Anti-Discrimination Learning: from Association to Causation

Lu Zhang
Yongkai Wu
Xintao Wu

Social Awareness and Intelligent Learning Lab
University of Arkansas

Detailed Outline

• Part I: Introduction (5 Min)
  – Context
  – Literature & Resource

• Part II: Correlation based Anti-Discrimination Learning (45 Min)
  – Measures
  – Algorithms
  – From Correlation to Causation

• Part III: Causal Modeling Background (40 Min, Video by Lu Zhang)
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference

• Break (9:30am – 10:am)
Detailed Outline

• Part IV: Causal Modeling based Anti-Discrimination Learning (60 Min)
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works

• Part V: Challenges and Directions for Future Research (30 Min)
  – Challenges (20 Min, Video by Lu Zhang)
  – Future Research

• Discussions and Wrap-up (30 Min)
Outline

• Part I: Introduction
  – Context
  – Literature & Resource

• Part II: Correlation based Anti-Discrimination Learning

• Part III: Causal Modeling Background

• Part IV: Causal Modeling based Anti-Discrimination Learning

• Part V: Challenges and Directions for Future Research
Introduction

• Discrimination refers to unjustified distinctions of individuals based on their membership in a certain group.

• Federal Laws and regulations disallow discrimination on several grounds:
  – Gender, Age, Marital Status, Race, Religion or Belief, Disability or Illness
  ......
  – These attributes are referred to as the **protected attributes**.
Discrimination Cases

• Discrimination in U.S. against people of color and women, especially before 1964

• COMPAS – Correctional Offender Management Profiling for Alternative Sanctions from Northpointe, Inc.
  – Predictive model for risk of recidivism
  – Prediction accuracy of recidivism for blacks and whites is about the same
  – However
  • Blacks who did not reoffend were classified as high risk twice as much as whites who did not reoffend
  • Whites who did reoffend were classified as low risk twice as much as blacks who did reoffend
Laws and Regulations

- Equal Pay Act of 1963
- Title VII of Civil Rights Act of 1964
- Age Discrimination in Employment Act of 1967
- Vietnam Era Vets Readjustment Act of 1974
- Pregnancy Discrimination Act of 1978
- Americans with Disabilities Act of 1990
- Revision of the Civil Rights Act (1991)
Introduction

May 2014

Big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups.
Introduction

February 2015

Pay attention to the potential for big data to facilitate discrimination

Expand technical expertise to stop discrimination

Deepen understanding of differential pricing
Introduction

May 2016

Support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks.
Anti-Discrimination Learning

Discover and remove discrimination from the training data

Build discrimination-free classifier
Anti-Discrimination Learning

• Discrimination Discovery/Detection
  – Unveil evidence of discriminatory practices by analyzing the historical dataset or the predictive model.

• Discrimination Prevention/Removal
  – Ensure non-discrimination by modifying the biased data (before building predictive models) or twisting the predictive model.
Discrimination Categorization

• From the perspective of in what way discrimination occurs, discrimination is legally divided into
  – **Direct**: explicitly based on the protected attributes.
    • E.g., rejecting a qualified female just because of her gender.
  – **Indirect**: based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group.
    • E.g., redlining, where the residential Zip Code of an individual is used for making decisions such as granting a loan.
Disparate Treatment vs. Impact

• Disparate treatment
  – Intentional effect on protected group
  – To enforce *procedural fairness*, the equality of treatments should prohibit the use of the protected attribute in the decision process.

• Disparate impact
  – Unintentional adverse impact on members of protected group
  – To guarantees *outcome fairness*, the equality of outcomes should be achieved.
Discrimination Categorization

- From the perspective of different level of granularity in studying, discrimination can be divided into
  - **System** level: the average discrimination across the whole system, e.g., all applicants to a university.
  - **Group** level: the discrimination that occurs in one particular subgroup, e.g., the applicants applying for a particular major, or the applicants with a particular score.
  - **Individual** level: the discrimination that happens to one particular individual, e.g., one particular applicant.
Discrimination Categorization

- Fairness measure for historical data
- Fairness measure for supervised learning
  - E.g., pedestrians are stopped on the suspicion of possessing an illegal weapon, having different weapon discovery rates for different races.
  - Equality of Opportunity
    - True positive rate of a predictor should be the same for all the groups.
Outline

• Part I: Introduction
  – Context
  – Literature and Resource

• Part II: Correlation based Anti-Discrimination Learning

• Part III: Causal Modeling Background

• Part IV: Causal Modeling based Anti-Discrimination Learning

• Part V: Challenges and Directions for Future Research
Resources

• Tutorials and keynotes
  – Hajian, S., Bonchi, F., Castillo, C. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. Tutorial of KDD 2016
  – Abiteboul, S., Miklau, G., Stoyanovich J. Data Responsibly: Fairness, Neutrality and Transparency in Data Analysis, Tutorial of EDBT 2016
  – Dwork, C. What’s Fair. Keynote of KDD 2017

• Survey papers and books
Resources

• Conferences/Workshops/Symposiums
  – ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)
  – Fairness, Accountability, and Transparency in Machine Learning (FATML)
  – AAAI/ACM Conference on AI, Ethics, and Society (AIES)
  – Workshop on Responsible Recommendation (FAT/Rec)
  – Workshop on Data and Algorithmic Bias (DAB)
  – Ethics in Natural Language Processing
  – Workshop on Fairness, Accountability, and Transparency on the Web (FAT/WEB)
  – Special Session on Explainability of Learning Machines
  – Workshop on Data and Algorithmic Transparency (DAT)
  – The Human Use of Machine Learning: An Interdisciplinary Workshop
  – International Workshop on Privacy and Discrimination in Data Mining
  – Machine Learning and the Law
  – Interpretable Machine Learning for Complex Systems
  – Workshop on Human Interpretability in Machine Learning
  – Workshop on the Ethics of Online Experimentation
  – Auditing Algorithms From the Outside: Methods and Implications
  – Discrimination and Privacy-Aware Data Mining
  – Workshop on Novelty and Diversity in Recommender Systems
• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
  – Measures
  – Algorithms
  – From Correlation to Causation
• Part III: Causal Modeling Background
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Notations

- Denote an attribute by an uppercase alphabet, e.g., $X$
- Denote a value of attribute $X$ by $x$
- Denote a subset of attributes by a bold uppercase alphabet, e.g., $\mathbf{X}$
- Denote a value assignment of attributes $\mathbf{X}$ by $\mathbf{x}$

- A binary protected attribute $C = \{c^+, c^-\}$ (sometimes use $A = \{a^+, a^-\}$ or $S = \{s^+, s^-\}$).
- A binary decision $E = \{e^+, e^-\}$ (sometimes use $Y = \{y^+, y^-\}$).
- Non-protected attributes $\mathbf{X}$ among which $\mathbf{R}$ are redlining attributes.
- A predictor of decision $\hat{E} = f(C, \mathbf{X})$ (sometimes use $\hat{Y} = f(C, \mathbf{X})$).
• Gender discrimination in college admission.

<table>
<thead>
<tr>
<th>No.</th>
<th>gender</th>
<th>major</th>
<th>score</th>
<th>height</th>
<th>weight</th>
<th>ad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>CS</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>CS</td>
<td>B</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>CS</td>
<td>A</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>CS</td>
<td>A</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>CS</td>
<td>C</td>
<td>low</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>CS</td>
<td>C</td>
<td>median</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>EE</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
</tbody>
</table>

\[
C \text{ is gender, } c^{-} = \text{female}, c^{+} = \text{male}. \\
E \text{ is admission, } e^{-} = \text{reject}, e^{+} = \text{admit}.\]
Measuring Discrimination

- Fairness through unawareness
- Disparate impact
- Individual fairness
- Statistical parity
- Equality of opportunity
- Calibration
- Metrics considering $X$
  - Conditional discrimination
  - $\alpha$-discrimination based on association rules
  - Multi-factor interactions
  - belift based on Bayesian networks
- Preference
Conditional Independence

- Two random variables $X$ and $Y$ are called independent, if for each values of $X$ and $Y$, $x$ and $y$,
  - $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ or
  - $P(X = x|Y = y) = P(X = x)$ or $P(Y = y|X = x) = P(Y = y)$
  - Denoted by $X \perp Y$

- Two random variables $X$ and $Y$ are called conditionally independent given $Z$, if for each values of $(X, Y, Z)$, $(x, y, z)$,
  - $P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$ or
  - $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$ or
  - $P(Y = y|X = x, Z = z) = P(Y = y|Z = z)$
  - Denoted by $X \perp Y|Z$

- Note: conditional independence neither implies nor is implied by independence.
A predictor is said to achieve fairness through unawareness if protected attributes $C$ are not explicitly used in the prediction process.

- The approach of being blind to counter discrimination.
- Prevent disparate treatment.
- Not a sufficient condition to avoid discrimination as $X$ can contain discriminatory information.

Disparate Impact

- Disparate Impact (DI) aims for unintentional bias
  - No rigid math formula
  - Feldman et al. define DI with risk ratio
    \[ DI = \frac{c/(a+c)}{d/(b+d)} \]
  - propose a test for DI based on how well the \( C \) can be predicted from \( X \)
    \( f: X \rightarrow C \) is a predictor of \( C \) from \( X \).
  - Balanced Error Rate (BER):
    \[ BER(f(X), C) = \frac{P(f(X) = c^- | C = c^+) + P(f(X) = c^+ | C = c^-)}{2} \]
  - A dataset is \( \epsilon \)-fairness if \( BER(f(X), C) > \epsilon \)

Individual Fairness

• Similar predictions to similar individuals
• Consistency for individual $i$
  – $Cons_i = 1 - \frac{1}{k} \sum_{j \in kNN(i)} |e_i - e_j|$
  – Compare the outcome of an individual with its $k$-nearest neighbors
  – Note that the similar individuals may be from the protected group and all are treated badly.
• Consistency for the whole data
  – $Cons = 1 - \frac{1}{NK} \sum_i \sum_{j \in kNN(i)} |e_i - e_j|$
• Distance function must be carefully chosen.

Situation Testing

• A legally grounded technique for analyzing the discriminatory treatment on an individual adopted both in the US and the EU.

• In responding to complaint about discrimination:
  1. Pairs of testers who are similar to the individual are sent out to participate in the same decision process (e.g., applying for the same job).
  2. For each pair, the two testers possess the same characteristics except the membership to the protected group.
  3. The distinction of decisions between the protected group and the non-protected group implies discriminatory behavior.
$k$NN-Based Situation Testing

- Given a individuals tuple $t$ with $c^-$ and $e^-$;
- Rank all the individuals according to their distances to $t$;
- Select the individuals that closest to $t$;
  - individuals with $c^+$ are added into set $S^+$
  - individuals with $c^-$ are added into set $S^-$;
- If $P(e^+|S^+) - P(e^+|S^-) > \tau$, then $t$ is considered as being discriminated.

Luong, B.T., Ruggieri, S., Turini, F.: $k$-NN as an implementation of situation testing for discrimination discovery and prevention. In: SIGKDD'11 (2011)
Statistical Parity

- Risk Difference (RD), UK law
- Risk Ration (RR), EU Court of Justice
- Relative Chance (RC)
- Odds Ratio (OR)
- Extended Risk Difference (ED)
- Extended Risk Ratio (ER)
- Extended Chance (EC)

Protected group vs. unprotected group

Protected group vs. entire population

\[
\begin{align*}
&\text{benefit} \\
&\begin{array}{ccc}
\text{group} & \text{granted} & \text{denied} \\
\text{unprotected} & a & b & n_1 \\
\text{protected} & c & d & n_2 \\
\end{array} \\
&m_1 & m_2 & n
\end{align*}
\]

\[
p_1 = \frac{a}{n_1} \quad p_2 = \frac{c}{n_2} \quad p = \frac{m_1}{n}
\]

\[
RD = p_1 - p_2 \quad RR = \frac{p_1}{p_2} \quad RC = \frac{1 - p_1}{1 - p_2} \quad OR = \frac{RR}{RC} = \frac{a/b}{c/d} \quad ED = p_1 - p \quad ER = \frac{p_1}{p} \quad EC = \frac{1 - p_1}{1 - p}
\]
Statistical Parity

• Naturally extend to subgroups, e.g., admission rate difference between female and male applying for CS
  \[ P(e^+ | c^+, X = x) - P(e^+ | c^-, X = x) \] where \( X \) can be \( \emptyset \).

• Individual fairness vs. group fairness
  – (Dwork et al.) show if a predictor satisfies Lipschitz property, it also achieves statistical parity with certain bias.

• Statistical parity is independent of the ground truth, i.e., the label information, when applied to the predictor.
  – Equal opportunity utilizes the ground truth.

Equality of Opportunity

- Equality of opportunity\[^{[1]}\], mistreatment parity\[^{[2]}\], predictive equality\[^{[3]}\]:
  - Target a classifier or predictive model $\hat{E}$.
  - Accuracy of predictions is equal across protected and non-protected groups.

