALUATION

We conduct experiments to evaluate the learned structures quantitatively and qualitatively on three fairness datasets. In our experiments, we illustrate the capability of Fair-A3SL to be able to: i) learn fair network and collective structures that bring out the modeling power of statistical relational models, ii) incorporate a wide range of fairness measures and learn model structures using them, and iii) learn model structures that outperform state-of-the-art fairness models both across performance and fairness metrics and are qualitatively meaningful. The Fair-A3SL code and the code for experiments will be made publicly available when the paper is accepted for publication. The best scores and those that are statistically indistinguishable from the best are typed in **bold** in all the results. All experiments use 5-fold cross-validation.

5.1 Results on relational paper review dataset

We first present results on a paper reviewing problem that can potentially be biased by the author's affiliation instead of the quality of the paper. We follow Farnadi et al. [10] to generate a similar dataset to theirs in order to facilitate a direct comparison. Table 2 gives the conditional probability distribution table (*left*) and the Bayesian network (*right*) that we use for generating the data. Two specific scenarios parametrized by P(H) that determine the degree of discrimination are: i) probability of the paper receiving a favorable rating given the paper is of high quality and the author is not from a top ranked institution ($\theta_1 = P(R_1|Q = T, H = F, S = T)$), and ii) probability of the paper receiving a favorable reviewer rating given the paper is of high quality and the author is from a top ranked institution ($\theta_2 = P(R_1|Q = T, H = T, S = T)$). We introduce bias in the data when the author is a student (S = T) by setting $\theta_1 = 0.5$ and $\theta_2 = 0.9$. We set $P(R_1|Q = T, H = F, S = F)$ and $P(R_1|Q = T, H = T, S = F)$ to 0.85. The train and test dataset both contain data generated using the Bayesian network comprising of 100 papers, 100 authors, 30 reviewers, and each paper is reviewed by 2 random reviewers.

Table 2: Generation model of the paper review dataset: *left* shows the joint probability distribution of variables and *right* shows the graphical model. Q: indicates whether or not the paper is high quality; H: indicates whether or not the author is affiliated with a top-rank institute; S: indicates whether or not the author is a student; R1, R2: indicates whether or not the first/second reviewer gives the paper a positive review.

Q	H	S	P(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 = 0.50
T	T	F	0.85
T	T	T	θ_2 = 0.90

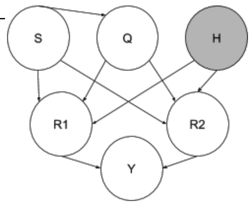

Table 3: Fairness-A3SL Model on Paper-Review Dataset

PSL Rules Learnt from Fair-A3SL
Author: A; Reviewer : R_1, R_2; Paper : P
Set A. Relational Rules:
λ_1: *submits(A,P)* $\wedge$ *student(A)* $\wedge$ *positiveReviews(*R_1*,P)* $\rightarrow$ $\neg$*positiveSummary(P)*
λ_2: *acceptable(P)* $\wedge$ *positiveReviews(*R_1*,P)* $\rightarrow$ *positiveSummary(P)*
λ_3: $\neg$*highQuality(P)* $\rightarrow$ $\neg$*positiveReviews(*R_1*,P)*
λ_4: *highQuality(P)* $\rightarrow$ *positiveReviews(*R_1*,P)*
Set B. Collective Rules:
λ_5: *highQuality(P)* $\wedge$ *positiveReviews(*R_1*,P)* $\wedge$ *reviews(*R_2*,P)* $\rightarrow$ *positiveReviews(*R_2*,P)*
λ_6: *positiveSummary(P)* $\wedge$ *positiveReview(*R_1*,P)* $\wedge$ *reviews(*R_2*,P)* $\rightarrow$ *positiveReviews(*R_2*, P)*

Table 3 gives the learnt rules the Fair-A3SL model on the paper review dataset. To enable a comparison with Farnadi et al. [10], we also enhance A3SL by adding the ability to encode collective rules. Collective rules jointly predict two or more target variables. Note that the learned model structure is expressive, learning different kinds of rules: network, collective, and combination of features.

