Unobserved Protected Classes & Fairness

Kiki Densel and Tamara Tran

CS 590, Spring 2023
Focus

**Fairness Under Unawareness:**
Assessing Disparity When Protected Class Is Unobserved
Chen et al, 2019

Using proxy variables to perform fairness assessments in cases when protected classes are unobserved

**Interventional Fairness with Indirect Knowledge of Unobserved Protected Attributes**
Galhotra et al, 2021

Identifying proxy variables responsible for biased outcomes when protected classes are unobserved
Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

Chen et al, 2019
1. Probabilistic proxy models and paper objectives
   a. Use in fairness assessments
   b. Overestimation issue
2. Threshold estimators for protected attribute classification
   a. Definition
   b. Decomposing bias
3. Weighted estimators
   a. Definition
   b. Single source of bias
4. Results
   a. Synthetic data
   b. Home Mortgage Disclosure Act dataset
Protected Classes

Ethical and legal obligations to demonstrate a lack of discrimination in:
- Employment
- Compensation
- University admissions
- Credit decisions

Illegal or operationally difficult to obtain this data

Protected class membership unobserved in decision-making models

Table 1: Protected classes defined under fair lending laws.

<table>
<thead>
<tr>
<th>Law</th>
<th>FHA[28]</th>
<th>ECOA[29]</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>color</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>disability</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>exercised rights under CCPA</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>familial status (household composition)</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>gender identity</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>marital status (single or married)</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>national origin</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>race</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>recipient of public assistance</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>religion</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>sex</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Discussion: What are the implications of designing a model when protected attributes are unobserved?
Challenge: If protected classes are unobserved, how can we verify that models are making fair decisions?

**Example:** Credit card and auto loan companies cannot ask applicants their race, but they *also* must demonstrate that the way they extend credit is not racially discriminatory.
Fairness Assessments with Probabilistic Proxy Models

- Proxy variables $\rightarrow$ probabilistic proxy model $\rightarrow$ impute protected class labels
- Perform fairness assessment with imputed data
Probabilistic Proxy Models Overestimate Disparity

- Probabilistic proxy models tend to overestimate disparity
  - Unreliable or controversial fairness assessments
  - Example: Bayesian Improved Surname Geocoding (BISG)
Paper Objectives

1. Improve understanding of demographic disparity in fairness assessments performed with probabilistic proxy models
   a. Derive the statistical bias for the commonly used *thresholded estimator* (method of applying probabilistic proxy models)
   b. Decompose bias into multiple sources → set of interpretable conditions

2. Design a *new weighted* estimator as an additional tool in fairness assessments

↑ Model Fairness
Outline

1. Probabilistic proxy models and paper objectives
   a. Use in fairness assessments
   b. Overestimation issue
2. Threshold estimators for protected attribute classification
   a. Definition
   b. Decomposing bias
3. Weighted estimators
   a. Definition
   b. Single source of bias
4. Results
   a. Synthetic data
   b. Home Mortgage Disclosure Act dataset
Notation

- **Binary decision** $Y$
  - $Y = 1$ represents a favorable outcome, $Y = 0$ represents an unfavorable outcome
- **Protected class** $A$
  - $A = a$ for the advantaged group, $A = b$ for the disadvantaged group
- **Proxy variable** $Z$
  - A set of covariates used to predict $A$ in a probabilistic proxy model
- **Demographic disparity** or **Calders-Verwer gap** $\delta$
  - the difference in mean group outcomes between the advantaged and disadvantaged groups: $\delta = \mu(a) - \mu(b)$
Assume that we have $N$ independent and identically distributed (iid) samples $(Y_i, Z_i)_{i=1}^N$.

The true membership $A_i$ of the $i$\textsuperscript{th} sample is unknown, but we have access to the probabilistic proxy estimates $\{P(A_i = u | Z_i) : u \in \{a, b\}, i \in \{1, \ldots, N\}\}$ by applying our proxy model to the observed proxy variable $Z_i$. 
Definitions: Threshold Estimator

Definition 2.3. Let $q \in \left[\frac{1}{2}, 1\right)$. Then the thresholded estimated membership $\hat{A}_i$ for the $i$th unit is

$$\hat{A}_i = \begin{cases} 
    a, & P(A_i = a \mid Z_i) > q, \\
    b, & P(A_i = b \mid Z_i) > q, \\
    \text{NA}, & \text{otherwise},
\end{cases}$$
Definitions: Threshold Estimator Disparity