- Equalized odds:
  $$P(\hat{E} = e^+ | C = c^-, E = e) = P(\hat{E} = e^+ | C = c^+, E = e), \quad e \in \{e^+, e^-\}$$

- Equal opportunity:
  $$P(\hat{E} = e^+ | C = c^-, E = e^+) = P(\hat{E} = e^+ | C = c^+, E = e^+ )$$
  - True positive rate should be the same for all the groups.

---


Test Fairness

• Test fairness (calibration)
  \[ P(E = e^+ | C = c^-, \hat{E} = e^+) = P(E = e^+ | C = c^+, \hat{E} = e^+) \]
  – Classifier precision should be the same for all the groups.

• COMPAS
  – ProPublica showed that COMPAS score used by Northpointe violated equalized odds, incurring racial discrimination.
  – Northpointe responded that COMPAS score satisfied calibration.

• Kleinberg et al. showed that Equalized Odds and Test Fairness cannot be satisfied at the same time except in special cases such as zero prediction error or if \( C \) independent of \( E \)

Fundamental Discrimination Criteria

- **Independence**
  - Data: $E$ independent of $C$ ($E \perp C$)
  - Prediction: $\hat{E}$ independent of $C$ ($\hat{E} \perp C$)

- **Separation**
  - $\hat{E}$ independent of $C$ conditional on $E$ ($\hat{E} \perp C \mid E$)

- **Sufficiency**
  - $E$ independent of $C$ conditional on $\hat{E}$ ($E \perp C \mid \hat{E}$)

Conditional Discrimination

- $diff = P(e^+ | c^+) - P(e^+ | c^-)$ is a sum of the explainable and the bad discrimination.
  - $D_{all} = D_{exp} + D_{bad} = P(e^+ | c^+) - P(e^+ | c^-)$

- Explainable Discrimination
  - $D_{exp} = \sum_i P(x_i | c^+)P^*(e^+ | x_i) - \sum_i P(x_i | c^-)P^*(e^+ | x_i)$
  - $P^*(e^+ | x_i) = \frac{P(e^+ | x_i, c^+) + P(e^+ | x_i, c^-)}{2}$
  - $X$ is an explanatory attribute and $x_i$ is its $i$-th domain value

### Examples

#### Example 1

<table>
<thead>
<tr>
<th>Major</th>
<th>Medicine</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td># of applicants</td>
<td>800</td>
<td>200</td>
</tr>
<tr>
<td>Acceptance rate (%)</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

\[
P(\text{accepted}|\text{male}) = 36\% \\
P(\text{accepted}|\text{female}) = 24\% \\
D_{\text{all}} = 12\% \\
D_{\text{exp}} = 12\% \\
D_{\text{bad}} = 0\%
\]

#### Example 2

<table>
<thead>
<tr>
<th>Major</th>
<th>Medicine</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td># of applicants</td>
<td>800</td>
<td>200</td>
</tr>
<tr>
<td>Acceptance rate (%)</td>
<td>15%</td>
<td>25%</td>
</tr>
</tbody>
</table>

\[
P(\text{accepted}|\text{male}) = 41\% \\
P(\text{accepted}|\text{female}) = 19\% \\
D_{\text{all}} = 22\% \\
D_{\text{exp}} = 12\% \\
D_{\text{bad}} = 10\%
\]
\(\alpha\)-Discrimination in Association Rules

- **Direct Discrimination**
  - \( C, X \rightarrow E \)
    - \( \text{elift}(C, X \rightarrow E) = \frac{\text{conf}(C, X \rightarrow E)}{\text{conf}(X \rightarrow E)} \geq \alpha \)
    - \( C \) is a protected attribute
    - \( X \) is a context attribute
    - \( E \) is a decision attribute

- **Indirect Discrimination**
  - \( X_1, X_2 \rightarrow E \)
    - \( X_1, X_2 \) are both context attributes
    - \( X_1, X_2 \) are strongly correlated with \( C \)
    - \( E \) is a decision attribute

Multi-Factor Interaction

• Build a loglinear model from categorical data
• Measure the discrimination based on the strength of interactions among categorical attributes in the fitted model

Data:
A 3-D table \((C, E, X)\) where a cell is denoted as \((i, j, k, m_{ijk})\)

\[
\log(m_{ijk}) = \gamma + \gamma_i^C + \gamma_j^E + \gamma_k^X + \gamma_{ij}^{CE} + \gamma_{ik}^{CX} + \gamma_{jk}^{XE} + \gamma_{ijk}^{CEX}
\]

\[
I_{ij|k}^{CE|X} = \gamma_{ik}^{CE} + \gamma_{ijk}^{CEX}
\]

\[
\log(OR) = I_{ij|k}^{CE|X} + I_{i'j'|k}^{CE|X} - I_{ij'k}^{CE|X} - I_{i'j'k}^{CE|X}
\]

• Extendable to multiple protected/decision attributes

**belift** Based on Bayesian networks

- \( belift = \frac{P(e^+|c_1, c_2, ..., c_l, x_1, x_2, ..., x_m, r_1, r_2, ..., r_n)}{P'(e^+|x_1, x_2, ..., x_m)} \)
  - \( C_i \) is a protected attribute
  - \( X_i \) is a non-protected attribute
  - \( R_i \) is a redlining attribute
  - \( belift = 1 \): perfect equality

- Two bayesian networks are built from data to calculate conditional probabilities.

Discrimination discovery using \textit{belift}

- Build a Bayesian network $G$ from training dataset $D$
- Build a relative Bayesian network $G'$ by removing protected attributes and any attribute directly connected to them in $G$
- For each instance in $D$
  - Compute $P(e^+|c_1, c_2, \ldots, c_l, x_1, x_2, \ldots, x_m, r_1, r_2, \ldots, r_n)$ over $G$
  - Compute $P'(e^+|x_1, x_2, \ldots, x_m)$ over $G'$
  - Calculate \textit{belift} and report discrimination if it exceeds a threshold
Preference-based Fairness

• Inspired by fair division and envy-freeness
• Preference-based notions relax stringent parity-based notations
  – Preferred treatment
    • Ensure each sensitive attribute group prefers the set of decisions over the set they would have received if they had been a different group.
  – Preferred impact
    • Ensure each sensitive attribute group prefers the set of decisions over the set they would have received under the criterion of impact parity.
  – Pareto-efficiency
    • A Pareto-efficient solution is such that there can be no increase in the benefit of one group without strictly decreasing the benefit of another group.

Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
  – Measures
  – Algorithms
  – From Correlation to Causation
• Part III: Causal Modeling Background
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Anti-Discrimination Learning

Build discrimination-free predictive model

- Pre-processing: modify the training data
- In-processing: adjust the learning process
- Post-processing: directly change the predicted labels
Anti-Discrimination Learning

• Pre-processing
  – Data modification
  – Fair data representation
  – Fair data generation

• In-processing
  – Regularization
  – Explicit constraints

• Post-processing
Discrimination Prevention

- Data manipulation (Pre-processing)
    - Suppression/Massaging/Reweighting/Sampling (uniform vs. preferential sampling)
Massaging

• Flip the decision of some individuals according to a ranker
  1. Learn a classifier and estimate the predicted probability of the positive decision of each individual
  2. Sort the individuals of four groups according to this probability
  3. Flip the decision of individuals that close to the bottom/top

\[ RD = \frac{6}{10} - \frac{4}{10} = 0.2 \]

\[ RD = \frac{5}{10} - \frac{5}{10} = 0.0 \]

Kamiran, F., Calders, T.: Classifying without discriminating. IC4’09 (2009)
Preferential Sampling

- Partition the data into 4 groups \((c^+ e^+, c^- e^-, c^- e^+, c^+ e^-)\) and two are under-sampled and two over-sampled
- Select and remove/duplicate the individuals close to the top/bottom

\[
RD = \frac{12}{20} - \frac{9}{20} = 0.15
\]

\[
RD = \frac{10}{20} - \frac{10}{20} = 0
\]

Conditional Discrimination

- $\text{diff} = P(e^+ | c^+) - P(e^+ | c^-)$ is a sum of the explainable and the bad discrimination.
  - $D_{all} = D_{exp} + D_{bad} = P(e^+ | c^+) - P(e^+ | c^-)$

- Explainable Discrimination
  - $D_{exp} = \sum_i P(x_i | c^+)P^*(e^+ | x_i) - \sum_i P(x_i | c^-)P^*(e^+ | x_i)$
  - $P^*(e^+ | x_i) = \frac{p(e^+ | x_i, c^+)+p(e^+ | x_i, c^-)}{2}$
  - $X$ is an explanatory attribute and $x_i$ is its $i$-th domain value

- Žliobaite et al. propose local massaging and local preferential sampling to removal bad discrimination

Removing Disparate Impact

• Modify the distribution of $X$ so that $C$ is not predictable from $X$.

Learning Fair Representation

• Find a good representation of the data
  – Encode the data as well as possible
  – Obfuscate the sensitive information

\[
X \xrightarrow{\text{Encode}} Z \xrightarrow{\text{Obfuscate}} Y = f(Z)
\]

Fair & Good

• Minimize the objective function
  \[
  L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y
  \]
  – \(L_z\) captures the statistical parity of the representation.
  – \(L_x\) constrains the re-construction error.
  – \(L_y\) requires the accurate prediction.

\(A_x, A_y, A_z\) are hyper-parameters

Fair Representations

• Learn fair representations for prediction task
  – Learn representations of data via auto-encoder.
  – An adversary tries to recover a sensitive attribute $C$ from the representation. The encoder tries to make $C$ impossible to recover.
  – As a result, the prediction based on the fair representations does not depend on sensitive attribute $C$.

• Loss function

\[
\min_{\theta, \eta, \phi} \max_{\phi} L = \alpha C_\theta (X, R) + \beta B_\eta (E, R) + \gamma D_{\theta, \phi} (C, R)
\]

Fair Representations

• Learn fair representations for prediction task
  – Learn representations of data via auto-encoder.
  – An adversary tries to recover a sensitive attribute $C$ from the representation. The encoder tries to make $C$ impossible to recover.
  – As a result, the prediction based on the fair representations does not depend on sensitive attribute $C$.

• Loss function

$$\min_{\theta, \eta} \max_{\phi} L = \alpha C_{\theta}(X, R) + \beta B_{\eta}(E, R) + \gamma \mathcal{D}_{\theta, \phi}(C, R)$$

Auto-encoder loss: $C_{\theta}(X, R) = \|X - \text{Dec}(R)\|_2^2$

Learn useful data representations
Fair Representations

• Learn fair representations for prediction task
  – Learn representations of data via auto-encoder.
  – An adversary tries to recover a sensitive attribute $C$ from the representation. The encoder tries to make $C$ impossible to recover.
  – As a result, the prediction based on the fair representations does not depend on sensitive attribute $C$.

• Loss function

\[
\min_{\theta, \eta} \max_{\phi} L = \alpha C_\theta(X, R) + \beta B_\eta(E, R) + \gamma D_{\theta, \phi}(C, R)
\]

Classification loss: $B_\eta(E, R) = -E \cdot \log(Pred(R)) - (1 - E) \cdot \log(1 - Pred(R))$

Train prediction model
Fair Representations

• Loss function

\[
\min_{\theta, \eta} \max_{\phi} L = \alpha C_\theta(X, R) + \beta B_\eta(E, R) + \gamma D_{\theta, \phi}(C, R)
\]

Adversarial training to remove the information of sensitive attribute

- Risk difference (Edwards, et al. 2016)

\[
\min_{\theta} \max_{\phi} D_{\theta, \phi}(C, R) = C \cdot \log(\text{Adv}(R)) + (1 - C) \cdot \log(1 - \text{Adv}(R))
\]

- Equalized odds (Madras, et al. 2018)

\[
\min_{\theta} \max_{\phi} D_{\theta, \phi}(C, R) = 2 - \sum_{(i, j) \in \{0, 1\}^2} \frac{1}{|V_i|} \sum_{(c, x) \in V_i^j} |\text{Adv}(R) - R|
\]

where \( V_i^j = \{(c, x, e) \in V | c = i, e = j\} \)
Discrimination Prevention

• Algorithm tweak (In-processing)
Fair Regularization for Classification

- Objective functions:

\[-\ell(D; \Theta) + \eta R(D, \Theta) + \frac{\lambda}{2} \|\Theta\|_2^2\]

- Define the discrimination regularization term using mutual information between \(Y\) and \(S\):

\[\text{PI} = \sum_{Y,S} \hat{P}_r[Y,S] \ln \frac{\hat{P}_r[Y,S]}{\hat{P}_r[S]\hat{P}_r[Y]}\]

\[= \sum_{Y,X,S} M[Y|X,S; \Theta]\hat{P}_r[X,S] \ln \frac{\hat{P}_r[Y,S]}{\hat{P}_r[S]\hat{P}_r[Y]}\]

Fairness Constraints for Classification

• Classification fairness is measured using risk ratio
  – Classifier $f(x)$ is learned by minimizing a loss function $L(\theta)$.
  – $f(x_i) = 1$ if $d_\theta(x_i) \geq 0$ and $f(x_i) = -1$ otherwise.