We compare Fair-A3SL with the following state-of-the-art baselines: i) Fair-PSL [10], manually-defined PSL rules with fairness constraints in inference, ii) Sensitive-PSL, manually-defined PSL rules with *no* fairness constraints, and iii) Sensitive-A3SL [26], a model structure learned using A3SL with *no* fairness constraints or priors. Additionally, we experiment with three versions of Fair-A3SL that use different combinations of fairness measures. Fair-A3SL$_1$ in-

Table 4: Comparison of Fair-A3SL with baselines on Area under PR curve and ROC curve

Model	AUC-PR Pro.	AUC-PR Unpro.	AUC-ROC
Sensitive-PSL	0.3490±0.1946	0.6112±0.0566	0.8354±0.0421
Sensitive-A3SL	0.3490±0.1946	0.6707±0.0443	**0.8544±0.0360**
Fair-PSL ([10])	0.4332±0.1104	0.6009±0.0463	0.7887±0.0306
Fair-A3SL$_1$	0.3490±0.1946	0.6112±0.0566	0.8037±0.0491
Fair-A3SL$_2$	**0.6208±0.2981**	**0.7279±0.0322**	0.8118±0.0076
Fair-A3SL$_3$	**0.6208±0.2981**	0.5469±0.1486	0.7396±0.0005

Table 5: Comparison of Fair-A3SL with baselines on fairness measures

Model	RD	RR	RC	Equal Odds Pos.	Equal Odds Neg.
Sensitive-PSL	0.2766±0.0768	0.2090±0.0935	1.4344±0.1543	0.5227±0.1799	0.1429±0.0389
Sensitive-A3SL	0.0661±0.0591	0.8723±0.1479	1.1094±0.1202	0.1550±0.1156	0.0317±0.0056
Fair-PSL [10]	**0.0005±0.0004**	**0.9980±0.0019**	**1.0007±0.0007**	0.1346±0.0234	0.0829±0.0550
Fair-A3SL$_1$	**0.0002±3.7e-5**	**1.0007±0.0002**	**0.9996±4.9e-5**	0.1968±0.1899	0.1286±0.0548
Fair-A3SL$_2$	**0.0059±0.0005**	**1.0009±0.0115**	**0.9989±0.0122**	**0.0096±0.0066**	**0.0093±0.0054**
Fair-A3SL$_3$	**7.9e-5±1.8e-5**	**0.9999±4.9e-5**	**1.1651±0.3413**	**0.0001±1.2e-5**	**6.6e-5±3.2e-5**

cludes fairness constraints without equalized odds priors. Fair-A3SL$_2$ includes fairness constraints along with equalized odds priors with $\alpha_{odds} = 0.1$. Fair-A3SL$_3$ includes fairness constraints along with equalized odds with $\alpha_{\text{odds}} = 0.5$. We set δ-fairness=0.1 for all fairness inference inequality constraints. The AUC-ROC values from the Sensitive-A3SL model can be considered an upper bound, as it is a purely data-driven model.

Our specific focus is on the prediction performance for protected/unprotected groups, especially for predicting a positive outcome in both these groups (Table 4). We report area under the AUC-PR curve for the positive class (*positiveSummary*). From the table, we can see that all A3SL versions outperform the human expert counterparts (Sensitive-A3SL vs. Sensitive-PSL, Fair-A3SL versions vs. Fair-PSL). We can see that the Fair-PSL model even when the fairness measures are included in the inference only achieves a prediction performance of ~ 0.4, while the Fair-A3SL models achieve > 0.6 for the protected group. The Fair-A3SL models also improve the prediction performance of the unprotected groups when compared to the Fair-PSL model. The combined AUC-ROC value for the Fair-A3SL models is also closer to the models that include sensitive attributes (Sensitive-PSL and Sensitive-A3SL). Similarly, all the Fair-A3SL models achieve better or comparable performance across all fairness metrics (RD, RR, RC, Equalized Odds Positive and Negative) when compared with Fair-PSL with manually defined rules (Table 5). Particularly, for the equalized odds measures, Fair-A3SL models clearly outperform Fair-PSL. We also observe that we get better results for the equalized odds fairness measure when we increase the value of α_{odds}. Thus, Fair-A3SL is able to achieve fairness without compromising on performance.

5.2 Results on COMPAS dataset

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool produces a risk score that predicts a person's likelihood of committing a crime in the next two years [19]. The output is a score between 1 to 10 that maps to low, medium, or high. We collapse this to a binary prediction: a score of 0 corresponds to a prediction of low risk according to COMPAS, while a score of 1 indicates high or medium risk. The dataset also contains information on recidivism for each person over the next two years, which we use as ground truth. Existing work shows that the COMPAS risk scores discriminate against black defendants, who were predicted to be far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk [19, 9].