Let \((Y_i, Z_i)^N_{i=1}\) be \(N\) iid samples, \(\{\hat{A}_i\}^N_{i=1}\) be the estimated labels according to the thresholding rule, and \(I(S)\) be the indicator function for some set \(S\)

\[
\hat{\mu}_q(u) = \frac{\sum^N_{i=1} I(\hat{A}_i = u)Y_i}{\sum^N_{i=1} I(\hat{A}_i = u)}, u \in \{a, b\},
\]

\[
\hat{\delta}_q = \hat{\mu}_q(a) - \hat{\mu}_q(b).
\]

Thresholded estimators for mean group outcomes and demographic disparity
Discussion: What issues do you see with the threshold estimator?
Discussion: What issues do you see with the threshold estimator?

- Under a probabilistic proxy model, the threshold classification rule inevitably misclassifies some individuals.
- A group of unclassified individuals are removed from the outcome disparity evaluation.
Example 1: Loan Approval

Threshold estimator with $q = 0.5$

assumes that the high-income neighborhood is occupied by the advantaged group $a$ and the low-income neighborhood occupied by the disadvantaged group $b$

Available variables:
- Income
- Geolocation
- Imputed race

Decision:
- Accepts all high-income people and rejects all low-income people
Example 1: Loan Approval

<table>
<thead>
<tr>
<th>Neighborhood</th>
<th>True race</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>high income</td>
<td>70 (100%)</td>
</tr>
<tr>
<td>low income</td>
<td>30 (0%)^ii</td>
</tr>
</tbody>
</table>

\[ \delta = \frac{1}{100}(70 \times 1 + 30 \times 0) - \frac{1}{100}(30 \times 1 + 70 \times 0) = 0.4 \]
\[ \hat{\delta}_{0.5} = \frac{1}{100}(70 \times 1 + 30 \times 1) - \frac{1}{100}(30 \times 0 + 70 \times 0) = 1 \]

^i Misclassified as race a in high-income neighborhood.
^ii Misclassified as race b in low-income neighborhood.

True disparity = 40%
Threshold estimator disparity = 100%

Discussion: Why does the threshold estimator overestimate disparity?
Race proxy is correlated with the loan approval outcome

<table>
<thead>
<tr>
<th></th>
<th>Advantaged group (a)</th>
<th>Disadvantaged group (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High income neighborhood</td>
<td>Classified as group a</td>
<td>Misclassified as group a</td>
</tr>
<tr>
<td>Low income neighborhood</td>
<td>Misclassified as group b</td>
<td>Classified as group b</td>
</tr>
</tbody>
</table>

Overestimates loan acceptance rate of the advantaged group

Underestimates loan acceptance rate of the disadvantaged group
Inter-Geolocation Outcome Variation

- **Inter-geolocation outcome variation**: Socioeconomic status, race proxy probabilities and loan acceptance rates vary across different neighborhoods
  - favorable outcome is positively correlated with the probability of belonging to the advantaged group
Example 2: Loan Approval with Affirmative Action

Threshold estimator with \( q = 0.5 \)

Assume that the high-income neighborhood is occupied by the advantaged group \( a \) and the low-income neighborhood occupied by the disadvantaged group \( b \).

Approves group \( b \) at a higher rate than group \( a \) with the same income level, but overall people with high income are still more likely to be accepted.
Example 2: Loan Approval with Affirmative Action

<table>
<thead>
<tr>
<th>Neighborhood</th>
<th>True race</th>
<th>$\delta$ = $\frac{1}{100}(70 \times 0.7 + 30 \times 0.2) - \frac{1}{100}(30 \times 0.8 + 70 \times 0.3) = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>$\frac{1}{100}(70 \times 0.7 + 30 \times 0.8) - \frac{1}{100}(30 \times 0.2 + 70 \times 0.3) = 0.46$.</td>
</tr>
<tr>
<td>high income</td>
<td>70 (70%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30 (80%)i</td>
<td></td>
</tr>
<tr>
<td>low income</td>
<td>30 (20%)ii</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70 (30%)</td>
<td></td>
</tr>
</tbody>
</table>

$^i$ Misclassified as race $a$ in high-income neighborhood.

$^ii$ Misclassified as race $b$ in low-income neighborhood.

