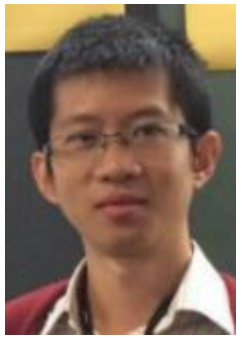
Anti-Discrimination Learning: from Association to Causation



Lu Zhang

Yongkai Wu

Xintao Wu

Social Awareness and Intelligent Learning Lab
University of Arkansas

KDD 2018 Tutorial, Aug 19, 2018. London, UK

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

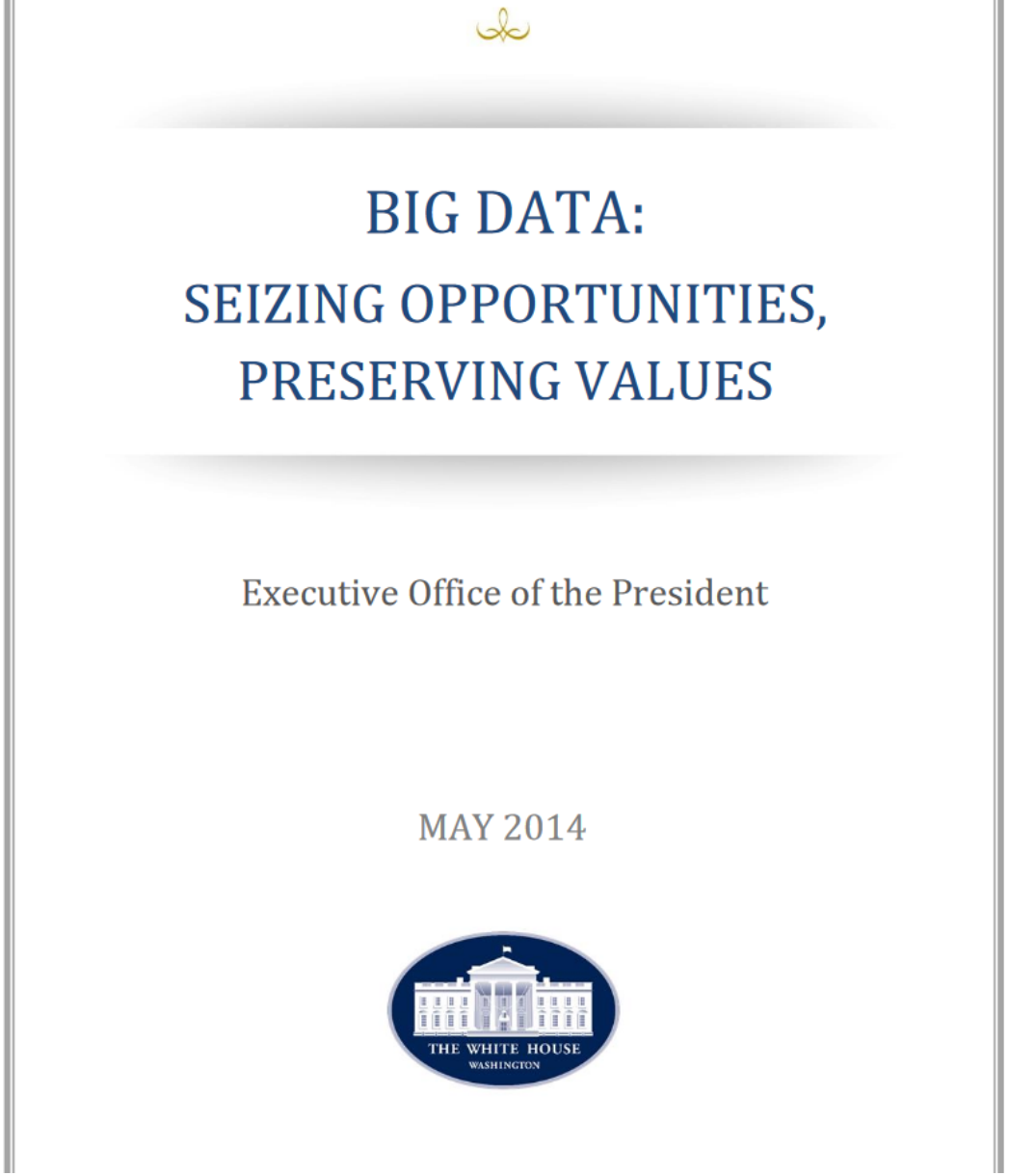
Laws and Regulations

- 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.