Table 6 gives the Sensitive-A3SL model. We can see that the model combines other recidivism signals of having committed prior felonies (*priors* and *priorFelony)* with the race attribute (*africanAmerican*), indicating how the race attribute and combinations with it are predictive of recidivism and are a natural albeit unfair and discriminatory choice for models that are solely performance driven. The rules learned by the Fair-A3SL model are given in Table 7. Parameter U represents user, I_i represents a felony instance. For example, *priorFelonHistory(U,I_1)* can be grounded with multiple historical felony instances I_1 for each user U. Fair-A3SL's transparency, interpretability, expressibility, along with fairness, makes it an ideal candidate for automatically learning prediction models for sensitive domains.

Table 6: Representative rules from Sensitive-A3SL model

Sensitive-A3SL Recidivism Model
U : users; I_i : felony instances.

priors(U, I_4) $\wedge$ *africanAmerican(U)* $\rightarrow$ *recidivism(U)*
priorFelony(U, I_5) $\wedge$ *africanAmerican(U)* $\rightarrow$ *recidivism(U)*
$\neg$*oldAge(U)* $\wedge$ *africanAmerican(U)* $\rightarrow$ *recidivism(U)*
africanAmerican(U) $\rightarrow$ *recidivism(U)*

Table 7: Rules from Fair-A3SL model

Fairness-A3SL Recidivism Model
U : users; I_i : felony instances.
Set A. Combining Local Features: λ_1: *oldAge(U)* $\rightarrow \neg$ *recidivism(U)* λ_2: $\neg$*oldAge(U)* $\wedge$ *longJailDay(U)* $\rightarrow$ *recidivism(U)* λ_3: $\neg$*longJailDay(U)* $\rightarrow$ *recidivism(U)* **Set B. Combining Jail History Features**: λ_4: *priorFelonHistory(U, I_1)* $\rightarrow$ *recidivism(U)* λ_5: *priorMisdemeanorHistory(U, I_2)* $\rightarrow$ *recidivism(U)* λ_6: *juvenileOtherHistory(U, I_3)* $\rightarrow$ *recidivism(U)* λ_7: *priors(U, I_4)* $\rightarrow$ *recidivism(U)* **Set C. Prior Rule**: λ_8: *user(U)* $\rightarrow \neg$ *recidivism(U)*

We compare Fair-A3SL with recently developed state-of-the-art fairness models: i) Calibrated Equalized Odds [22], ii) Prejudice Remover [16], iii) Optimized Pre-processing [6], iv) Adversarial Debiasing [25], and v) Line-FERM [8], where Calibrated Equalized Odds, Prejudice Remover, and Optimized Preprocessing use logistic regression as the backend model; Adversarial Debiasing uses a deep learning neural network model; and FERM uses SVM as the underlying model. Table 8 gives the 5-fold cross-validation results and shows

Table 8: AUC-PR curve and ROC values for state-of-the-art fairness models and Fair-A3SL for COMPAS dataset.

Model	AUC-PR Pro.	AUC-PR Unpro.	AUC-ROC
COMPAS Scores [19]	0.6168±0.0177	0.5118±0.0173	0.6530±0.0158
Line-FERM [8]	0.6281±0.0077	0.5103±0.0167	0.6482±0.0145
Calibrated Equalized Odds [22]	0.7030±0.0371	0.3882±0.0096	0.6553±0.0207
Prejudice Remover [16]	0.6801±0.0274	0.5494±0.0259	0.6859±0.0040
Optimized Preprocessing [6]	0.7039±0.0191	0.5903±0.0399	**0.7131±0.0123**
Adversarial Debiasing [25]	0.6654±0.0334	0.5174±0.0356	0.6545±0.0241
Fair-A3SL (our approach)	**0.7262±0.0165**	**0.6080±0.0229**	**0.7103±0.0109**

Table 9: Comparison of performance of Fair-A3SL with state-of-the-art fairness models on different fairness metrics for COMPAS dataset