True disparity = 10%
Threshold estimator disparity = 46%

Discussion: Why does the threshold estimator overestimate disparity?
# Differing approval rates

<table>
<thead>
<tr>
<th>Income Level</th>
<th>High income neighborhood</th>
<th>Low income neighborhood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>group a</td>
<td>group b</td>
</tr>
<tr>
<td>High income</td>
<td>Classified as group a</td>
<td>Misclassified as group a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium income</td>
<td>Classified as group a</td>
<td>Misclassified as group a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low income</td>
<td>Classified as group a</td>
<td>Misclassified as group a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Raises average approval rate of the advantaged group a
- Lowers the average approval rate of disadvantaged group b
Intra-Geolocation Outcome Variation

- **Intra-geolocation outcome variation:** People from different protected groups living in the same locations have different chance of getting the favorable outcome.
Bias Calculations for Threshold Estimator Disparity

- $\Delta_1(u)$ is the outcome mean discrepancy for two different protected attributes within the same proxy probability range
- $\Delta_2(u)$ is the outcome mean discrepancy across different proxy probability ranges for the same protected group

$\Delta_1(u) = \mathbb{E}[Y \mid \mathbb{P}(A = u \mid Z) > q, A = u^c] - \mathbb{E}[Y \mid \mathbb{P}(A = u \mid Z) > q, A = u]$

$\Delta_2(u) = \mathbb{E}[Y \mid \mathbb{P}(A = u \mid Z) \leq q, A = u] - \mathbb{E}[Y \mid \mathbb{P}(A = u \mid Z) > q, A = u]$, 

Intra-Geolocation Variation

Inter-Geolocation Variation
**Theorem 3.3.** Let $A$ be a binary protected class with values $a$ and $b$. The bias for the thresholded estimator $\hat{\delta}_q$ in Definition 2.4 is:

$$\hat{\delta}_q - \delta = [\hat{\mu}_q(a) - \mu(a)] - [\hat{\mu}_q(b) - \mu(b)],$$

and as $N \to \infty$, for $u \in \{a, b\}$ and $u^c \in \{a, b\}$ as the class opposite to $u$,

$$\hat{\mu}_q(u) - \mu(u) \to \Delta_1(u)C_1(u) - \Delta_2(u)C_2(u)$$
$$+ (\Delta_1(u) - \Delta_2(u))C_3(u),$$

where

$$C_1(u) = \mathbb{P}(\hat{A} = u \mid A = u)\mathbb{P}(A = u^c \mid \hat{A} = u),$$

$$C_2(u) = \mathbb{P}(A = u \mid \hat{A} = u)\mathbb{P}(\hat{A} \neq u \mid A = u),$$

and

$$C_3(u) = \mathbb{P}(\hat{A} \neq u \mid A = u)\mathbb{P}(A = u^c \mid \hat{A} = u).$$

Here $\hat{A} \neq u$ means that $\hat{A} = u^c$ or $\hat{A} = NA$. 

These conditions characterize the nature/ the source of threshold estimator bias.
Conditions Characterize Threshold Estimator Bias

\[ \Delta_1(a) \geq 0, \quad (i) \]
\[ -\Delta_1(b) \geq 0, \quad (ii) \]
\[ -\Delta_2(a) \geq 0, \quad (iii) \]
\[ \Delta_2(b) \geq 0, \quad (iv) \]

- (i) and (ii) capture intra-geolocational outcome variation
  - on higher proportion of disadvantaged group receive the favorable outcome than the advantaged group
- (iii) and (iv) capture inter-geolocational outcome variation
  - (iii) \( \rightarrow \) decision outcome is positively correlated with belonging to the advantaged group
  - (iv) \( \rightarrow \) decision outcome is negatively correlated with belong to the disadvantaged group
Bias Calculations for Threshold Estimator Disparity

\[ P(A = a | Z) \]

- **disfavored mean outcome**: \( 0, 1 - q \)
- **favored mean outcome**: \( q, 1 \)
- **advantaged group**: \( A = \hat{A} \)

\[ \Delta_1(a) \geq 0 \]

\[ \Delta_2(a) \geq 0 \]

**incorrectly classified** \( A \neq \hat{A} \)

\[ \Delta_1(b) \geq 0 \]

\[ \Delta_2(b) \geq 0 \]

- **disadvantaged group**: \( A = a \)
- **classified correctly**: \( A = b \)
Outline