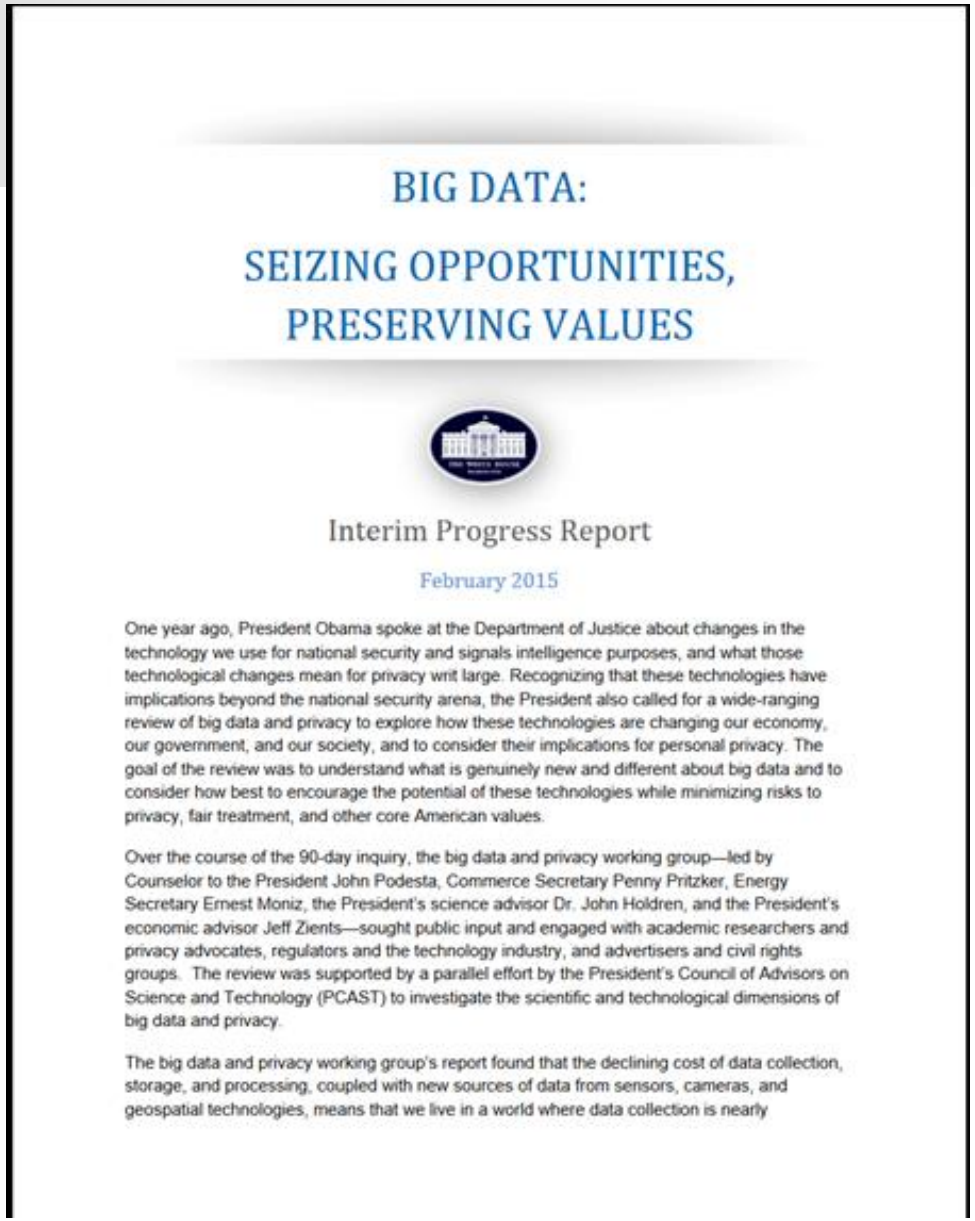
Introduction

February 2015

Pay attention to the potential for big data to facilitate discrimination

Expand technical expertise to stop discrimination

Deepen understanding of differential pricing



BIG DATA:

SEIZING OPPORTUNITIES, PRESERVING VALUES

Interim Progress Report

February 2015

One year ago, President Obama spoke at the Department of Justice about changes in the technology we use for national security and signals intelligence purposes, and what those technological changes mean for privacy writ large. Recognizing that these technologies have implications beyond the national security arena, the President also called for a wide-ranging review of big data and privacy to explore how these technologies are changing our economy, our government, and our society, and to consider their implications for personal privacy. The goal of the review was to understand what is genuinely new and different about big data and to consider how best to encourage the potential of these technologies while minimizing risks to privacy, fair treatment, and other core American values.

Over the course of the 90-day inquiry, the big data and privacy working group—led by Counselor to the President John Podesta, Commerce Secretary Penny Pritzker, Energy Secretary Ernest Moniz, the President's science advisor Dr. John Holdren, and the President's economic advisor Jeff Zients—sought public input and engaged with academic researchers and privacy advocates, regulators and the technology industry, and advertisers and civil rights groups. The review was supported by a parallel effort by the President's Council of Advisors on Science and Technology (PCAST) to investigate the scientific and technological dimensions of big data and privacy.

The big data and privacy working group's report found that the declining cost of data collection, storage, and processing, coupled with new sources of data from sensors, cameras, and geospatial technologies, means that we live in a world where data collection is nearly