Model	RD	RR	RC	Equal Odds Pos.	Equal Odds Neg.
COMPAS Scores [19]	0.2632±0.0228	1.8170±0.1165	0.6106±0.0232	0.2261±0.0284	0.2285±0.0187
Line-FERM [8]	0.1450±0.0647	1.5485±0.2704	0.7936±0.0961	0.1147±0.0774	0.1063±0.0609
Calibrated Equalized Odds [22]	0.1350±0.0145	1.3480±0.0515	0.7986±0.0322	0.1946±0.0180	0.0698±0.0160
Prejudice Remover [16]	0.0541±0.0089	1.1438±0.0306	0.9125±0.0124	0.0772±0.0194	0.0583±0.0118
Optimized Pre-processing [6]	0.0517±0.0102	1.1218±0.0259	0.9099±0.0168	0.0325±0.0178	0.0361±0.0088
Adversarial Debiasing [25]	0.0511±0.0096	1.1176±0.0243	0.9094±0.0161	0.0539±0.0089	0.0307±0.0089
Fair-A3SL (our approach)	**0.0035±0.0003**	**1.0047±0.0006**	**0.9851±0.0030**	**0.0039±0.0051**	**0.0160±0.0105**

that Fair-A3SL is able to achieve a better prediction performance for both the protected and unprotected groups, individually (AUC-PR for protected and unprotected groups) and combined (AUC-ROC). We use the IBM AI Fairness 360 tool [1] for running the existing state-of-the-art models. We also demonstrate that our learned model outperforms the state-of-the-art fairness models in the fairness metrics as well, achieving the best scores across all metrics (Table 9).

5.3 Results on Movielens dataset

In the third experiment, we consider another important domain for fairness, recommender systems. To evaluate the effectiveness of Fair-A3SL in recommender systems, we use the *MovieLens* 100k dataset. It consists of ratings from 1 to 5 by 943 users for 1682 movies. The users are annotated with demographic variables such as gender, and the movies are each annotated with a set of genres. For convenience, we convert the ratings to range between values 0 and 1. From Table 10, we can see that women rate musical and romance films higher and more frequently than men. Men rate Sci-Fi and crime films higher and more frequently than women. Women and men both give action films an almost equal rating, but men rate these films more frequently.

Table 10: Gender-based statistics of movie genres in MovieLens data.

	Romance	Action	Sci-Fi	Musical	Crime
Count	14202	19141	9577	3765	5835
Avg Count per Female	24.74	23.13	11.57	7.32	7.41
Avg Count per Male	20.43	31.11	15.64	6.30	9.67
Avg Rating by Female	0.73	0.70	0.70	0.73	0.71
Avg Rating by Male	0.72	0.70	0.71	0.69	0.73

Table 11: Rules from Fair-A3SL model for MovieLens data

Fair-A3SL Recommender Model

U_i : users; I_i : items.

λ_1: $\textit{rating}_{MF}(U, I) \rightarrow \textit{rating(U,I)}$

λ_2: $\textit{avgUserRating(U)} \wedge \textit{reviews(U,I)} \rightarrow \textit{rating(U,I)}$

λ_3: $\textit{itemPearsonSim(I, I}_2\textit{)} \wedge \textit{rating}_{MF}\textit{(U,I)} \wedge \textit{rating(U,I)}$
$\rightarrow \textit{rating(U,I}_2\textit{)}$

λ_4: $\textit{userPearsonSim(U, U}_2\textit{)} \wedge \textit{rating(U}_2\textit{,I)} \wedge \textit{avgUserRating(U)}$
$\rightarrow \textit{avgItemRating(I)}$

Table 12: Mean square errors (MSE) results on state-of-the-art fairness models and Fair-A3SL on MovieLens dataset

Model	Err Pro.	Err Unpro.	Error
HyPER [18]	0.04530±0.00212	0.03887±7.4e-5	0.04043±0.00046
MF [17]	0.03909±0.00233	0.03258±0.00014	0.03415±0.00067
Fair-HyPER[11]	0.03947±0.00222	0.03297±0.00015	0.03455±0.00065
Fair-MF [23]	0.03942±0.00215	0.03249±0.00015	0.03415±0.00063
Fair-A3SL	**0.03779±0.00203**	**0.03189±0.00025**	**0.03331±0.00068**

Table 13: Overall fairness measurements of state-of-the-art fairness models and Fair-A3SL on MovieLens dataset

Model	Non-Parity	Overestimation
HyPER [18]	0.00424±0.00033	**0.0349±0.00338**
MF [17]	0.00473±0.0005	0.06294±0.00475
Fair-HyPER [11]	0.00465±0.00037	0.05346±0.00380
Fair-MF [23]	0.00076±0.00055	0.06101±0.00402
Fair-A3SL	**9.2e-5±5.9e-5**	0.05914±0.00307

Following Kouki et al. [18], we extract features that combines multiple different sources of information, including similarity between pairs of users (*userPearsonSim*(U, U_2)), similarity between items (*itemPearsonSim*(I, I_2)), average rating with respect to users and items to serve as priors (*avgUserRating*(U) and *avgItemRating*(I)), and leveraging predictions from existing recommendation algorithms as a feature ($\textit{rating}_{MF}(U, I)$) to enable an appropriate comparison. Table 11 gives the rules learned by Fair-A3SL.