1. Probabilistic proxy models and paper objectives
   a. Use in fairness assessments
   b. Overestimation issue
2. Threshold estimators for protected attribute classification
   a. Definition
   b. Decomposing bias
3. Weighted estimators
   a. Definition
   b. Single source of bias
4. Results
   a. Synthetic data
   b. Home Mortgage Disclosure Act dataset
Definitions: Weighted Estimator

Unlike the threshold estimator, the new weighted estimator accounts for the soft classification generated by the probabilistic proxy model.

\[ \hat{\mu}_W(u) = \frac{\sum_{i=1}^{N} P(A_i = a \mid Z_i) Y_i}{\sum_{i=1}^{N} P(A_i = a \mid Z_i)}, u \in \{a, b\} \]

\[ \hat{\delta}_W = \hat{\mu}_W(a) - \hat{\mu}_W(b) \]
Bias Calculation for Weighted Estimator

**Theorem 3.1.** Let $A$ be a binary protected class with values $a$ and $b$. The bias of the weighted estimator $\hat{\delta}_W$ in Definition 2.5 for demographic disparity $\delta$ in (1) is

$$
\hat{\delta}_W - \delta = [\hat{\mu}_W(a) - \mu(a)] - [\hat{\mu}_W(b) - \mu(b)],
$$

where as $N \to \infty$, the biases in the weighted estimators for the mean group outcomes $\hat{\mu}_W(u)$, for $u \in \{a, b\}$, converge almost surely to

$$
\hat{\mu}_W(u) - \mu(u) \overset{a.s.}{\longrightarrow} - \frac{\mathbb{E}[\text{Cov}(\mathbb{I}(A = u), Y | Z)]}{\mathbb{P}(A = u)}.
$$

(4)
Breakdown of Weighted Estimator Bias

**In theory:** If \( Y \) is independent of \( A \) conditionally on \( Z \), then \( \delta_w \) is asymptotically unbiased.

**In practice:** If \( Y \) covaries with \( A \) for a given \( Z \), then \( \delta_w \) will be biased.

- Disadvantaged group more likely to have favorable outcome → overestimates bias
- Advantaged group more likely to have favorable outcome → underestimate bias

Intra-geolocation variation
Outline

1. Probabilistic proxy models and paper objectives
   a. Use in fairness assessments
   b. Overestimation issue
2. Threshold estimators for protected attribute classification
   a. Definition
   b. Decomposing bias
3. Weighted estimators
   a. Definition
   b. Single source of bias
4. Results
   a. Synthetic data
   b. Home Mortgage Disclosure Act dataset
Results: Synthetic Data for Loan Approval

- $\lambda$: extent to which loan approval $Y$ depends on income $X$
- $d$: discrepancy of income $X$ between the two race groups within the same geolocation

Conditions (i) - (iv) reflect bias as expected.
Results: HMDA (Home Mortgage Disclosure Act) data set

Thresholding estimator underestimates loan approval of Hispanic/Black populations → overestimates demographic disparity
Results: HMDA (Home Mortgage Disclosure Act) data set

Condition (iv) strictly holds → loan approval is negatively correlated with Hispanic/Black prevalence across location
Summary

1. Derived the bias from a probabilistic proxy model used with the standard thresholded estimator that has been described.

2. Provided sufficient conditions to understand when this methodology is biased and to what extent.

3. Proposed a weighted estimator which propagates the uncertainty resulting from the probabilistic proxy onto the final estimand.
   a. Estimation bias of this weighted estimator has only one bias source.
   b. Useful new method to incorporate into outcome disparity evaluations when proxy models are used.

How do we identify proxy variables? → next paper
Interventional Fairness with Indirect Knowledge of Unobserved Protected Attributes

Galhotra et al, 2021
Causal interventional fairness / paper objectives

Algorithm 1 to identify proxy variables
   a. Problem set-up
   b. Conditional independence tests

Algorithm 2 to identify proxy variables
   a. Constraint satisfaction problem
   b. Constraints

Results
   a. Synthetic data
   b. Real-world datasets
Causal Interventional Fairness

**Definition 1 (Causal Interventional Fairness).** For a given set of admissible attributes $A$, a classifier is considered fair if for any collection of values $a$ of $A$ and output $Y'$, the following holds: 

$$Pr(Y' = y | \text{do}(S) = s, \text{do}(A = a)) = Pr(Y' = y | \text{do}(S) = s', \text{do}(A = a))$$ for all values of $A, S$ and $Y'$.