• Use the covariance as the measure of fairness.

\[
\text{Cov}(c, d_\theta(x)) = E[(c - \bar{c})d_\theta(x)] - E[(c - \bar{c})] \bar{d}_\theta(x) \approx \frac{1}{N} \sum_{i=1}^{N} (c_i - \bar{c})d_\theta(x_i)
\]

• Two formulations

| Minimize $L(\theta)$ | Minimize $\left| \frac{1}{N} \sum_{i=1}^{N} (c_i - \bar{c})d_\theta(x_i) \right|$ |
|----------------------|---------------------------------------------------------------|
| Subject to
\[
\frac{1}{N} \sum_{i=1}^{N} (c_i - \bar{c})d_\theta(x_i) \leq \tau
\]
| Subject to $L(\theta) \leq (1 + \gamma)L(\theta^*)$ |

Discrimination Prevention

• Prediction changing (Post-processing)
Post-processing: Manipulation

• Some in-processing techniques work for post-processing
  – Massaging
  – Uniform/preferential sampling

Massaging

• Flip the decision of some individuals according to a ranker
  1. Learn a classifier and estimate the predicted probability of the positive decision of each individual
  2. sort the individuals of four groups according to this probability
  3. Flip the decision of individuals that close to the bottom/top

\[
\begin{align*}
\text{probability: } & c^+ e^- & c^- e^- & c^+ e^+ & c^- e^+ \\
\text{RD: } & \frac{6}{10} - \frac{4}{10} = 0.2 \\
\text{probability: } & c^+ e^- & c^- e^- & c^+ e^+ & c^- e^+ \\
\text{RD: } & \frac{5}{10} - \frac{5}{10} = 0.0
\end{align*}
\]

Kamiran, F., Calders, T.: Classifying without discriminating. IC4’09 (2009)

Preferential Sampling

• Partition the data into 4 groups \((c^+ e^+, c^- e^-, c^- e^+, c^+ e^-)\) and two are under-sampled and two over-sampled
• Select and remove/duplicate the individuals close to the top/bottom

\[
\begin{align*}
\text{Probability: } & c^+ e^- & c^+ e^+ \\
\text{RD: } & \frac{12}{20} - \frac{0}{20} = 0.15 \\
\text{Probability: } & c^- e^- & c^- e^+ \\
\text{RD: } & \frac{10}{20} - \frac{10}{20} = 0
\end{align*}
\]

Decision Theory for Discrimination-aware Classification

• Hypothesis: discrimination decisions are made close to the decision boundary
  
  – Reject Option based Classification (ROC)
    • For probabilistic classifiers, \( P(\hat{y} = + | x) = 0.5 \Rightarrow \text{discrimination.} \)
    • Define \( \max[P(\hat{y} = + | x) - P(\hat{y} = - | x)] < \theta \) as the critical region.
    • Relabel the prediction of individuals in the critical regions.
  
  – Discrimination-Aware Ensemble (DAE)
    • For ensemble methods, larger disagreement of classifiers \( \Rightarrow \text{discrimination} \)
    • Define disagreement
      \[
      \text{disagr}_D = \frac{\left| \{X_i \mid \exists (j, k) \ F_j(X_i) \neq F_k(X_i) \} \right|}{\left| \{X_i \} \right|}
      \]
    • Relabel the prediction of individuals with large disagreement

Construct Equalized Odds Predictor

• Derive a non-discriminatory predictor $\tilde{Y}$ from a learned predictor $\hat{Y}$ by flipping the prediction:

$$p_{ya} = P(\tilde{Y} = 1 \mid \hat{Y} = y, A = a)$$

- These four parameters, $p = (p_{00}, p_{01}, p_{10}, p_{11})$, together specify the derived predictor $\tilde{Y}_p$.

• Finding the optimal, non-discriminatory predictor $\tilde{Y}_p$ is a linear optimization problem:

$$\min_p \mathbb{E}_\ell(\tilde{Y}_p, Y)$$

s.t. $\gamma_0(\tilde{Y}_p) = \gamma_1(\tilde{Y}_p)$

$\forall y, a \ 0 \leq p_{ya} \leq 1$

- Make $\tilde{Y}_p$ close to $Y$
- Ensure equalized odds
- Ensure that $\tilde{Y}_p$ is derived from $\hat{Y}$

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
  – Measures
  – Algorithms
  – From Correlation to Causation
• Part III: Causal Modeling Background
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Correlation vs. Causation

- Correlation means two variables are related but does not tell why.
- A strong correlation does not necessarily mean that changes in one variable causes changes in the other.
- $X$ and $Y$ are correlated
  - $X$ causes $Y$ or $Y$ causes $X$
  - $X$ and $Y$ are caused by a third variable $Z$

- In order to imply causation, a true experiment must be performed where subjects are randomly assigned to different conditions.
Gap Between Association and Causation

• Association does not mean causation, but discrimination is causal.
  – whether an individual would receive the same decision had the individual been of a different race (sex, age, religion, etc.)

• Knowledge about relationships between all attributes should be taken into consideration.

• The golden rule of causal analysis: no causal claim can be established by a purely statistical method.
  – Need causal-aware methods in discovering and preventing discrimination.
Causal based Discrimination Discovery

- Preliminary
- Causal Modelling
- Path-specific
- Counterfactual
Causal based Discrimination Discovery

• Preliminary work

• Causal-modeling-based
Causal based Discrimination Discovery

- **Path-specific-effect-based**
Causal based Discrimination Discovery

• Counterfactual-based
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Techniques in Causal Modeling

• Causal model and causal graph
  – Markovian model and semi-Markovian model
  – Conditional independence and $d$-separation

• Causal inference
  – Intervention and $do$-operator
  – Path-specific effect
  – Counterfactual analysis

How to construct causal graph is omitted.
Lessons of Causal Inference (*Pearl*)

1. No cause in — no cause out
2. Data
   \[ \text{Causal assumptions/knowledge} \]
   \[ \Rightarrow \text{causal conclusions} \]
4. Need ways of encoding causal assumptions/knowledge mathematically and test their implications.
From Statistics to Causal Modeling

- Traditional statistical inference paradigm:

  \[ Q(D) = P_D(E = 'A' | H = 1) \]

- What is the probability of getting Grade A for the students who study 1 hour each day?

  Estimate: \[ Q(D) = P_D(E = 'A' | H = 1) \]

  - \( E \) (Exam Grade)
  - \( H \) (Hour of Study)
  - \( I \) (Interest)
  - \( W \) (Working Strategy)
From Statistics to Causal Modeling

• What is the probability of getting Grade A if a new policy requires all students to study 2 hours each day?
  – The question cannot be solved by statistics.

Estimate \( Q(D') = P_{D'}(E = 'A') \)

\( D' \) represents the joint distribution after adopting the new policy.
What is the probability of getting Grade A if a new policy requires all students to study 2 hours each day?

- The question cannot be solved by statistics.

\[ P_{D'}(E = 'A') \neq P_D(E = 'A' | H = 2) \]

The probability of getting Grade A of the students who study 2 hours each day at the first place.
From Statistics to Causal Modeling

- Causal inference

\[ M \rightarrow \text{Joint Distribution} \rightarrow D \]

\[ D \rightarrow \text{Data Generating Model} \rightarrow M \rightarrow Q(M) \]

\( M \) – Data generation model that encodes the causal assumptions/knowledge.

\( D \) – model of data, \( M \) – model of reality
From Statistics to Causal Modeling

- Causal inference

\[ D \rightarrow \text{Joint Distribution} \rightarrow D' \]

\[ \text{Data Generating Model} \rightarrow M \]

\[ \text{Data Generating Model} \rightarrow M' \]

\[ Q(M') \]

\[ \text{change} \]
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Structural Causal Model

• A theory of inferred causation.
• Describe how causal relationships can be inferred from nontemporal statistical data if one makes certain assumptions about the underlying process of data generation.
• Developed since 1988, still growing at an increasing speed.
Structural Causal Model

• A causal model is triple $\mathcal{M} = < U, V, F >$, where
  – $U$ is a set of exogenous (hidden) variables whose values are determined by factors outside the model;
  – $V = \{X_1, \ldots, X_i, \ldots\}$ is a set of endogenous (observed) variables whose values are determined by factors within the model;
  – $F = \{f_1, \ldots, f_i, \ldots\}$ is a set of deterministic functions where each $f_i$ is a mapping from $U \times (V \setminus X_i)$ to $X_i$. Symbolically, $f_i$ can be written as
    \[
    x_i = f_i(pa_i, u_i)
    \]
    where $pa_i$ is a realization of $X_i$’s parents in $V$, i.e., $Pa_i \subseteq V$, and $u_i$ is a realization of $X_i$’s parents in $U$, i.e., $U_i \subseteq U$. 

Causal Graph

• Each causal model $\mathcal{M}$ is associated with a direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where
  – $\mathcal{V}$ is the set of nodes representing the variables $U \cup V$ in $\mathcal{M}$;
  – $\mathcal{E}$ is the set of edges determined by the structural equations in $\mathcal{M}$: for $X_i$, there is an edge pointing from each of its parents $Pa_i \cup U_i$ to it.
    • Each direct edge represents the potential direct causal relationship.
    • Absence of direct edge represents zero direct causal relationship.

• Assuming the acyclicity of causality, $\mathcal{G}$ is a directed acyclic graph (DAG).

• Standard terminology
  – parent, child, ancestor, descendent, path, direct path
A Causal Model and Its Graph

Observed Variables $V = \{I, H, W, E\}$

Hidden Variables $U = \{U_I, U_H, U_W, U_E\}$

Model ($M$)

\[ i = f_I(u_I) \]
\[ h = f_H(i, u_H) \]
\[ w = f_W(h, u_W) \]
\[ e = f_E(i, h, w, u_E) \]

Assume $U_I$ and $U_H$ are correlated.
Markovian Model

- A causal model is Markovian if
  1. The causal graph is a DAG;
  2. All variables in $U$ are mutually independent.

Equivalent expression

Each node $X$ is conditionally independent of its non-descendants given its parents $Pa_X$.

Known as the local Markov condition (e.g., in Bayesian network), or causal Markov condition in the context of causal modeling.
A Markovian Model and Its Graph

Model (M)

\[ i = f_I(u_I) \]
\[ h = f_H(i, u_H) \]
\[ w = f_W(h, u_W) \]
\[ e = f_E(i, h, w, u_E) \]

Graph (G)

Assume \( U_I, U_H, U_W, U_E \) are mutually independent.
Causal Graph of Markovian Model

Each node is associated with a observable conditional probability table (CPT) $P(x_i | pa_i)$
Conditional Independence

- We can read off from the causal graph all the conditional independence relationships encoded in the causal model (graph) by using a graphical criterion called $d$-separation.

- Two random variables $X$ and $Y$ are called conditionally independent given $Z$, if for each values of $(X, Y, Z)$, $(x, y, z)$,
  - $P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$
  - Denoted by $X \perp Y|Z$ or $(X \perp Y|Z)_D$
**d-Separation**

- **Definition of d-separation**

  - A path $q$ is said to be blocked by conditioning on a set $Z$ if
    - $q$ contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node $m$ is in $Z$, or
    - $q$ contains a collider $i \rightarrow m \leftarrow j$ such that the middle node $m$ is not in $Z$ and such that no descendant of $m$ is in $Z$.

- $Z$ is said to $d$-separate $X$ and $Y$ if $Z$ blocks every path from $X$ to $Y$, denoted by $(X \perp Y|Z)_G$
**d-Separation**

- **Example (blocking of paths)**
  
  ![Diagram](image)

  - Path from $X$ to $Y$ is blocked by conditioning on $\{U\}$ or $\{Z\}$ or both $\{U, Z\}$

- **Example (unblocking of paths)**
  
  ![Diagram](image)

  - Path from $X$ to $Y$ is blocked by $\emptyset$ or $\{U\}$
  - Unblocked by conditioning on $\{Z\}$ or $\{W\}$ or both $\{Z, W\}$
\section*{d-Separation}

\begin{itemize}
  \item Example (\textit{d}-separation)
    \begin{center}
      \begin{tikzpicture}
        \node[shape=circle] (x) at (0,0) {$X$};
        \node[shape=circle] (z) at (1,0) {$Z$};
        \node[shape=circle] (u) at (2,0) {$U$};
        \node[shape=circle] (y) at (3,0) {$Y$};
        \node[shape=circle] (v) at (1,-1) {$V$};
        \node[shape=circle] (w) at (2,-1) {$W$};
        \draw[->] (x) to (z);
        \draw[->] (z) to (u);
        \draw[->] (u) to (y);
        \draw[->] (z) to (v);
        \draw[->] (u) to (w);
        \end{tikzpicture}
    \end{center}
  \item We have following \textit{d}-separation relations
    \begin{itemize}
      \item $(X \perp Y|Z)_G, (X \perp Y|U)_G, (X \perp Y|ZW)_G$
      \item $(X \perp Y|ZW)_G, (X \perp Y|UW)_G, (X \perp Y|ZUW)_G$
      \item $(X \perp Y|ZUW)_G$
    \end{itemize}
  \item However we do NOT have
    \begin{itemize}
      \item $(X \perp Y|VZU)_G$
    \end{itemize}
\end{itemize}
Factorization Formula

- In a Markovian model, the joint distribution over all attributes can be computed using the factorization formula:

\[ P(\mathbf{v}) = \prod_{x \in \mathbf{V}} P(x|\mathbf{pa}_x) \]

\[ P(i, h, w, e) = P(i)P(h|i)P(w|h)P(e|i, h, w) \]

\[ P(e) = \sum_{I, H, W} P(i)P(h|i)P(w|h)P(e|i, h, w) \]
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference
    • Intervention and \textit{do}-operator
    • Path-specific effect
    • Counterfactual analysis
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Statistical Inference

• What is the probability of getting grade A if we see that the study hour is 1?