We compare our approach to the state-of-the-art recommender systems baseline models: i) HyPER [18], which is a PSL model and includes hybrid recommender systems feature,; ii) matrix factorization based collaborative filtering model [17], iii) Fair-HyPER [11], which defines additional latent variable rules to abstract the rating of unprotected and protected groups in order to ensure there is no overestimation unfairness, iv) baseline model Fair-MF [23], which considers overestimation and non-parity unfairness as regularization terms. Table 12 shows Fair-A3SL achieves the best overall performance for both the protected and unprotected groups. Table 13 shows that our Fair-A3SL model gets a comparable value in the overestimation unfairness measure, and the best value in the non-parity fairness measure. The model learned by Fair-A3SL achieves comparable performance to Fair-HyPER even without the inclusion of carefully designed latent variables that provide additional complexity.

6 CONCLUSION

In this work, we developed Fair-A3SL, a general purpose fair structure learning algorithm for HL-MRFs and demonstrated that it learns fair, semantically interpretable, and expressive relational structures while achieving good prediction performance. Fair-A3SL is capable of encoding various different measures of fairness both as constraints and priors and we demonstrate its effectiveness across three different domains and modeling scenarios. Further, Fair-A3SL has minimal pre-processing requirements (only those posed by the underlying fairness measures) and can seamlessly be utilized to learn models for any sensitive prediction problem including those that require complex relational structures. Fair-A3SL's joint qualities of fairness, interpretability, and performance make it lucrative for many downstream applications (e.g., bank loans, student admissions) to adopt it.

REFERENCES

[1] IBM AI fairness 360 open source toolkit. https://aif360.mybluemix.net/.

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

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

[4] Danah Boyd, Karen Levy, and Alice Marwick, 'The networked nature of algorithmic discrimination', *Data and Discrimination: Collected Essays. Open Technology Institute*, (2014).

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al., 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends® in Machine learning*, 1–122, (2011).

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney, 'Optimized pre-processing for discrimination prevention', in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, (2017).

[7] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi, 'Classification with fairness constraints: A meta-algorithm with provable guarantees', in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, (2019).

[8] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil, 'Empirical risk minimization under fairness constraints', in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, (2018).

[9] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science advances*, **4**(1), eaao5580, (2018).

[10] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor, 'Fairness in relational domains', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, (2018).

[11] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor, 'A fairness-aware hybrid recommender system', in *RecSys Workshop on FATREC*, (2018).

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, (2015).

[13] Moritz Hardt, Eric Price, Nati Srebro, et al., 'Equality of opportunity in supervised learning', in *Proceedings of the Conference on Advances in neural information processing systems (NIPS)*, (2016).

[14] Faisal Kamiran and Toon Calders, 'Data preprocessing techniques for classification without discrimination', *Knowledge and Information Systems (KAIS)*, 1–33, (2012).

[15] Faisal Kamiran, Asim Karim, and Xiangliang Zhang, 'Decision theory for discrimination-aware classification', in *Proceedings of the International Conference on Data Mining (ICDM)*, (2012).

[16] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma, 'Fairness-aware classifier with prejudice remover regularizer', in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (2012).

[17] Yehuda Koren, Robert Bell, and Chris Volinsky, 'Matrix factorization techniques for recommender systems', *Computer*, (8), 30–37, (2009).

[18] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor, 'Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems', in *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, (2015).

[19] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, 'How we analyzed the COMPAS recidivism algorithm', in *ProPublica*, (2016).

[20] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu, 'Asynchronous methods for deep reinforcement learning', in *Proceedings of the International Conference on Machine Learning (ICML)*, (2016).

[21] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini, 'A study of top-k measures for discrimination discovery', in *Proceedings of the Annual ACM Symposium on Applied Computing*, (2012).

[22] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger, 'On fairness and calibration', in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, (2017).

[23] Sirui Yao and Bert Huang, 'Beyond parity: Fairness objectives for collaborative filtering', in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, (2017).

[24] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, 'Learning fair representations', in *Proceedings of the International Conference on Machine Learning (ICML)*, (2013).

[25] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, 'Mitigating unwanted biases with adversarial learning', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, (2018).

[26] Yue Zhang and Arti Ramesh, 'Learning interpretable relational structures of hinge-loss markov random fields', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, (2019).