(i.e.) the probability distribution of the classifier output $Y'$ is independent of the protected attributes when we intervene on the admissible attributes.

Crucial that predictive models are fair when used for decision-making.
Discussion: Even when protected attributes aren’t included in a model, why might that model not meet the requirements for casual interventional fairness?
Challenge to Causal Interventional Fairness: Proxy Variables

Proxy variables introduce bias
Challenge to Causal Interventional Fairness: Proxy Variables

Even when protected attributes are hidden and proxy variables are unknown, auditors can use indirect knowledge to flag potentially biased results.

<table>
<thead>
<tr>
<th>Name</th>
<th>Degree</th>
<th>School</th>
<th>GPA</th>
<th>Zip Code</th>
<th>Recommended for Interview?</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>BS</td>
<td>Duke</td>
<td>3.3</td>
<td>01234</td>
<td>Yes</td>
</tr>
<tr>
<td>Jane Williams</td>
<td>BS</td>
<td>UNC</td>
<td>2.9</td>
<td>01234</td>
<td>No</td>
</tr>
<tr>
<td>Carla Lopez</td>
<td>MS</td>
<td>MIT</td>
<td>3.4</td>
<td>56789</td>
<td>No</td>
</tr>
</tbody>
</table>
Paper Objective

How do we identify proxy attributes when we have no knowledge of the protected attribute or model structure?

Feedback-based framework that uses indirect knowledge to isolate proxy variables
Framework Approach

Auditors review model decisions and flag potentially biased outcomes → Flagged samples guide conditional independence (CI) tests → CI tests reveal proxy variables responsible for bias
Outline

1. Causal interventional fairness / paper objectives
2. Algorithm 1 to identify proxy variables
   a. Problem set-up
   b. Conditional independence tests
3. Algorithm 2 to identify proxy variables
   a. Constraint satisfaction problem
   b. Constraints
4. Results
   a. Synthetic data
   b. Real-world datasets
# Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>Unobserved protected attribute</td>
</tr>
<tr>
<td>$\mathcal{V}$</td>
<td>Set of attributes (also known as variables of the causal graph)</td>
</tr>
<tr>
<td>$D$</td>
<td>Input dataset containing $\mathcal{V}$ attributes</td>
</tr>
<tr>
<td>$\mathcal{Y}$</td>
<td>Prediction attribute</td>
</tr>
<tr>
<td>$\mathcal{Y}'$</td>
<td>Classifier output</td>
</tr>
<tr>
<td>$F$</td>
<td>Feedback attribute</td>
</tr>
<tr>
<td>$D'$</td>
<td>Feedback set</td>
</tr>
<tr>
<td>$\mathcal{V}' \subseteq \mathcal{V}$</td>
<td>Proxy attributes</td>
</tr>
<tr>
<td>$\mathcal{V}_F \subseteq \mathcal{V}$</td>
<td>Parents of $F$ in the causal graph</td>
</tr>
</tbody>
</table>
Problem Set-Up

- **Known**
  - Training dataset D contains $V = \{X_1, X_2, X_3\}$
  - Feedback dataset $D'$ which is equivalent to conditioning $F = 1$

- **Unknown:**
  - Model structure
  - Unobserved protected attribute $S$
  - The attributes that impact $F$, $V^F = \{X_1, X_3\}$
  - The proxy attributes, $V' = \{X_1, X_2\}$
Feedback Attributes

- An auditor inspects model records and flags biased outcomes, which get denoted with the extra attribute, $F$
  - $F$ is a function of a subset $V' \subseteq V$ and the learned target $Y'$ such that $F = 1$ refers to a biased prediction
Complaint Set

The set of complaints $D'$ is comprised of attributes $V$ for a small subset where $F = 1$
- Not all biased samples, only the flagged samples
- Any CI test of the form $A \perp_{D'} B|C$ on the sample $D'$ is equivalent to conditioning on the attribute $F$ along with $C$
- **Assumption:** all individuals in the complaint set are from the same marginalized group