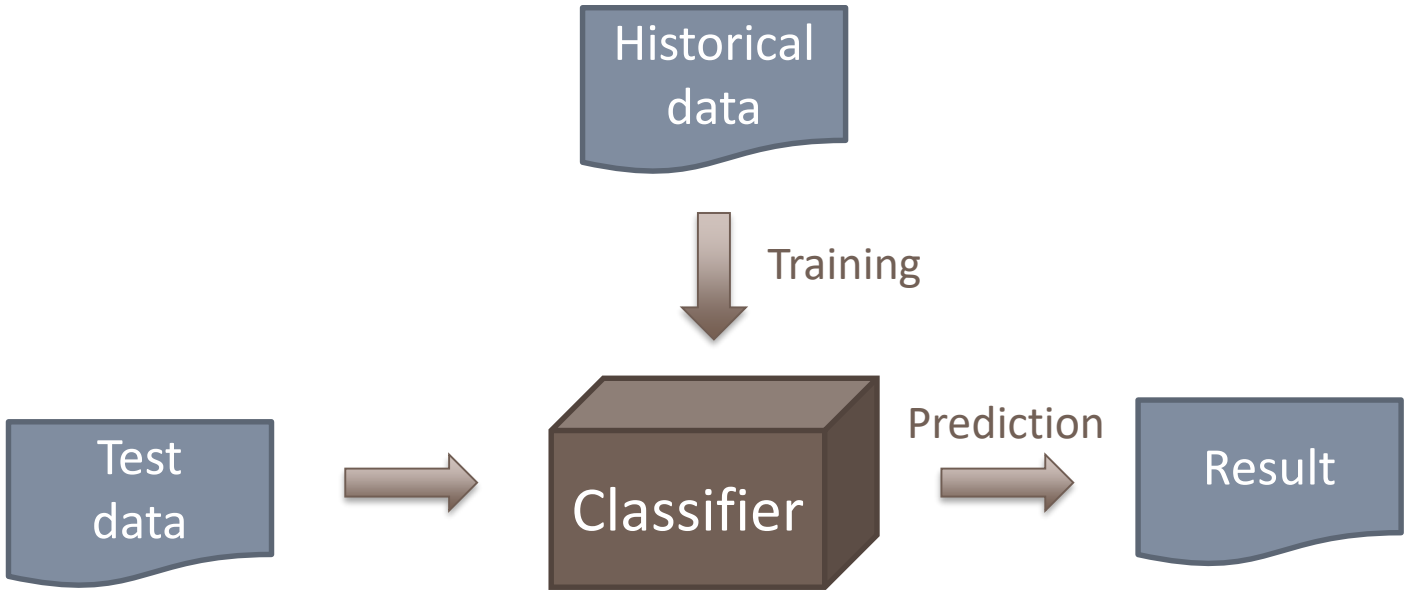
Introduction

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 Transparecy in Data Analysis, Tutorial of EDBT 2016
  - Dwork, C. What's Fair. Keynote of KDD 2017
  - Barocas, S., Hardt, M.: Fairness in machine learning. Tutorial of NIPS 2017
- Survey papers and books
  - Magnani, L., Board, E., Longo, G., Sinha, C., Thagard, P.: Discrimination and privacy in the information society. Springer (2013)
  - Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowl. Eng. Rev. 29(05), 582–638 (2014)
  - Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. 4(1), 1-16 (2017)

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

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

Notations

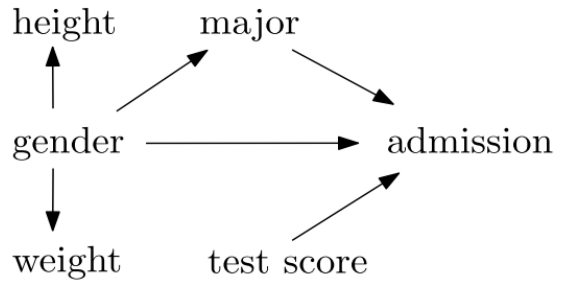
- Denote an attribute by an uppercase alphabet, e.g., X
- Denote a value of attribute X by x
- Denote a subset of attributes by a bold uppercase alphabet, e.g., $\boldsymbol{X}$
- Denote a value assignment of attributes $\boldsymbol{X}$ by $\boldsymbol{x}$

- 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 $\boldsymbol{X}$ among which $\boldsymbol{R}$ are redlining attributes.
- A predictor of decision $\hat{E} = f(C, \boldsymbol{X})$ (sometimes use $\hat{Y} = f(C, \boldsymbol{X})$).

Illustrative Example

- Gender discrimination in college admission.

No.	gender	major	score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject



C is gender, c^- = female, c^+=male.
E is admission, e^-= reject, e^+=admit.

Measuring Discrimination

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

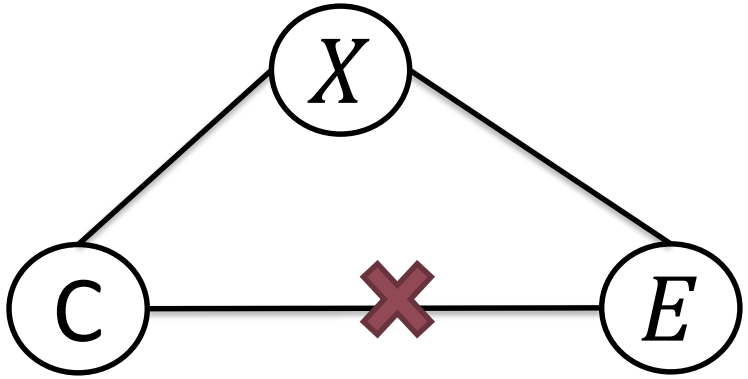
Conditional Independence

- 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.