• Find $P(E = 'A' | H = 1)$
Causal Inference

- What is the probability of getting grade A if we change the study hour to 2?
- The above probability does not equal to \( P(E = 'A'|H = 2) \), i.e., the conditional probability of getting grade A given study hour equals to 2.
Intervention and \textit{do}-Operator

• The basic operation of manipulating a causal model.
  – Simulate the manipulation of the physical mechanisms by some physical interventions or hypothetical assumptions.
  – Forces some observed variables $X \in V$ to take certain constants $x$.

• Mathematically formulated as $do(X = x)$ or simply $do(x)$.

• For an observed variable $Y$ disjoint with $X$, its interventional variant under intervention $do(x)$ is denoted by $Y_{X \leftarrow x}$ or $Y_x$.

• The \textbf{effect of intervention} on all other observed variables $Y = V \setminus X$ is represented by the \textit{post-intervention distribution} of $Y$.
  – Denoted by $P(Y = y | do(X = x))$ or simply $P(y | do(x))$;
  – Or equivalently $P(Y_{X \leftarrow x} = y)$ or simply $P(y_x)$. 
Intervention and \textit{do}-Operator

• In causal model $\mathcal{M}$, intervention $\textit{do}(x^*)$ is defined as the substitution of structural equation $x = f_X(pa_X, u_X)$ with value $x^*$. The causal model after performing $\textit{do}(x^*)$ is denoted by $\mathcal{M}_{x^*}$.

\[
\begin{align*}
\mathcal{M} : & \quad x = f_X(pa_X, u_X) & \xrightarrow{\textit{do}(x^*)} & \mathcal{M}_{x^*} : & \quad x = x^*
\end{align*}
\]

• From the point of view of the causal graph, performing $\textit{do}(x^*)$ is equivalent to setting the node $X$ to value $x^*$ and removing all the incoming edges in $X$. 
Intervention in Markovian Model

• In the Markovian model, the post-intervention distribution $P(y|do(x))$ can be calculated from the CPTs, known as the truncated factorization:

$$P(y|do(x)) = \prod_{Y \in Y} P(y|Pa_Y) \delta_{X \leftarrow x}$$

– where $\delta_{X \leftarrow x}$ means assigning attributes in $X$ involved in the term ahead with the corresponding values in $x$.

• Specifically, for a single attribute $Y$ given an intervention on a single attribute $X$,

$$P(y|do(x)) = \sum_{V \setminus \{X,Y\}} \prod_{V \in V \setminus \{X\}} P(v|Pa_V) \delta_{X \leftarrow x}$$
Intervention Example

• What is the probability of getting grade A if we change the study hour to 2?

Graph \((G)\)

\[
\begin{align*}
I \text{ (Interest)} & \\
H \text{ (Hour of Study)} & \\
W \text{ (Working Strategy)} & \\
E \text{ (Exam Grade)} & 
\end{align*}
\]

Model \((M)\)

\[
\begin{align*}
i &= f_I(u_I) \\
h &= f_H(i, u_H) \\
w &= f_W(h, u_W) \\
e &= f_E(i, h, w, u_E)
\end{align*}
\]
• What is the probability of getting grade A if we change the study hour to 2, i.e., $do(H = 2)$?

• Find $P(E = 'A'|do(H = 2))$
Intervention Example

Graph ($G'$)

- $I$ (Interest)
- $W$ (Working Strategy)
- $E$ (Exam Grade)

$do(H = 2)$
(Hour of Study)

Model ($M'$)

$$i = f_i(u_I)$$
$$h = 2$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, k, u_E)$$

$$P(y|do(x)) = \sum_{V \setminus \{x,Y\}} \prod_{v \in V \setminus \{x\}} P(v|Pa_v)\delta_{X=x}$$

$$P(E = 'A'|do(H = 2)) = \sum_{I,W} P(i)P(w|H = 2)P(E = 'A'|i, H = 2, w)$$
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference
    • Intervention and do-operator
    • Path-specific effect
    • Counterfactual analysis
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
**Total Causal Effect**

- The total causal effect of $X$ on $Y$ is given by

$$TE(x_2, x_1) = P(y|do(x_2)) - P(y|do(x_1))$$

- Measures the causal effect transmitted along all causal paths from $X$ to $Y$.

$$P(e | do(H = 2))$$

$$P(w | H = 2)$$

\[
TE(H = 2, H = 1) = P(E = 'A'|do(H = 2)) - P(E = 'A'|do(H = 1)) = \sum_{i,W} P(i)P(w|H = 2)P(E = 'A'|i, H = 2, w) - \sum_{i,W} P(i)P(w|H = 1)P(E = 'A'|i, H = 1, w)
\]
• **Path-specific effect** measures the causal effect transmitted along certain paths.

• Given a subset of causal paths $\pi$, the causal effect of $X$ on $Y$ transmitted along $\pi$ is denoted by

$$SE_\pi(x_2, x_1) = P(y|do(x_2|\pi)) - P(y|do(x_1))$$

- $P(y|do(x_2|\pi))$ denotes the distribution of $Y$ after an intervention of changing $X$ from $x_1$ to $x_2$ with the effect transmitted along $\pi$. 
Path-Specific Effect

- The causal effect of Study Hour on Exam Grade while keeping the Working Strategy **unchanged**.
- Measures the causal effect of $H$ on $E$ transmitted along the direct edge ($\pi$).

$$P(e | do(H = 2|\pi))$$

$$SE_\pi(H = 2, H = 1) = P(E = A | do(H = 2|\pi)) - P(E = A | do(H = 1))$$

$$= \sum_{I,W} P(i)P(w | H = 2)P(E = A | i, H = 1, w) - \sum_{I,W} P(i)P(w | H = 1)P(E = A | i, H = 1, w)$$
Path-Specific Effect

- **Identifiability**: The path-specific effect can be computed from the observational data if and only if the recanting witness criterion is not satisfied.

- **Recanting witness criterion**:

- Refer to (Avin et al., 2005).

Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
  – From Statistics to Causal Modeling
  – Structural Causal Model and Causal Graph
  – Causal Inference
    • Intervention and do-operator
    • Path-specific effect
    • Counterfactual analysis
• Part IV: Causal Modeling based Anti-Discrimination Learning
• Part V: Challenges and Directions for Future Research
Counterfactual Analysis

• Counterfactual analysis deals with interventions while we also have certain observations, or evidence $e$.

• General form of a counterfactual query: “what would we expect the value of $Y$ had $X$ been $x$, given that we observe $E = e$?”

$$P(Y_{X\leftarrow x} = y \mid E = e) \text{ or } P(y_x \mid e)$$

• Example: Whether “gender is male” is the necessary and sufficient condition for “being hired”?
  
  – Probability of necessity: $P(H_{G\leftarrow f} = n \mid G = m, H = y)$
  – Probability of sufficiency: $P(H_{G\leftarrow m} = y \mid G = f, H = n)$
  – Probability of necessity and sufficiency: $P(H_{G\leftarrow m} = y, H_{G\leftarrow f} = n)$
Counterfactual Analysis

- Counterfactual $P(y_x | e)$ considers both the actual world $\mathcal{M}$, and the counterfactual world $\mathcal{M}_x$.
- Two worlds share background before the intervention.
- Example: $P(y'_x, | x, y)$

\[
\begin{align*}
\mathcal{M} & : x = f_X(u_X) \\
& : y = f_Y(x, u_Y)
\end{align*}
\]

\[
\begin{align*}
\mathcal{M}_x & : x = x' \\
& : y = f_Y(x, u_Y)
\end{align*}
\]

Counterfactual graph: depicts together the actual world and counterfactual worlds invoked by the counterfactual query.
Intervention vs. Counterfactual

**Intervention:**
Questions regarding a *single* world

**Counterfactual:**
Questions regarding *multiple* worlds

\[
P(y_x) \quad \text{Actual World} \quad \text{Counterfactual World}
\]

\[
P(y_x | e) \quad \text{Actual World} \quad \text{Counterfactual World}
\]
Counterfactual Analysis

\[ P(y_x \mid e) = \sum_u P(y_x \mid e, u) P(u \mid e) = \sum_u P(y_x \mid u) P(u \mid e) \]

• Principled procedure for computing \( P(y_x \mid e) \):
  – **Abduction**: Update \( P(u) \) by the evidence \( e \) to obtain \( P(u|e) \);
  – **Action**: Perform intervention \( do(x) \) on causal model \( \mathcal{M} \) to obtain \( \mathcal{M}_x \);
  – **Prediction**: Compute the probability of \( Y = y \) using \( \mathcal{M}_x \) and \( P(u|e) \).

• Usually don’t know \( P(u) \).
Identifiability of Counterfactual

- May be non-identifiable without complete knowledge of causal model (structure equations and \( P(\mathbf{u}) \)), even in Markovian model.

- “W-graph”: the simplest non-identifiable counterfactual graph structure.

\[
P(y'_{x'} | x, y) \text{ is non-identifiable for any causal model}
\]

Identifiability of Counterfactual

- Complete identification algorithm: ID* (Shpitser et al., 2008)
- Possible to be identifiable under certain assumptions.
  - Example: In linear Gaussian models, $\mathbb{E}[y_x \mid e]$ is identifiable for any $Y, X, E$, given by (Pearl et al., 2017)

\[
\mathbb{E}[Y_x \mid e] = \mathbb{E}[Y \mid e] + \tau(x - \mathbb{E}[X \mid e])
\]

where $\tau = \frac{\partial}{\partial x} \mathbb{E}[Y \mid do(x)]$


Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part V: Challenges and Directions for Future Research
Main Ideas

• Use **causal model** and **causal graph** to capture the causal structure of the data.

• Employ **do-operator** to simulate the intervention of changing an individual from protected group to non-protected group and vice versa.

• Adopt **path-specific effect** technique to identify direct/indirect discrimination as the causal effects transmitted along different paths in the causal graph.

• Utilize **counterfactual** to measure discrimination in sub-groups and for individuals.
### Causal Model

**Observed Variables** \( V = \{C, \cdots, X_i, \cdots, E\} \)

**Hidden Variables** \( U \)

**Causal Model** \( M \)

\[
\begin{align*}
    c &= f_C(p_aC, u_C) \\
    x_i &= f_i(p_a_i, u_i), \ i = 1, \cdots, m \\
    e &= f_E(p_a_E, u_E)
\end{align*}
\]

**Observed Variables**  
**Hidden Variables**

\( U_C, \cdots, U_i, \cdots, U_E \) are mutually independent  
(Markovian Assumption)

**Causal Graph** \( G \)

\[
\begin{align*}
    P(x_i|p_a_i) \quad C & \rightarrow X_i \\
    P(c|p_aC) \quad C & \rightarrow C \\
    P(e|p_aE) \quad E & \rightarrow E \\
    \cdots \quad R \quad \cdots
\end{align*}
\]
Motivating Examples (ME1)

- How to deal with indirect discrimination due to redlining attributes?
- Assume a bank makes loan decisions based on the areas of residence of the applicants.
Motivating Examples (ME2)

- How to answer “what if” questions?
  - E.g., a female applicant is rejected when applying for a job. What if the applicant is a male?

- Refer to as the counterfactual question, since it asks about the result NOT in the actual world but in a counterfactual world.
  - Results in the counterfactual world cannot be observed in any way.
Motivating Examples (ME3)

- Data discrimination-free vs. Model discrimination-free

- Assumption: a classifier learned from a discrimination-free training data will also be discrimination-free.
- Whether and to what extent this assumption holds?
Motivating Examples (ME4)

• How to ensure non-discrimination in data release under all possible scenarios?
• How to identify meaningful partitions?