<table>
<thead>
<tr>
<th>Name</th>
<th>Degree</th>
<th>School</th>
<th>GPA</th>
<th>Zip Code</th>
<th>Recommended?</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>BS</td>
<td>Duke</td>
<td>3.3</td>
<td>01234</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>Joe Smith</td>
<td>BS</td>
<td>UNC</td>
<td>2.9</td>
<td>01234</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Carla Lopez</td>
<td>MS</td>
<td>MIT</td>
<td>3.4</td>
<td>56789</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>BS</td>
<td>UNC</td>
<td>3.9</td>
<td>23786</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Allison Williams</td>
<td>MS</td>
<td>Duke</td>
<td>3.5</td>
<td>12345</td>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>
Goal: Identify Proxy Variables

Due to two properties of causal directed acyclic graphs (DAG), conditional independence tests can be performed on pairs of variables to isolate the proxy variables, V'
$X_1$ and $X_2$ can be d-separated with respect to $D$ and $D'$ if $X_1, X_2$ are proxy attributes

**Lemma 2.** Consider a pair of attributes $X_1$ and $X_2 \in \mathcal{V}$ with $(X_1, X_2) \notin E$. $X_1, X_2 \in \mathcal{V}'$, and at least one of $X_1$ and $X_2$ does not belong to $\mathcal{V}_F$ iff

1. $X_1 \notin X_2 | A$ for all $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$ and
2. $X_1 \perp X_2 | A, F$ for some $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$

If a pair of variables is d-separated:

- They are proxy attributes
- At least one of them does not impact the feedback attributed
$X_1$ and $X_2$ can never be d-separated in $D'$ if one of $X_1$ and $X_2$ is not a proxy attribute

**Lemma 3.** For a pair of attributes $X_1$ and $X_2 \in \mathcal{V}$ with $(X_1, X_2) \notin E$, $X_1, X_2 \in \mathcal{V}_F$, and at least one of $X_1$ and $X_2$ does not belong to $\mathcal{V}'$ iff

1. $X_1 \perp X_2 | A$ for some $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$
2. $X_1 \not\perp X_2 | A$, $F$ for all $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$

If a pair of variables is can’t be d-separated:

- At least one of them is not a proxy attribute
# Conditional Independence Properties

**Table 2.** Conditional independence properties for a pair of attributes $X_1, X_2 \in \mathcal{V}$ such that $(X_1, X_2) \notin E$ where the output of conditional independence tests varies based on the set that $X_1, X_2$ belong to and vice versa. For example, $X_1, X_2 \in \mathcal{V'} \cap \mathcal{V}_F$ iff $X_1 \not\perp X_2|A$ and $X_1 \not\perp X_2|A, F$ for all $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$.

<table>
<thead>
<tr>
<th>Conditions on $X_1, X_2$</th>
<th>Conditioning on $D$</th>
<th>Conditioning on $D'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1, X_2 \in \mathcal{V'} \cap \mathcal{V}_F$</td>
<td>$X_1 \not\perp X_2</td>
<td>A$ for all $A \subseteq \mathcal{V} \setminus {X_1, X_2}$</td>
</tr>
<tr>
<td>$X_1, X_2 \in \mathcal{V}_F$ and (X_1 \notin \mathcal{V'} and/or X_2 \notin \mathcal{V'}) (Lemma 3) and</td>
<td>$X_1 \perp X_2</td>
<td>A$ for some $A \subseteq \mathcal{V} \setminus {X_1, X_2}$</td>
</tr>
<tr>
<td>$X_1, X_2 \in \mathcal{V'}$ and (X_1 \notin \mathcal{V}_F and/or X_2 \notin \mathcal{V}_F) (Lemma 2)</td>
<td>$X_1 \not\perp X_2</td>
<td>A$ for all $A \subseteq \mathcal{V} \setminus {X_1, X_2}$</td>
</tr>
<tr>
<td>$X_1 \notin \mathcal{V'} \setminus \mathcal{V}_F$ and X_2 \in \mathcal{V}_F \setminus \mathcal{V'}</td>
<td>$X_1 \perp X_2</td>
<td>A$ for some $A \subseteq \mathcal{V} \setminus {X_1, X_2}$</td>
</tr>
<tr>
<td>$X_1 \notin \mathcal{V'} \cup \mathcal{V}_F$</td>
<td>$X_1 \perp X_2</td>
<td>A$ for some $A \subseteq \mathcal{V} \setminus {X_1, X_2}$</td>
</tr>
</tbody>
</table>
Algorithm 1 Proxy identification.