Fairness through Unawareness

- 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 $\boldsymbol{X}$ can contain discriminatory information.



Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS'12 (2012)

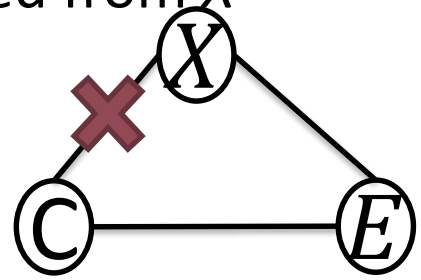
Disparate Impact

- Disparate Impact (DI) aims for unintentional bias
  - No rigid math formula
  - Feldman et al. define DI with risk ratio $\mathrm{DI} = \frac{c/(a+c)}{d/(b+d)}$

	$C = c^-$	$C = c^+$
$E = e^-$	a	b
$E = e^+$	c	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):

$$\mathrm{BER}(f(X), C) = \frac{P(f(X) = c^- | C = c^+) + P(f(X) = c^+ | C = c^-)}{2}$$

  - A dataset is ϵ-fairness if $\mathrm{BER}(f(X), C) > \epsilon$

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15 (2015)

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.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML'13 (2013)

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.

kNN-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 $\boldsymbol{S}^+$
  - individuals with c^- are added into set $\boldsymbol{S}^-$;
- If $P(e^+|\boldsymbol{S}^+) - P(e^+|\boldsymbol{S}^-) > \tau$, 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)

Protected group vs. unprotected group

- Extended Risk Difference (ED)
- Extended Risk Ratio (ER)
- Extended Chance (EC)

Protected group vs. entire population

	benefit		
group	granted	denied	
unprotected	a	b	n_1
protected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

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

Statistical Parity

- Naturally extend to subgroups, e.g., admission rate difference between female and male applying for CS
  - $P(e^+ \mid c^+, \boldsymbol{X} = \boldsymbol{x}) - P(e^+ \mid c^-, \boldsymbol{X} = \boldsymbol{x})$ where $\boldsymbol{X}$ 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.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS'12 (2012)

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:

$$P\left(\hat{E} = e^{+} \mid C = c^{-}, E = e\right) = P\left(\hat{E} = e^{+} \mid C = c^{+}, E = e\right), \qquad e \in \{e^{+}, e^{-}\}$$

- Equal opportunity:

$$P\left(\hat{E} = e^{+} \mid C = c^{-}, E = e^{+}\right) = P\left(\hat{E} = e^{+} \mid C = c^{+}, E = e^{+}\right)$$

  - True positive rate should be the same for all the groups.

[1] Hardt M., Price E., Srebro N.: Equality of opportunity in supervised learning. In: NIPS'16 (2016)
[2] Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW'17 (2017)
[3] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: SIGKDD'17 (2017)

Test Fairness

- Test fairness (calibration)
  - $P\left(E = e^{+} \mid C = c^{-}, \hat{E} = e^{+}\right) = P\left(E = e^{+} \mid C = c^{+}, \hat{E} = e^{+}\right)$
  - 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

Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: ITCS'17 (2017)

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 $\hat{E}$ ($E \perp C|\hat{E}$)

Barocas, S., Hardt, M.: Fairness in machine learning. Tutorial, NIPS'17 (2017)

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

Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM'11 (2011)

Examples

Example 1				
Major	Medicine		Computer	
Gender	female	male	female	male
# of applicants	800	200	200	800
Acceptance rate (%)	20%	20%	40%	40%

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

Example 2				
Major	Medicine		Computer	
Gender	female	male	female	male
# of applicants	800	200	200	800
Acceptance rate (%)	15%	25%	35%	45%

$P(accepted|male) = 41\%$
$P(accepted|female) = 19\%$
$D_{all} = 22\%$
$D_{exp} = 12\%$
$D_{bad} = 10\%$