<table>
<thead>
<tr>
<th>gender</th>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>admission (%)</td>
<td>37%</td>
<td>47%</td>
</tr>
</tbody>
</table>

\[ P(e^+|c^+) - P(e^+|c^-) = 0.1 \]

<table>
<thead>
<tr>
<th>major</th>
<th>CS</th>
<th>EE</th>
</tr>
</thead>
<tbody>
<tr>
<td>test score</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>gender</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>admission (%)</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

\[ P(e^+|c^+, \{CS,L\}) - P(e^+|c^- , \{CS,L\}) = 0 \]

<table>
<thead>
<tr>
<th>gender</th>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>admission (%)</td>
<td>43%</td>
<td>43%</td>
</tr>
</tbody>
</table>

\[ P(e^+|c^+) - P(e^+|c^-) = 0 \]

<table>
<thead>
<tr>
<th>major</th>
<th>CS</th>
<th>EE</th>
</tr>
</thead>
<tbody>
<tr>
<td>test score</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>gender</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>admission (%)</td>
<td>30%</td>
<td>36%</td>
</tr>
</tbody>
</table>

\[ P(e^+|c^+, \{CS,L\}) - P(e^+|c^- , \{CS,L\}) = 0.06 \]
Motivating Examples (ME5)

- How to find paired individuals for situation testing in individual discrimination?

<table>
<thead>
<tr>
<th>No.</th>
<th>gender</th>
<th>major</th>
<th>score</th>
<th>height</th>
<th>weight</th>
<th>ad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>CS</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>CS</td>
<td>B</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>CS</td>
<td>A</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>CS</td>
<td>A</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>CS</td>
<td>C</td>
<td>low</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>CS</td>
<td>C</td>
<td>median</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>EE</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
</tbody>
</table>

- Which one is closest to 1? 2 or 3 or 7?
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part V: Challenges and Directions for Future Research
Direct and Indirect Discrimination

- **Direct**: explicitly based on the protected attribute $C$.
  - E.g., rejecting a qualified female just because of her gender.

- **Indirect**: based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group.
  - E.g., redlining, where the residential Zip Code of an individual is used for making decisions such as granting a loan.
  - Redlining attributes $R$: non-protected attributes that can cause indirect discrimination.
Direct and Indirect Discrimination Discovery and Removal

• How to deal with indirect discrimination due to redlining attributes?

• Modeling direct and indirect discrimination using the causal model.

• Quantitative discrimination measure and criterion.

• Algorithm for removing direct and indirect discrimination from a dataset.

Direct and indirect discrimination can be captured by the causal effects of $C$ on $E$ transmitted along different paths.

- Direct discrimination: causal effect along direct edge from $C$ to $E$.
  - Denoted by $SE_{\pi_d}(c^+, c^-)$ where $\pi_d$ is the path $C \rightarrow E$.

- Indirect discrimination: causal effect along causal paths that pass through redlining attributes.
  - Denoted by $SE_{\pi_i}(c^+, c^-)$ where $\pi_i$ contains all the causal paths from $C$ to $E$ through redlining attributes $R$. 

![Diagram](image.png)
Quantitative Measuring

• $\pi_d$-specific effect:

$$SE_{\pi_d}(c^+, c^-) = \sum_q (P(e^+|c^+, q)P(q|c^-)) - P(e^+|c^-)$$

$Q$: $E$’s parents except $C$

• $\pi_i$-specific effect:

$$SE_{\pi_i}(c^+, c^-) = \sum_{v'} \left( P(e^+|c^-, q) \prod_{G \in S_{\pi_i}} P(g|c^+, pa_G \backslash \{C\}) \right) \prod_{H \in S_{\pi_i} \backslash \{E\}} P(h|c^-, pa_H \backslash \{C\}) \prod_{o \in V \backslash Ch_C} P(o|pa_o)) - P(e^+|c^-)$$

$S_{\pi_i}$: $C$’s children that lie on paths in $\pi_i$

$\bar{S}_{\pi_i}$: $C$’s children that don’t lie on paths in $\pi_i$
Illustrative Example

- A bank makes loan decisions based on the Zip Codes, races, and income of the applicants.
  - Race: protected attribute
  - Loan: decision
  - Zip Code: redlining attribute
  - Income: non-protected attribute

- Race: protected attribute
- Loan: decision
- Zip Code: redlining attribute
- Income: non-protected attribute
Illustrative Example

\[
SE_{\pi_d}(c^+, c^-) = \sum_{Z,I} \left( P(e^+|c^+, z, i) - P(e^+|c^-, z, i) \right) P(z|c^-) P(i|c^-)
\]

\[
SE_{\pi_i}(c^+, c^-) = \sum_{Z,I} P(e^+|c^-, z, i) \left( P(z|c^+) - P(z|c^-) \right) P(i|c^-)
\]
Causal Effect vs. Risk Difference

- The total causal effect of $C$ (changing from $c^-$ to $c^+$) on $E$ is given by

$$TE(c^+, c^-) = P(e^+|do(c^+)) - P(e^+|do(c^-))$$

- transmitted along all causal paths from $C$ to $E$.

- Connection with the risk difference

$$TE(c^+, c^-) = P(e^+_c) - P(e^+_c|c^-)$$
Total Causal Effect vs. Path-Specific Effect

- For any $\pi_d$ and $\pi_i$, we don’t necessarily have

$$SE_{\pi_d}(c^+, c^-) + SE_{\pi_i}(c^+, c^-) = SE_{\pi_d \cup \pi_i}(c^+, c^-)$$

- If $\pi_i$ contains all causal paths from $C$ to $E$ except $\pi_d$, then

$$TE(c^+, c^-) = SE_{\pi_d}(c^+, c^-) - SE_{\pi_i}(c^-, c^+)$$

“reverse” $\pi_i$-specific effect
Discrimination Discovery and Removal Algorithms

- **Path-Specific Effect based Discrimination Discovery (PSE-DD) algorithm**
  - Build causal graph
  - Compute $SE_{\pi_d}$ and $SE_{\pi_i}$

- **Path-Specific Effect based Discrimination Removal (PSE-DR) algorithm**
  - Modify the CPT of $E$ so that no discrimination exists.
  - Generate a new dataset using the modified graph.
  - Minimize the distance of the joint distributions: quadratic programming.

\[
\begin{align*}
\text{minimize} & \quad \sum_V \left( P'(v) - P(v) \right)^2 \\
\text{subject to} & \quad SE_{\pi_d}(c^+, c^-) \leq \tau, \quad SE_{\pi_d}(c^-, c^+) \leq \tau, \\
& \quad SE_{\pi_i}(c^+, c^-) \leq \tau, \quad SE_{\pi_i}(c^-, c^+) \leq \tau, \\
& \quad \forall Pa(E), \quad P'(e^-|Pa(E)) + P'(e^+|Pa(E)) = 1, \\
& \quad \forall Pa(E), e, \quad Pr'(e|Pa(E)) \geq 0,
\end{align*}
\]
Empirical Evaluation

- Data: Adult dataset

  protected attribute: sex  
decision: income  
redlining attribute: marital_status

Tool: TETRAD for building the causal graph (using the classic PC algorithm)

Threshold = 0.05

\[ SE_{\pi_d}(c^+, c^-) = 0.025 \]
\[ SE_{\pi_i}(c^+, c^-) = 0.175 \]

Correlation-based methods cannot correctly identify either direct or indirect discrimination.
Comparison of Removal Methods

- Evaluated algorithms:
  - PSE-DR (Zhang et al. IJCAI 2017)
  - Local massaging (LMSG) and local preferential sampling (LPS) algorithms (Žliobaite et al. ICDM 2011)
  - Disparate impact removal algorithm (DI) (Feldman et al. KDD 2015)

- Local massaging (LMSG) and local preferential sampling (LPS) algorithms still have discrimination.

- Disparate impact removal algorithm (DI) incurs more utility loss.

<table>
<thead>
<tr>
<th>Remove Algorithm</th>
<th>PSE-DR</th>
<th>DI</th>
<th>LMSG</th>
<th>LPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>0.013</td>
<td>0.001</td>
<td>-0.142</td>
<td>-0.142</td>
</tr>
<tr>
<td>Indirect</td>
<td>0.049</td>
<td>0.050</td>
<td>0.288</td>
<td>0.174</td>
</tr>
<tr>
<td>$\chi^2 (\times 10^4)$</td>
<td>1.038</td>
<td>4.964</td>
<td>1.924</td>
<td>1.292</td>
</tr>
</tbody>
</table>
Fair Inference on Outcomes

• Infer a fair distribution \( P^*(C, X, E) \) from a sample \( D \) drawn from the original distribution \( P(C, X, E) \).

• Approximate \( P^*(C, X, E) \) by solving a constrained maximum likelihood problem using path-specific effects

\[
\hat{\alpha} = \operatorname{argmax}_\alpha L_{C,X,E}(D; \alpha) \quad \text{Subject to } \epsilon_l \leq g(D) \leq \epsilon_u
\]

- \( D \): finite samples drawn from \( P(C, X, E) \)
- \( L_{C,X,E}(D; \alpha) \): likelihood function parameterized by \( \alpha \)
- \( g(D) \): estimator of the path-specific effect

Variants of Indirect Discrimination

• Two definitions of **indirect** discrimination:
  – Unresolved discrimination: if there exits a directed path from $C$ to $E$ that is not blocked by a resolving variable (explainable variable).
  – Potential proxy discrimination: if there exists a directed path from $C$ to $E$ that is blocked by a proxy variable $R$ (redlining variable).
    • No proxy discrimination if $P(E \mid do(R = r)) = P(E \mid do(R = r'))$

• Pros:
  – Use intervention rather than path-specific effect to define indirect discrimination, avoid non-identifiability issue.

• Cons:
  – Can only qualitatively determine the existence of the discrimination, but cannot quantitatively measure the amount of discriminatory effects as the path-specific effects do

Variants of Indirect Discrimination

• Develop procedures for avoiding discrimination in the predictor under linearity assumptions about the causal model.

• Example:

\[
P = \alpha_P A + N_P \\
X = \alpha_X A + \beta P + N_X \\
R_\theta = \lambda_P P + \lambda_X X
\]

Result: any predictor of the form 
\[
R_\theta = \lambda_X (X - \beta P)
\]
with free parameter \(\lambda_X\) exhibits no proxy discrimination.

\[
E = \alpha_E A + N_E \\
X = \alpha_X A + \beta E + N_X \\
R_\theta = \lambda_A A + \lambda_P P + \lambda_X X
\]

Result: any predictor of the form 
\[
R_\theta = \lambda_X (X - \alpha_X A) + \lambda_E E
\]
with free parameters \(\lambda_X, \lambda_E\) exhibits no unresolved discrimination.
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part V: Challenges and Directions for Future Research
Counterfactual Fairness

- Protected attribute: $C \rightarrow A$
- Non-protected attribute: $X \rightarrow X$
- Decision attribute: $E \rightarrow Y$
- Predictor: $\hat{E} \rightarrow \hat{Y} = f(x, a)$

Counterfactual Fairness

(Russell et al., 2017) For any predictor $\hat{Y} = f(x, a)$:

- Counterfactual fair if $f(x_{A\leftarrow a}, a) = f(x_{A\leftarrow a}, a')$ for any input with $X = x$ and $A = a$;

- $(\epsilon, 0)$-approximate counterfactual fair if $|f(x_{A\leftarrow a}, a) - f(x_{A\leftarrow a}, a')| \leq \epsilon$ for any input with $X = x$ and $A = a$;

- $(\epsilon, \delta)$-approximate counterfactual fair if $P(|f(x_{A\leftarrow a}, a) - f(x_{A\leftarrow a}, a')| \leq \epsilon | x, a) \geq 1 - \delta$

- Also works for a dataset:

  Counterfactual fair if for any context $U = u, X = x$ and $A = a$,
  
  $$P(Y_a(u) = y | X = x, A = a) = P(Y_{a'}(u) = y | X = x, A = a),$$

  for all $y$ and $a'$.

Toy Examples

The Red Car

- $A$ is independent of $Y$;
- $\hat{Y} = f(x)$ does not use $A$ but is not counterfactually fair;
- $\hat{Y} = f(x, a)$ is counterfactual fair if $f(\cdot)$ is a regression.
  - Equivalent to regressing on $U$.

High Crime Regions

- Locations $X$ with more police resources have larger $Y$;
- Not because different races are any more or less likely to break the law;
- Algorithms enforcing EO will not remedy unfairness.
Constructing Counterfactually Fair Predictors

• Lemma: $\hat{Y}$ will be counterfactually fair if it is a function of non-descendants of $A$.

• Three levels of conditions for counterfactually fair predictors:
  1. $\hat{Y}$ is built using only non-descendants of $A$;
  2. $\hat{Y}$ is built on latent variables $U$ whose distribution, i.e., $P(u|x, a)$, is known based on explicit domain knowledge;
  3. $\hat{Y}$ is built on latent variables $U$ where the causal model is postulated, e.g., $x_i = f_i(pa_i) + u_i$ with given types of function $f_i(\cdot)$. 
Dataset: 21,790 law students with their race, sex, entrance exam scores (LSAT), grade-point average (GPA) prior to law school, and first year average grade (FYA).

Counterfactual unfair predictors:
- Full model: which is built on all attributes;
- Unaware model: which is built on attributes other than race and sex;

Counterfactual fair predictors:
- Fair K: which is built on K, a postulated hidden variable whose distribution is estimated from data;
- Fair Add: which assumes an additive causal model, and is built on the error terms of the additive model.
Illustrative Example

• Counterfactual fairness:
  – Both counterfactual fair predictors can achieve fairness;
  – while counterfactual unfair predictors cannot.