1: Input: attributes $\mathcal{V}, F$
2: $\mathcal{V}' \leftarrow \emptyset$
3: for $X_1 \in \mathcal{V} \setminus \mathcal{V}'$ do
4:   for $X_2 \in \mathcal{V}$ do
5:     if $\exists A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$ $|$ $(X_1 \perp X_2 | F, A)$ then
6:       if $\forall A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$ $|$ $(X_1 \not \perp X_2 | A)$ then
7:         $\mathcal{V}' \leftarrow \mathcal{V}' \cup \{X_1, X_2\}$
8: return $\mathcal{V}'$

Lemma 4. An attribute $X \in \mathcal{V}'$ is correctly identified to belong to $\mathcal{V}'$ if $\exists X' \in \mathcal{V}'$ such that $(X, X') \not \in E$ and $|\mathcal{V}_F \cap \{X, X'\}| \leq 1.$
Discussion: What are the potential drawbacks of Algorithm 1?
Discussion: What are the potential drawbacks of Algorithm 1?

- In dense graphs, there may exist an attribute \( X \in V' \) such that \( E \in V' \) where \((X, X') \in E\). These attributes may not be identified by Algorithm 1.

- The conditional independence test of the form \( X_1 \perp X_2 | A, \forall A \subseteq V \setminus \{X_1, X_2\} \) requires us to test the conditional dependence for every subset \( A \), requiring an exponential number of conditional independence tests.
Outline

1. Causal interventional fairness / paper objectives
2. Algorithm 1 to identify proxy variables
   a. Problem set-up
   b. Conditional independence tests
3. Algorithm 2 to identify proxy variables
   a. Constraint satisfaction problem
   b. Constraints
4. Results
   a. Synthetic data
   b. Real-world datasets
Constraint Satisfaction Formulation

- For each attribute $X \in V$ there are two binary variables $X^F$ and $X^S \in \{0, 1\}$
  - $X^F = 1$ if $X \in V^F$ and 0 otherwise
  - $X^S = 1$ if $X \in V'$ and 0 otherwise

- $X_1^S = 1$, $X_1^F = 1$
- $X_2^S = 1$, $X_2^F = 0$
- $X_3^S = 0$, $X_3^F = 1$
**Constraint 1**

**Lemma 2.** Consider a pair of attributes $X_1$ and $X_2 \in \mathcal{V}$ with $(X_1, X_2) \notin E$. $X_1, X_2 \in \mathcal{V}'$, and at least one of $X_1$ and $X_2$ does not belong to $\mathcal{V}_F$ iff

1. $X_1 \not\perp X_2 | A$ for all $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$ and
2. $X_1 \perp X_2 | A, F$ for some $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$

\[ X_1^S = X_2^S = 1 \text{ and } X_1^F + X_2^F \leq 1 \]
Constraint 2

Lemma 3. For a pair of attributes $X_1$ and $X_2 \in \mathcal{V}$ with $(X_1, X_2) \notin E$, $X_1, X_2 \in \mathcal{V}_F$, and at least one of $X_1$ and $X_2$ does not belong to $\mathcal{V}'$ iff
1. $X_1 \perp X_2 | A$ for some $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$
2. $X_1 \not\perp X_2 | A, F$ for all $A \subseteq \mathcal{V} \setminus \{X_1, X_2\}$

$X_1^F = X_3^F = 1$ and $X_1^S + X_3^S \leq 1$
Constraint 3

\[ x_1^F + x_1^S + x_2^F + x_2^F \leq 2 \]
Efficient Implementation

Algorithm 2 improves efficiency by leveraging Lemma 5

**Lemma 5.** Consider a pair $X_1$ and $X_2$ such that $(X_1, X_2) \notin E$ and at least one of the two attributes does not belong to $\mathcal{V}'$. The following conditions hold:

1. $X_1$ and $X_2$ are independent when conditioned on all other attributes $(X_1 \perp X_2 | \mathcal{V} \setminus \{X_1, X_2\})$ iff there does not exist $X' \in \mathcal{V}$ such that $X_1 \rightarrow X' \leftarrow X_2$ form a collider path.
2. $\exists V_1$ such that $X_1 \perp X_2 | V_1$ where $|V_1| \geq n - t$ iff the number of attributes in set $\mathcal{V}'$ is less than $t$, where $\mathcal{V}'$ contains all attributes $X \in \mathcal{V}$ that form a length-2 collider path $X_1 \rightarrow X \leftarrow X_2$ or $X$ is a descendant of some attribute $X' \in \mathcal{V}'$, where $X'$ forms a length-2 collider path.
**Algorithm 2** Proxy identification.