α-Discrimination in Association Rules

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

$$\{Race = black, Income = high\} \longrightarrow Loan = reject$$

- 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

$$\{ZipCode = 70201, Income = high\} \longrightarrow Loan = reject$$

Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.* 25(7), 1445-1459 (2013)

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})

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

$C_i = female$
$X_j = CS$
$E_k = reject$

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

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

- Extendable to multiple protected/decision attributes

Wu,Y., Wu,X.: Using loglinear model for discrimination discovery and prevention. In: DSAA'16 (2016)

$belift$ Based on Bayesian networks

- $belift = \frac{P(e^+|c_1,c_2,\ldots c_l, x_1,x_2,\ldots x_m, r_1,r_2,\ldots r_n)}{P'(e^+|x_1,x_2,\ldots x_m)}$
  - 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.

Mancuhan, K., Clifton, C.: Combating discrimination using Bayesian networks. *Artif. Intell. Law* 22(2), 211–238 (2014)

Discrimination discovery using $belift$

- 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 $P(e^+|c_1, c_2, \dots c_l, x_1, x_2, \dots x_m, r_1, r_2, \dots r_n)$ over G
  - Compute $P'(e^+|x_1, x_2, \dots x_m)$ over G'
  - Calculate $belift$ 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.

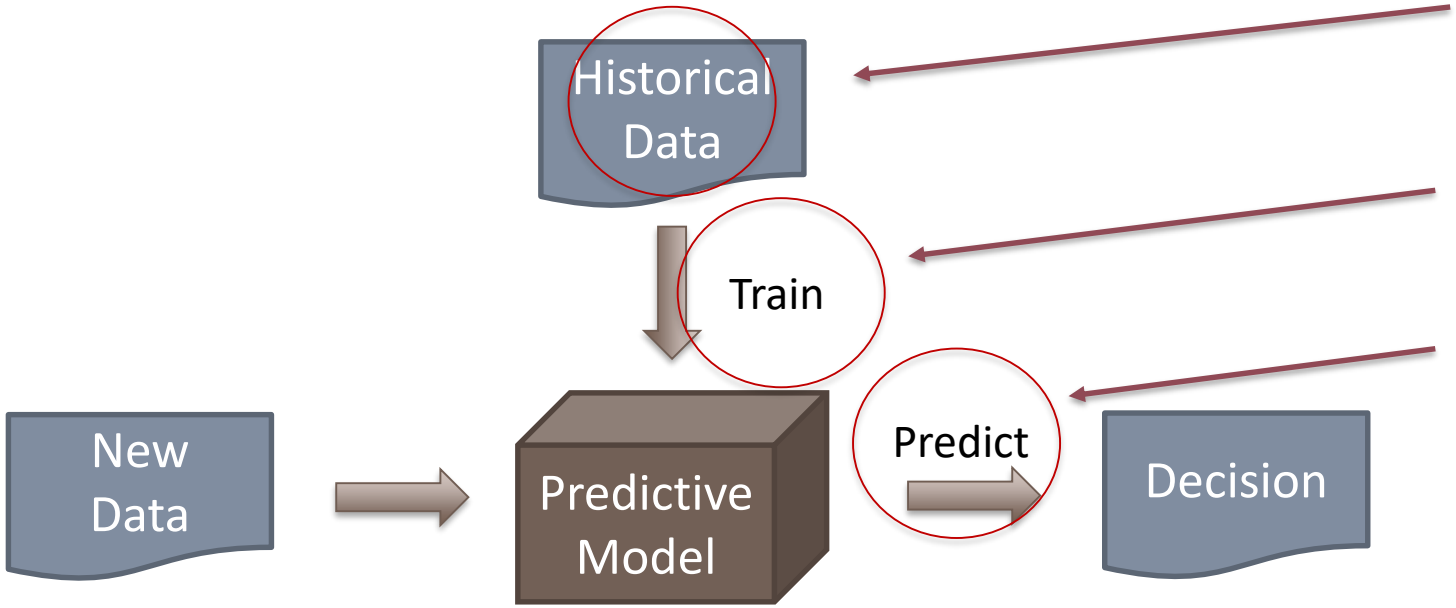
Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., Weller, A. : From parity to preference-based notations of fairness in classification. In: NIPS'17 (2017)
Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. Preprint (2018)