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Unaware</th>
<th>Fair $\bar{K}$</th>
<th>Fair Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.873</td>
<td>0.894</td>
<td>0.929</td>
<td>0.918</td>
</tr>
</tbody>
</table>
**Counterfactual Direct/Indirect Discrimination**

- **Protected attribute:** $C \rightarrow X$
- **Non-protected attribute:** $X \rightarrow Z, W$
- **Decision attribute:** $E \rightarrow Y$

**Effect of Treatment On the Treated (ETT):** Effect of intervention $X = x_1$ on $Y = y$ conditioned on $X = x_0$

$$ETT_{x_0,x_1}(y) = P(y_{x_1} \mid x_0) - P(y \mid x_0)$$

The probability of $Y$ would be $y$ had $X$ been $x_1$ (counterfactually), given that in the actual world $X = x_0$.

---

Counterfactual Direct/Indirect Discrimination

- Define discrimination as the direct/indirect ETT.
- Discrimination measures:
  - Counterfactual direct effect (Ctf-DE): Direct effect of intervention $X = x_1$ on $Y$ (with baseline $x_0$) conditioned on $X = x$

$$DE_{x_0,x_1}(y|x) = P\left(y_{x_1,W_{x_0}}|x\right) - P(y_{x_0}|x)$$

The value of $W$ which would have attained had $X$ been $x_0$

The value of $Y$ would be had $X$ been $x_1$, while $W$ is kept at the same value that it would have attained had $X$ been $x_0$, given that $X$ was actually equal to $x$
Counterfactual Direct/Indirect Discrimination

- Discrimination measures:
  - Counterfactual direct effect (Ctf-DE):
    \[ DE_{x_0,x_1}(y|x) = P(y_{x_1,w_{x_0}}|x) - P(y_{x_0}|x) \]
  - Counterfactual indirect effect (Ctf-IE): Indirect effect of intervention \( X = x_1 \) on \( Y \) (with baseline \( x_0 \)) conditioned on \( X = x \)
    \[ IE_{x_0,x_1}(y|x) = P(y_{x_0,w_{x_1}}|x) - P(y_{x_0}|x) \]

The value of \( W \) which would have attained had \( X \) been \( x_1 \)

The value of \( Y \) would be had \( X \) been \( x_0 \), while changing \( W \) to whatever level it would have obtained had \( X \) been \( x_1 \), given that \( X \) was actually equal to \( x \)
Counterfactual Direct/Indirect Discrimination

- Discrimination measures:
  - Counterfactual direct effect (Ctf-DE):
    \[ DE_{x_0,x_1}(y|x) = P\left( y_{x_1,w_{x_0}} \mid x \right) - P(y_{x_0} \mid x) \]
  - Counterfactual indirect effect (Ctf-IE):
    \[ IE_{x_0,x_1}(y|x) = P\left( y_{x_0,w_{x_1}} \mid x \right) - P(y_{x_0} \mid x) \]
  - Counterfactual spurious effect (Ctf-DE) (NOT discrimination): Capture spurious associations between \( X \) and \( Y \)
    \[ SE_{x_0,x_1}(y) = P\left( y_{x_0} \mid x_1 \right) - P(y \mid x_0) \]

The value of \( Y \) would be had \( X \) been \( x_0 \), given that \( X \) was actually equal to \( x_1 \)

The probability difference in \( Y = y \) had \( X \) been \( x_0 \) for the individuals that would naturally choose \( X \) to be \( x_0 \) versus \( x_1 \).
Graphical Properties

1. If $X$ has no direct causal path connecting $Y$ in the causal graph, then $DE_{x_0,x_1}(y|x) = 0$, for any $x, y, x_0 \neq x_1$.

2. If $X$ has no indirect causal path connecting $Y$ in the causal graph, then $IE_{x_0,x_1}(y|x) = 0$, for any $y, x, x_0 \neq x_1$.

3. If $X$ has no back-door path connecting $Y$ in the causal graph, then $SE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$. 
Relationship

- **Show relationships among counterfactual effects**
  
  **Theorem 1** (Causal Explanation Formula). The total variation, counterfactual spurious, direct, and indirect effects obey the following relationships

  \[
  TV_{x_0,x_1}(y) = SE_{x_0,x_1}(y) + IE_{x_0,x_1}(y|x_1) \\
  - DE_{x_1,x_0}(y|x_1) \quad (9) \\
  TV_{x_0,x_1}(y) = DE_{x_0,x_1}(y|x_0) \\
  - SE_{x_1,x_0}(y) - IE_{x_1,x_0}(y|x_0) \quad (10)
  \]

- **Summary of different discrimination measures**

<table>
<thead>
<tr>
<th>Fairness</th>
<th>Measures</th>
<th>Direct</th>
<th>Indirect</th>
<th>Spurious</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
<td>TV</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ETT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>NIE</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Procedure</strong></td>
<td>NDE</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>QII</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>CDE</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>Our Approach</strong></td>
<td>Ctf-DE</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Ctf-IE</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Ctf-SE</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>total variation</td>
<td>( P(y</td>
</tr>
<tr>
<td>TE</td>
<td>total effect</td>
<td>( P(y_{x_1}) - P(y_{x_0}) )</td>
</tr>
<tr>
<td>ETT</td>
<td>effect of the treatment on the treated</td>
<td>( P(y_{x_1}</td>
</tr>
<tr>
<td>NIE</td>
<td>natural indirect effect</td>
<td>( P(y_{x_0}, w_{x_1}) - P(y_{x_0}) )</td>
</tr>
<tr>
<td>NDE</td>
<td>natural direct effect</td>
<td>( P(y_{x_1}, w_{x_0}) - P(y_{x_0}) )</td>
</tr>
<tr>
<td>QII</td>
<td>quantitative input influence</td>
<td>( E[Y] - E[Y_{X \sim P(x), Z \sim P(z), W \sim P(z, w)}] )</td>
</tr>
<tr>
<td>CDE</td>
<td>controlled direct effect</td>
<td>( P(y_{x_1,x}, w) - P(y_{x_0,x}, w) )</td>
</tr>
</tbody>
</table>
Estimating from Observational Data

- Expressions under the “standard model”

\[ DE_{x_0,x_1}(y|x), IE_{x_0,x_1}(y|x), SE_{x_0,x_1}(y) \] are given by

\[
\sum_{z,w} \left( P(y|x_1, w, z) - P(y|x_0, w, z) \right) P(w|x_0, z) P(z|x),
\]

\[
\sum_{z,w} P(y|x_0, w, z) \left( P(w|x_1, z) - P(w|x_0, z) \right) P(z|x),
\]

\[
\sum_{z,w} P(y|x_0, w, z) P(w|x_0, z) \left( P(z|x_1) - P(z|x_0) \right).
\]
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part V: Challenges and Directions for Future Research
Achieving Non-Discrimination in Prediction

• Will a classifier learned from a discrimination-free training data also be discrimination-free?

• The gap between the discrimination-free training data and the discrimination-free classifier

• Mathematically bound the discrimination in predictions in terms of the training data and the classifier performance.

Causal Modeling-Based Anti-Discrimination Framework

- \( C \): protected attr
- \( Z \): non-protected attrs
- \( L \): decision attr

Population = Causal Model \( \mathcal{M} \)

Causal Model \( \mathcal{M} \)
\[
c = f_C(\text{pa}_C, u_C)
\]
\[
\forall z_i \in Z, \quad z_i = f_i(\text{pa}_i, u_i)
\]
\[
l = f_L(\text{pa}_L, u_L)
\]

Historical data \( \mathcal{D} \)

Train

Classifier \( h(C, Z) \)

Predict

Predicted data \( \mathcal{D}_h \)

Generate

Causal Model \( \mathcal{M}_h \)

Causal Model \( \mathcal{M}_h \)
\[
c = f_C(\text{pa}_C, u_C)
\]
\[
\forall z_i \in Z, \quad z_i = f_i(\text{pa}_i, u_i)
\]
\[
l = h(c, z)
\]
Measure of Discrimination

• Whether the decision of an individual would be different had the individual been of a different protected/non-protected group?
• For each individual \( u \), his/her label under intervention \( do(c^+) \): \( L_{c^+}(u) \)
• Expectation of differences in labels under \( do(c^+) \) and \( do(c^-) \):
\[
\mathbb{E}[L_{c^+}(u) - L_{c^-}(u)]
\]
• Derived causal measures of discrimination:
  - \( DE_M = \mathbb{E}[L_{c^+}(u) - L_{c^-}(u)] = P(l^+|c^+) - P(l^+|c^-) \)
  - \( DE_D = \hat{P}(l^+|c^+) - \hat{P}(l^+|c^-) \) Coincident with risk difference
  - \( DE_{Mh} = \mathbb{E}[h(c^+, Z_{c^+}(u)) - h(c^-, Z_{c^-}(u))] = P(\tilde{l}^+|c^+) - P(\tilde{l}^+|c^-) \)
  - \( DE_{Dh} = \hat{P}(\tilde{l}^+|c^+) - \hat{P}(\tilde{l}^+|c^-) \)
\[
= \sum_z \mathbb{I}[h(c^+, z) = l^+] \hat{P}(z|c^+) - \sum_z \mathbb{I}[h(c^-, z) = l^+] \hat{P}(z|c^-)
\]
Problem Definition

• **Problem 1** (*Discover Discrimination in Prediction*). Given a causal measure of discrimination defined on $\mathcal{M}$, i.e., $DE_\mathcal{M}$, a sample dataset $\mathcal{D}$ and a classifier $h$ trained on $\mathcal{D}$, compute analytic approximation to the true discrimination in prediction, i.e., $DE_\mathcal{M}h$.

• **Problem 2** (*Remove Discrimination in Prediction*). Given $DE_\mathcal{M}$, $\mathcal{D}$ and $h$, tweak $\mathcal{D}$ and/or $h$ in order to make $DE_\mathcal{M}h$ be bounded by a user-defined threshold $\tau$. 
The probability of the difference between $DE_M$ and $DE_D$ no larger than $t$ is bounded by

$$P(|DE_M - DE_D| \leq t) > 1 - 4e^{-rac{n^+n^-}{n}t^2}$$
The probability of the difference between $DE_{\mathcal{M}_h}$ and $DE_{\mathcal{D}_h}$ no larger than $t$ is bounded by

$$P\left( |DE_{\mathcal{M}_h} - DE_{\mathcal{D}_h}| \leq t \right) > 1 - \delta(t)$$

where

$$\delta(t) = \begin{cases} 
\frac{4|\mathcal{H}|^2 e^{-\frac{n^++n^-}{n}t^2}}{(2en^+)^d + (2en^-)^d} e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is finite} \\
\frac{4}{d^d} e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is infinite}
\end{cases}$$
Bound Discrimination in Prediction

Historical data \( \mathcal{D} \) 

Causal Model \( \mathcal{M} \)

Generate

Train

Classifier \( h(C, Z) \)

Predict

Predicted data \( \mathcal{D}_h \)

Generate

Causal Model \( \mathcal{M}_h \)

\[ DE_{\mathcal{D}_h} - DE_{\mathcal{D}} = \varepsilon_{h,\mathcal{D}} \]

where

\[ \varepsilon_{h,\mathcal{D}} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-) \]

% of false positives on data with \( c^+ \) and \( c^- \)

% of false negatives on data with \( c^+ \) and \( c^- \)
**Theorem 1.** Given a causal model $\mathcal{M}$, a sample dataset $\mathcal{D}$ and a classifier $h$ trained on $\mathcal{D}$, $DE_{\mathcal{M}_h}$ is bounded by

$$P \left( |DE_{\mathcal{M}_h}| \leq |DE_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}| + t \right) \geq 1 - \delta(t)$$
Remove Discrimination in Prediction

- Removing discrimination from training data ONLY is NOT enough as discrimination in prediction depends on $DE_{D} + \varepsilon_{h,D}$.

- Two-phase framework for non-discrimination in prediction:
  1. (Data modification) Modify training dataset $D$ to obtain a modified dataset $D^*$ such that $|DE_{D^*}| \leq \tau$;
  2. (Classifier tweaking) Train a classifier $h^*$ on $D^*$ (and tweak it) such that $|DE_{D^*} + \varepsilon_{h^*,D^*}| \leq \tau$.