1: \textbf{Input:} attributes \( \mathcal{V}, F \)
2: \( U \leftarrow \mathcal{V}, C \leftarrow \phi \)
3: \( X^s, X^f \leftarrow \{0, 1\}, \forall X \in \mathcal{V} \)
4: \( t \leftarrow |\mathcal{V}| \)
5: \textbf{while} \( t \geq 0 \) and \( U \neq \phi \) \textbf{do} 
6: \( T \leftarrow \text{IDENTIFYSUBSET}(\mathcal{V}, t) \)
7: \( C \leftarrow C \cup \text{PairwiseConstraints}(\mathcal{V}, T) \)
8: \( \text{SolveCSP}(\mathcal{V}, C) \)
9: \( U \leftarrow \{X : 0,1 \in X^s, X \in \mathcal{V} \} \)
10: \( t \leftarrow t - 1 \)
11: \( \mathcal{V}' \leftarrow \{X : X^s = \{1\} \} \)
12: \textbf{return} \( \mathcal{V}' \)

**Algorithm 3** Pairwise constraints.

\textbf{Input:} Attributes \( \mathcal{V}, F, T \)

\( C \leftarrow \phi \)

\textbf{for} \( (X_1, X_2) \in \mathcal{V} \times \mathcal{V} \) \textbf{do}  
if \( \exists T \in T \mid X_1 \perp X_2 \mid T \setminus \{X_1, X_2\} \) and \( X_1 \not\perp X_2 \mid T \setminus \{X_1, X_2\}, F \forall T \in \mathcal{T} \) then  
\( C \leftarrow C \cup \{X_1^f, X_2^f \leftarrow 1\} \)
\( C \leftarrow C \cup \{X_1^s + X_2^s \leq 1\} \)

if \( X_1 \not\perp X_2 \mid T \setminus \{X_1, X_2\} \) and \( X_1 \perp X_2 \mid T \setminus \{X_1, X_2\}, F \) then  
\( C \leftarrow C \cup \{X_1^f \leftarrow 1\} \)
\( C \leftarrow C \cup \{X_1^f + X_2^f \leq 1\} \)

if \( X_1 \perp X_2 \mid T \setminus \{X_1, X_2\} \) and \( X_1 \not\perp X_2 \mid T \setminus \{X_1, X_2\}, F \) then  
\( C \leftarrow C \cup \{X_1^s + X_2^s \leq 2\} \)

\textbf{return} \( C \)
Outline

1. Causal interventional fairness / paper objectives
2. Algorithm 1 to identify proxy variables
   a. Problem set-up
   b. Conditional independence tests
3. Algorithm 2 to identify proxy variables
   a. Constraint satisfaction problem
   b. Constraints
4. Results
   a. Synthetic data
   b. Real-world datasets
# Result: Real-world datasets

Table 3. Comparison of accuracy (Acc), average precision (AvgP), and absolute odds difference (AOD).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OurApproach</th>
<th>All</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>AvgP</td>
<td>OD</td>
</tr>
<tr>
<td>Adult</td>
<td>0.79</td>
<td>0.78</td>
<td>0.025</td>
</tr>
<tr>
<td>Adult-balanced</td>
<td>0.78</td>
<td>0.71</td>
<td>0.068</td>
</tr>
<tr>
<td>MEPS</td>
<td>0.85</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>MEPS-balanced</td>
<td>0.77</td>
<td>0.67</td>
<td>0.25</td>
</tr>
<tr>
<td>German</td>
<td>0.74</td>
<td>0.7</td>
<td>0.075</td>
</tr>
<tr>
<td>German-balanced</td>
<td>0.70</td>
<td>0.66</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Results: Synthetic Data

Figure 2. Complexity comparison of our techniques for varying dataset sizes.
Summary

- Formalized a feedback based framework for interventional fairness in settings where the protected attribute is unobserved and where auditors or users report issues in the classifier.
- Developed efficient techniques that use CI testing over the observational data to formulate a constraint satisfaction problem, which identifies the proxy variables.
- Demonstrated efficacy on real-world and synthetic datasets.