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

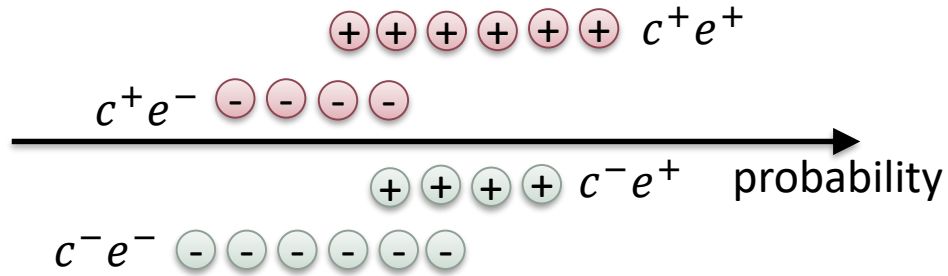
- Pre-processing
  - Data modification
  - Fair data representation
  - Fair data generation
- In-processing
  - Regularization
  - Explicit constraints
- Post-processing

Discrimination Prevention

- Data manipulation (Pre-processing)
  - Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1), 1-33 (2012)
    - Suppression/Massaging/Reweighting/Sampling (uniform vs. preferential sampling)
  - Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.* 25(7), 1445–1459 (2013)
  - Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML'13 (2013)
  - Mancuhan, K., Clifton, C.: Combating discrimination using Bayesian networks. *Artif. Intell. Law* 22(2), 211–238 (2014)
  - Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15 (2015)
  - Edwards, H., Storkey, A.: Censoring representations with an adversary. In: ICLR'16 (2016)
  - Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: ICML'18 (2018)