- What methods can be employed in the framework?
  - Only label-modifying data modification can achieve the guarantee.
  - If any attribute other than the label is modified, the testing data and the training data are from different distributions, and hence no guarantee.
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part IV: Challenges and Directions for Future Research
Suppes-Bayes Causal Network (SBCN)

• Each node represents an assignment attribute value
• Each arc $v \rightarrow u$ represents the existence of a relation satisfying Suppes’ constraints
  – Let $v$ denote cause, $u$ denote effect
  – Temporal priority: $t_v < t_u$
  – Probability raising: $P(u|v) > P(u|\neg v)$
• Each arc is labeled with a positive weight $p(u|v) - p(u|\neg v)$

A SBCN Example
Discrimination Score using SBCN

• Discrimination score
  \[ ds^-(v) = \frac{rw_{v \rightarrow e^-}}{n} \]
  - \( v \) is a node of SBCN (e.g. female), \( e^- \) is the node of negative decision, \( rw_{v \rightarrow e^-} \) is the number of random walks from \( v \) to \( e^- \) that earlier than \( e^+ \), \( n \) is the number of random walks from \( v \) to \( e^+ \) and from \( v \) to \( e^- \).

• Generalized score for individual and subgroup discrimination
  \[ gds^-(v_1, ..., v_n) = \frac{ppr(e^-|v_1, ..., v_n)}{ppr(e^-|v_1, ..., v_n) + ppr(e^+|v_1, ..., v_n)} \]
  - \( ppr(e^-|v_1, ..., v_n) \) is output of personalized PageRank.

• Limitations
  - The constructor of SBCN is impractical with large attribute-value pairs.
  - It is unclear how the number of random walks is related to meaningful discrimination metric.
Achieving Non-Discrimination in Data Release

• An organization/data-owner aims to achieve a non-discrimination guarantee against all possible lawsuits.

• Terminology:
  – Partition: a set of attributes are used to partition data
  – Group: a set of individuals who have the same values in terms of one partition

• Risk difference for group discrimination
  – $\Delta P|_s = P(e^+|c^+, s) - P(e^+|c^-, s)$
  – $\tau$: an user-defined threshold for discrimination detection depending on laws and regulations (e.g., 0.05).
  – If $\Delta P|_s < \tau$ holds across all possible partitions and their values $s$, then no discrimination.

Achieving Non-Discrimination in Data Release

- Achieve a non-discrimination guarantee against all possible lawsuits for all meaningful subgroups.

<table>
<thead>
<tr>
<th>No.</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>...</td>
</tr>
<tr>
<td>major</td>
<td>...</td>
</tr>
<tr>
<td>score</td>
<td>...</td>
</tr>
<tr>
<td>height</td>
<td>...</td>
</tr>
<tr>
<td>weight</td>
<td>...</td>
</tr>
<tr>
<td>admission</td>
<td>...</td>
</tr>
</tbody>
</table>

| gender | Female | Male |
| admission (%) | 43% | 43% |

| major | CS | EE |
| gender | Female | Male | Female | Male |
| admission (%) | 38% | 38% | 47% | 47% |

| major | CS | EE |
| test score | Low | High | Low | High |
| gender | Female | Male | Female | Male | Female | Male |
| admission (%) | 30% | 36% | 50% | 40% | 40% | 45% | 60% | 50% |
Achieving Non-Discrimination in Data Release

- A node set $B$ forms a meaningful partition:
  - $B$ $d$-separates $C$ and $E$ in the graph (deleting $C \rightarrow E$)
  - None of $E$’s children is in $B$
  - $B$ is called a block set

- Ensure $|\Delta P|_b < \tau$ for each $b$ of each $B$.
  - $\Delta P|_b = P(e^+|c^+, b) - P(e^+|c^-, b)$

- Let $Q = Pa(E) \setminus \{C\}$, if $|\Delta P|_q < \tau$ holds, it is guaranteed $|\Delta P|_b < \tau$ holds.
Discrimination Removal

- Modifying the causal graph (MGraph)
  - Modify the CPT of $E$ so that non-discrimination is achieved over its distribution and graph.
  - Generate a new dataset using the modified graph.
  - Minimize the distance of the joint distributions: quadratic programming.

- Modifying the dataset (MData)
  - If $\Delta P|_q \geq \tau$, randomly select a number of individuals from the $\{c^- e^-\}$ group and change decision from $e^-$ to $e^+$.
  - If $\Delta P|_q \leq -\tau$, do the similar modification.
  - As a result, ensure that $|\Delta P|_q \leq \tau$ holds for each $q$. 
Empirical Evaluation

- Data: Adult and Dutch Census
- Evaluated algorithms:
  - MGraph, MData (Zhang et al. SIGKDD 2017)
  - Local massaging (LM) and local preferential sampling (LPS) algorithms (Žliobaite et al. ICDM 2011)
  - Disparate impact removal algorithm (DI) (Feldman et al. SIGKDD 2015)
- Result
  - MGraph and MData totally remove discrimination over all meaningful subgroups.
  - LM, LPS, DI still have discriminated subgroups.
  - MGraph and MData well-preserve data utility.
Individual Discrimination Discovery

• Individual-level discrimination discovery deals with the discrimination that happens to one particular individual.

• Situation testing-based approach:
  – Select pairs of similar individuals to the target from both the protected ($c^-$) group and the unprotected ($c^+$) group.
  – Check whether difference is significant between the decisions of the selected protected and non-protected individuals.

• How to find similar individuals for situation testing?

Individual Discrimination Discovery

- Situation testing: find similar individuals for the target.

<table>
<thead>
<tr>
<th>No.</th>
<th>gender</th>
<th>major</th>
<th>score</th>
<th>height</th>
<th>weight</th>
<th>ad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>CS</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>CS</td>
<td>B</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>CS</td>
<td>A</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>CS</td>
<td>A</td>
<td>median</td>
<td>median</td>
<td>admit</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>CS</td>
<td>C</td>
<td>low</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>CS</td>
<td>C</td>
<td>median</td>
<td>median</td>
<td>reject</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>EE</td>
<td>B</td>
<td>low</td>
<td>low</td>
<td>reject</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
Individual Discrimination Discovery

• The distance function between two individuals $t$ and $t'$ is defined as:

$$d(t, t') = \sum_{k=1}^{|Q|} |CE(q_k, q_k') \cdot VD(q_k, q_k')|$$

• $CE(q_k, q_k')$ measures the causal effect of each attribute $Q_k \in Q$ on the decision when the value of $Q_k$ changes from $q_k$ to $q_k'$. Using the do-operator, it is computed with:

$$CE(q_k, q_k') = P(\epsilon^+ | do(q)) - P(\epsilon^+ | do(q', q \backslash \{q_k\}))$$

• $VD(q_k, q_k')$ measures the difference between two values $q_k$ and $q_k'$ of each attribute $Q_k \in Q$.

$$VD(q_k, q_k') = \begin{cases} \text{Manhattan}(q_k, q_k') & \text{if } Q_k \text{ is ordinal/interval} \\ \text{Overlap}(q_k, q_k') & \text{if } Q_k \text{ is categorical} \end{cases}$$
Empirical Evaluation

• Data: Dutch Census of 2001

• Comparison of Different Methods
  – CBN-based situation testing (CBN-DD) (Zhang et al. IJCAI 2017)
  – KNN-based situation testing (KNN-DD) (Luong et al. SIGKDD 2011)

• Result:
  – KNN-DD and CBN-DD are significantly different.
  – CBN-DD outperforms KNN-DD over the synthetic data.

Accuracy

<table>
<thead>
<tr>
<th>K</th>
<th>CBN-DD</th>
<th>KNN-DD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>TN</td>
</tr>
<tr>
<td>10</td>
<td>73.3</td>
<td>63.1</td>
</tr>
<tr>
<td>50</td>
<td>85.3</td>
<td>77.6</td>
</tr>
<tr>
<td>90</td>
<td>81.5</td>
<td>83.9</td>
</tr>
</tbody>
</table>

• Clean the dataset by “shuffling” gender
• Manually change decision from $e^+$ to $e^-$ for 100 female individuals.
• Use these individuals and another random 100 individuals without discrimination as the targets.
Summary

1. Preliminary
   - Causal Modelling
   - Path-specific
   - Counterfactual

2. Descriptions
   - On discrimination discovery using causal networks
   - Exposing the probabilistic causal structure of discrimination
   - Situation testing-based discrimination discovery: a causal inference approach
   - A causal framework for discovering and removing direct and indirect discrimination
   - Anti-discrimination learning: a causal modeling-based framework
   - Fairness in decision-making: the causal explanation formula
   - Pooling of causal models under counterfactual fairness via causal judgement aggregation
   - On discrimination discovery and removal in ranked data using causal graph
   - Fair inference on outcomes
   - Achieving non-discrimination in data release
   - Achieving non-discrimination in prediction
   - Avoiding discrimination through causal reasoning
   - Counterfactual fairness
   - When worlds collide: integrating different counterfactual assumptions in fairness
Outline

• Part I: Introduction
• Part II: Correlation based Anti-Discrimination Learning
• Part III: Causal Modeling Background
• Part IV: Causal Modeling-Based Anti-Discrimination Learning
  – Direct and Indirect Discrimination
  – Counterfactual Fairness
  – Data discrimination vs. model discrimination
  – Other Works
• Part V: Challenges and Directions for Future Research
Challenges

• Dealing with non-identifiability of path-specific effects
• Causal modeling implementation for mixed-type variables
• Relaxing Markovian assumption
• Dealing with multiple causal models
• Group/Individual-level indirect discrimination
Identifiability

• **Identifiability**: The path-specific effect can be computed from the observational data if and only if the **recanting witness criterion** is NOT satisfied.

• **Recanting witness criterion**: 

![The “kite” structure](image)

\[ \pi \]

\[ C \rightarrow W \rightarrow E \]
Unidentifiable Situation

- When the recanting witness criterion is satisfied, indirect discrimination $SE_{\pi_i}$ cannot be computed from observational data.
- Example:

\[ \pi_i = \{(C, A_2, E), (C, W, A_1, E)\} \]

The "kite" structure
Dealing with Unidentifiable Situation

- Principled approaches for dealing with non-identifiable path-specific effects (Nabi et al., 2018)
  - Measure hidden variables $U$ or obtain reliable proxies for them, if possible.
  - Consider a path-specific effect that is identifiable, which includes the paths of interest and some other paths.
    - The path-specific effect which includes more paths should be an upper bound of the path-specific effect of interest.
  - **Derive theoretical bounds for the non-identifiable path-specific effect.**
    - Zhang et al. TKDE18
    - Some tight bounds may be possible. Tian & Pearl 2000.


Bounding $\pi_i$-specific Effect

\[ \pi_i = \{(C, A_2, E), (C, W, A_1, E)\} \]

The “kite” structure

\[ P(e^+|do(c^+|\pi_i)) = \sum_{A_1, A_2, B, W^+, W^-} P(e^+|c^-, a_1, a_2, b)P(a_1|w^+)P(a_2|c^+)P(b|w^-)P(w^+_c, w^-_c) \]

- Counterfactual: $W$ would be $w^+$ if $C = c^+$ and $W$ would be $w^-$ if $C = c^-$.
- Generally unidentifiable from observational data or even controlled experiment.
- Bounded by condition \[ \sum_{w^-} P(w^+_c, w^-_c) = P(w^+_c) \]

Bounding $\pi_i$-specific Effect

- $SE_{\pi_i}(c^+, c^-) = P(e^+ | do(c^+ | \pi_i)) - P(e^+ | do(c^-))$
- Upper bound of $P(e^+ | do(c^+ | \pi_i))$

$$\sum_{a_2, b, w^-} \max\{P(e^+ | c^-, q)\} \prod_{A \in A_2} P(a|pa_A^+) \prod_{B \in B} P(b|pa_B^-) \prod_{W \in W} P(w^- | pa_W^-)$$

- Lower bound

- Notations:
  - $W$: witness nodes
  - $A_1$: nodes in $\pi_i$ not in $W$ but involved in “kite pattern”
  - $A_2$: nodes in $\pi_i$ not in $W$ and not involved in “kite pattern”
  - $B$: nodes not in $\pi_i$
Using Bounds for Discrimination Discovery and Removal

• Utilize lower and upper bounds for identifying indirect discrimination.
  – If upper bound < threshold, non-discrimination for certain.
  – If lower bound ≥ threshold, discrimination for certain.
  – If lower bound < threshold ≤ upper bound, uncertain.

• For removal, replace \( SE_{\pi_i}(c^+, c^-) \) with its upper bound in constraints of quadratic programming.
  – The solution of the “simple” method is a feasible solution of the above quadratic programming problem.
Empirical Evaluation

protected attribute: **sex**  
decision: **income**  
redlining attribute: **edu**

The “kite” structure
Challenges

• Dealing with non-identifiability of path-specific effects
• Causal modeling implementation for mixed-type variables
• Relaxing Markovian assumption
• Dealing with multiple causal models
• Group/Individual-level Indirect Discrimination
Causal Modeling for Mixed-type Variables

- Most existing works construct causal graph for categorical variables.
- For mixed-type variables, one option is Conditional Linear Gaussian (CLG) Bayesian network

\[
\begin{align*}
\text{discrete variable: CPT} & \quad \text{continuous variable: CLG distribution} \\
\Pr(x) & \quad Y \sim N(\mu_Y, \sigma_Y^2) \\
\text{Parameters: } a(\cdot), b(\cdot), c(\cdot) & \quad Z|x, y \sim N\left(a(x) + b(x)y, c(x)\right)
\end{align*}
\]

- Limitation: discrete variables can only have discrete parents.

• (Kocaoglu et al. 2018) uses neural network architecture to represent causal graph.

Challenges

• Dealing with non-identifiability of path-specific effects
• Causal modeling implementation for mixed-type variables
• Relaxing Markovian assumption
• Dealing with multiple causal models
• Group/Individual-level indirect discrimination
Relaxing Markovian Assumption

• A causal model is Markovian if
  1. The causal graph is acyclic;
  2. All variables in $U$ are mutually independent.
Relaxing Markovian Assumption

• A causal model is semi-Markovian if
  1. The causal graph is acyclic;
  2. All variables in $U$ are NOT mutually independent.

• Hidden confounders are known to exist in the system.

• The causal graph of the semi-Markovian model is commonly represented by the acyclic directed mixed graph (ADMG).
  – The bidirected arrow $\leftrightarrow$ implies the presence of unobserved confounder(s) between variables.
Relaxing Markovian Assumption

Markovian model

Semi-Markovian model

Directed Acyclic Graph (DAG)

Acyclic Directed Mixed Graph (ADMG)
Intervention in Semi-Markovian Model

• Intervention also applies to semi-Markovian model.

• Unlike in the Markovian model, $do$-operations may not be able to be calculated from observational data (i.e., identifiable) due to unobserved confounders.

• “Bow-arc graph”: the simplest non-identifiable graph structure.

$P(y \mid do(x))$ is non-identifiable in this graph
Intervention in Semi-Markovian Model

- Graphical criterion of identification:
  - Sufficient condition: *back-door criterion*
    - $P(y \mid \text{do}(x))$ is identifiable if exists a set of observed variables $Z$ that blocks all back-door paths from $X$ to $Y$
  - Sufficient condition: *front-door criterion*
    - $P(y \mid \text{do}(x))$ is identifiable if exists a set of observed variables $Z$ such that:
      - $Z$ blocks all causal paths from $X$ to $Y$;
      - There is no back-door path from $X$ to $Z$;
      - All back-door paths from $Z$ to $Y$ are blocked by $X$.
  - Complete criterion: *hedge criterion* (Shpitser et al., 2008)

- Complete identification algorithm: ID (Shpitser et al., 2008)

Relaxing Markovian Assumption

- **ID algorithm** for identification of interventions.
- **ID* algorithm** for identification of counterfactuals.
- Generalize the $d$-separation to $m$-separation.
- For path-specific effect, generalize recanting witness criterion to recanting district criterion.

Any anti-discrimination method designed for semi-Markovian models must be adapted to the differences in the causal inference techniques.
Challenges

• Dealing with non-identifiability of path-specific effects
• Causal modeling implementation for mixed-type variables
• Relaxing Markovian assumption
• Dealing with multiple causal models
• Group/Individual-level indirect discrimination
Fairness under Causal Model Aggregation

• Sometimes there may be multiple plausible causal models.
  – Provided by different experts.
  – Learned from data as the Markov equivalent class.

• Make predictions that are approximately fair with respect to multiple possible causal models.

• Potential solutions:
  – Opinion pooling
  – Aggregated fairness constraints

Challenges

- Dealing with non-identifiability of path-specific effects
- Causal modeling implementation for mixed-type variables
- Relaxing Markovian assumption
- Dealing with multiple causal models
- Group/Individual-level indirect discrimination
Group and Individual-Level Indirect Discrimination

- (Zhang et al. IJCAI 2017): direct/indirect discrimination at the system-level using **path-specific effect**.

- (Zhang et al. AAAI 2018): direct/indirect discrimination in protected and non-protected groups using **path-specific effect** (limited to direct/indirect effects) and **counterfactual** (limited to conditioning on protected attribute).

Group and Individual-Level Indirect Discrimination

• In general, dealing with group and individual-level indirect discrimination requires path-specific effect (any set of paths) and counterfactual (conditioning on any set of attributes), i.e.,

\[ \text{Path-specific counterfactual quantity} \ (Y_x|\pi \mid e) \]

• Has identifiability issues regarding both path-specific effect and counterfactual.

• Find assumptions for path-specific counterfactual quantity to be identifiable
  – E.g., causal linear models.
Future Directions

• Building Non-discrimination Predictors
  – Causal effects as constraints for classification
  – Direct/indirect discrimination: data vs. model
  – Trade-off between non-discrimination and accuracy

• Discrimination in tasks beyond classification
  – Ranking and recommendation
  – Generative adversarial network (GAN)
  – Dynamic data and time series
  – Text and image

• Transparency in learning process
Causal Effects as Constraints for Classification

- Classifier learning with fairness constraints

\[
\min_{h \in \mathcal{H}} \mathbb{L}(h) \quad \text{Minimize the loss function}
\]

\[
s.t. \ C(h) \leq \tau \quad \text{Subject to fairness constraints}
\]

- Challenges:
  - For computational tractability, how to transform causal effect-based fairness constraints to convex constraints?
  - How to deal with estimation errors due to the use of surrogate functions?
• Classification fairness is measured using risk difference
  \[ RD(f) = \mathbb{E}_{X|S=s^+}[1_{f(x)=1}] - \mathbb{E}_{X|S=s^-}[1_{f(x)=1}] \]

• Learn a classifier with fairness constraints
  \[ \min_{h \in \mathcal{H}} \mathbb{L}_{\phi}(h) \]
  \[ \text{subject to } RD_{\kappa}(h) \leq c_1, \quad -RD_{\delta}(h) \]
  – For computational feasibility, the loss function, fairness constraints are surrogated by convex/concave functions \( \phi, \kappa, \delta \).

• Bounding fairness constraints with surrogate function
  \[ \min_{h \in \mathcal{H}} \mathbb{L}_{\phi}(h) \]
  \[ \text{subject to } RD_{\kappa}(h) \leq \psi_{\kappa}(\tau - RD^-) + RD_{\kappa}^- \]
  \[ -RD_{\delta}(h) \leq \psi_{\delta}(\tau + RD^+) + RD_{\delta}^+ \]

\(-\tau \leq RD(h^*) \leq \tau\) is guaranteed

Direct/Indirect Discrimination: Data vs. Model

- Zhang et al. IJCAI 2018: target total effect.
Trade-Off

- How to balance the trade-off between non-discrimination and utility loss?

**Diagram:**
- Historical data $D$
- Causal Model $M$
- Discrimination removal
- Train
- Predict
- Predicted data $D_h$
- Generate
- Infer
- Causal Model $M_h$
Discrimination in Tasks Beyond Classification

• Currently mainly focus on classification problems.
• Tasks beyond classification:
  – Recommendation: a list of recommended items
  – Ranking: ranking positions of candidates
  – Generative adversarial network (GAN): a learned representation
  – Dynamic and time series data
  – Text and image
  – ...
• Transparency in learning process
Fairness-aware Recommendation

- Fairness-aware Recommendation

- No causal modeling based method
Fairness-aware Ranking

- Fairness-aware ranking
Fair Ranking

- Decisions are given in permutation rather than binary decisions.

<table>
<thead>
<tr>
<th>ID</th>
<th>Race</th>
<th>Zip Code</th>
<th>Interview</th>
<th>Education</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>W</td>
<td>72701</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>U2</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>U3</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>U4</td>
<td>W</td>
<td>72701</td>
<td>4</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>U5</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>U6</td>
<td>B</td>
<td>72701</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>U7</td>
<td>B</td>
<td>72702</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>U8</td>
<td>B</td>
<td>72702</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>U9</td>
<td>B</td>
<td>72701</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>U10</td>
<td>B</td>
<td>72702</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Causal graphs cannot be built directly for ranked data:
- Causal graphs must be built for random variables,
- But ranking is a permutation of a series of unique, concatenating integers.

## Fair Ranking

<table>
<thead>
<tr>
<th>ID</th>
<th>Race</th>
<th>Zip Code</th>
<th>Interview</th>
<th>Education</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>W</td>
<td>72701</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>U2</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>U3</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>U4</td>
<td>W</td>
<td>72701</td>
<td>4</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>U5</td>
<td>W</td>
<td>72701</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>U6</td>
<td>B</td>
<td>72701</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>U7</td>
<td>B</td>
<td>72702</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>U8</td>
<td>B</td>
<td>72702</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>U9</td>
<td>B</td>
<td>72701</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>U10</td>
<td>B</td>
<td>72702</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

**Bradley-Terry Model**

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>25.45</td>
</tr>
<tr>
<td>14.13</td>
</tr>
<tr>
<td>35.62</td>
</tr>
<tr>
<td>45.26</td>
</tr>
<tr>
<td>100.00</td>
</tr>
<tr>
<td>74.54</td>
</tr>
<tr>
<td>85.86</td>
</tr>
<tr>
<td>54.74</td>
</tr>
<tr>
<td>64.38</td>
</tr>
</tbody>
</table>
Fair Ranking

- Map ranking positions to continuous score using Bradley-Terry Model.
  \[ s_i - s_j = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right), \]
  \[ P(\omega | M) \propto \sum_{(i, j): \omega_i < \omega_j} p_{ij} \]

- Build a mixed-variable causal graph using conditional Gaussian distributions.
• Derive direct and indirect discrimination measure in mixed-variable causal graph.

\[
D E_{\pi_d}(c^+, c^-) = \frac{D E_{\pi_d}(c^+, c^-)}{E[S|c^+]} = \frac{\sum_{Z,E,I}(\mu_{c^+, z, e, i} - \mu_{c^-, z, e, i}) P(z, e, i|c^-)}{E[S|c^+]} \]

\[
D E_{\pi_i}(c^+, c^-) = \frac{D E_{\pi_i}(c^+, c^-)}{E[S|c^+]} = \frac{\sum_{Z,E,I} \mu_{c^-}, z, e, i P(e, i)[P(z|c^+) - P(z|c^-)]}{E[S|c^+]} \]

• Identify the relationship between discrimination in ranking and discrimination in binary decision.

– Assume that the decision is made based on a cut-off point \( \theta \) of the score. If then \( \theta \geq \mu_{c^+, q} \geq \mu_{c^-, q} \),

\[
SE^{\text{Bin}}_{\pi_d} \leq \tau \implies SE_{\pi_d} \leq \frac{2\sqrt{2}(\tau - \beta)\sigma}{\alpha}
\]

\[
SE^{\text{Bin}}_{\pi_i} \leq \tau \implies SE_{\pi_i} \leq \frac{2\sqrt{2}(\tau - c)\sigma}{\alpha}
\]
Fair Generative Adversarial Networks

- Fair Generative Adversarial Networks
• Instead of modifying the training data to remove discriminatory effect, FairGAN can directly generate fair data.

• Generative adversarial networks (GANs) is able to generate high quality synthetic data that are similar to real data.

• Besides generating synthetic samples that match the distribution of real data, FairGAN also aim to prevent the discrimination (with no risk difference) in the generated dataset.

Fair Data Generation

- The **first minimax game** ensures the generated data close to the real data
- The **second minimax game** ensures fairness by removing the correlation between unprotected attributes, decision and the protected attributes

\[
\min_{G_{Dec}} \max_{D_1, D_2} V(G_{Dec}, D_1, D_2) = V_1(G_{Dec}, D_1) + \lambda V_2(G_{Dec}, D_2),
\]

where

\[
V_1(G_{Dec}, D_1) = \mathbb{E}_{c \sim P_{data}(c), (x, e) \sim P_{data}(x, e|c)} \left[ \log D_1(x, e, c) \right] \\
+ \mathbb{E}_{s \sim P_G(s), (\hat{x}, \hat{e}) \sim P_G(x, e|c)} \left[ \log(1 - D_1(\hat{x}, \hat{e}, \hat{c})) \right]
\]

\[
V_2(G_{Dec}, D_2) = \mathbb{E}_{(\hat{x}, \hat{e}) \sim P_G(x, e|c=1)} \left[ \log D_2(\hat{x}, \hat{e}) \right] \\
+ \mathbb{E}_{(\hat{x}, \hat{e}) \sim P_G(x, e|c=0)} \left[ \log(1 - D_2(\hat{x}, \hat{e})) \right]
\]
Dealing with dynamic data and time series

- Structural causal model mainly deals with non-temporal data.
- Causal relationship in time series: Granger causality
  - One time series is useful in predicting another
  - Granger causality is not necessarily true causality

- How to integrate the Granger causality with the structural causal model?
Future Directions

• Building Non-discrimination Predictors
  – Causal effects as constraints for classification
  – Direct/indirect discrimination: data vs. model
  – Trade-off between non-discrimination and accuracy

• Discrimination in tasks beyond classification
  – Ranking and recommendation
  – Generative adversarial network (GAN)
  – Dynamic data and time series
  – Text and image

• Transparency in learning process
Thank you

Lu Zhang  lz006@uark.edu
Yongkai Wu  yw009@uark.edu
Xintao Wu  xintaowu@uark.edu

This work is supported by NSF 1646654.

Slides is available at: http://www.csce.uark.edu/~xintaowu/kdd18-tutorial/
References

References