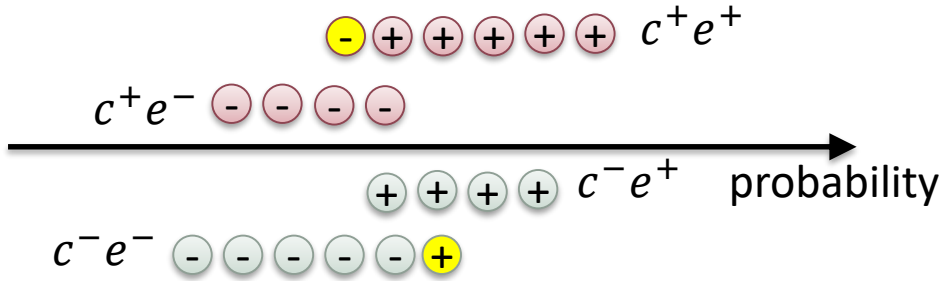
Massaging

- 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



$$RD = \frac{6}{10} - \frac{4}{10} = 0.2$$

  3. Flip the decision of individuals that close to the bottom/top

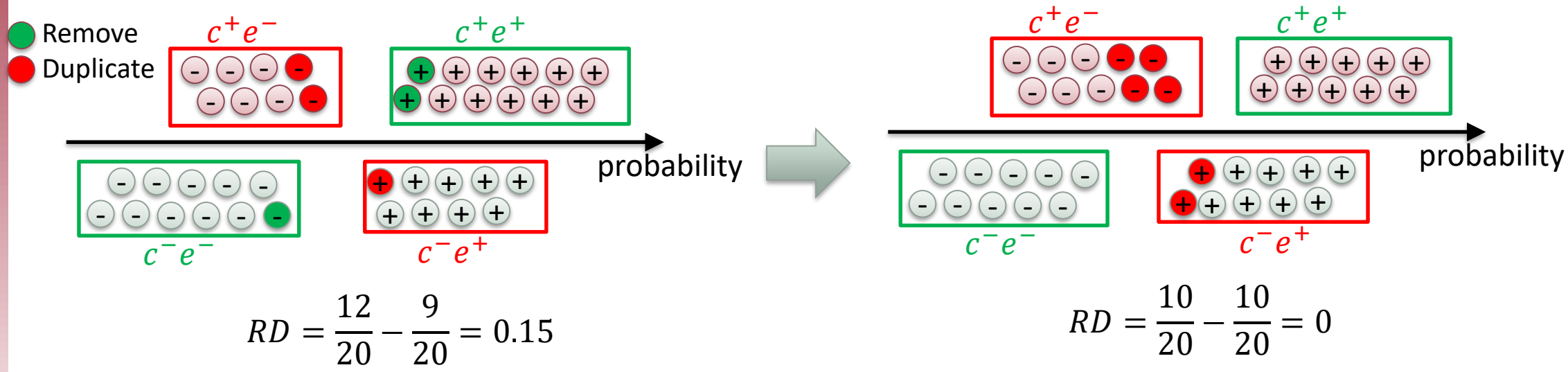


$$RD = \frac{5}{10} - \frac{5}{10} = 0.0$$

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



$$RD = \frac{12}{20} - \frac{9}{20} = 0.15$$

$$RD = \frac{10}{20} - \frac{10}{20} = 0$$

Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1), 1–33 (2012)

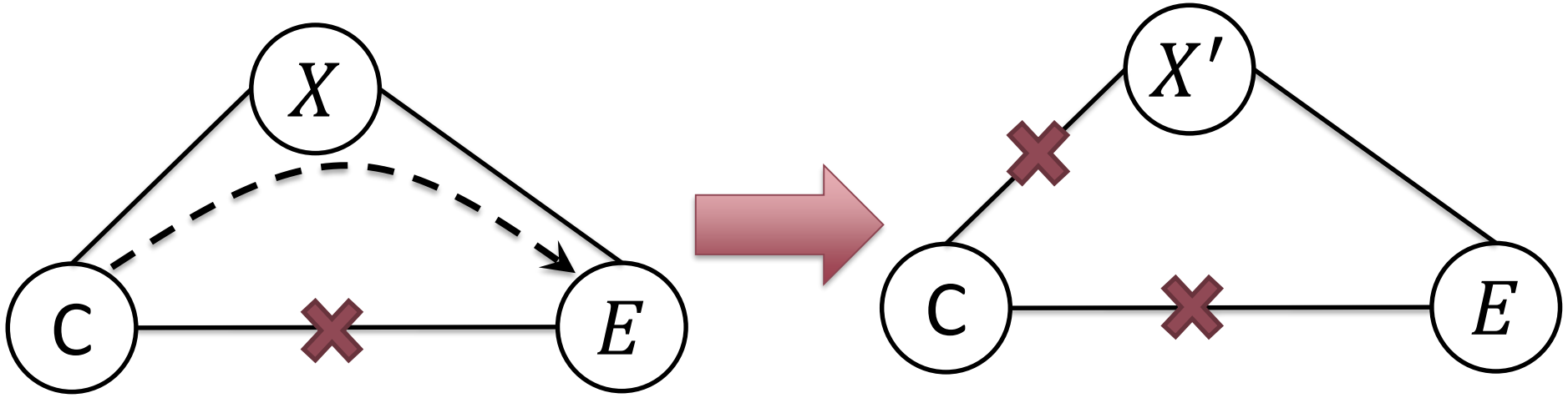
Conditional Discrimination

- $\text{diff} = P(e^+|c^+) - P(e^+|c^-)$ 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
- Zlibobaite et al. propose local massaging and local preferential sampling to removal bad discrimination

Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM'11 (2011)

Removing Disparate Impact

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



Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15 (2015)

Learning Fair Representation

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

- Minimize the objective function

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

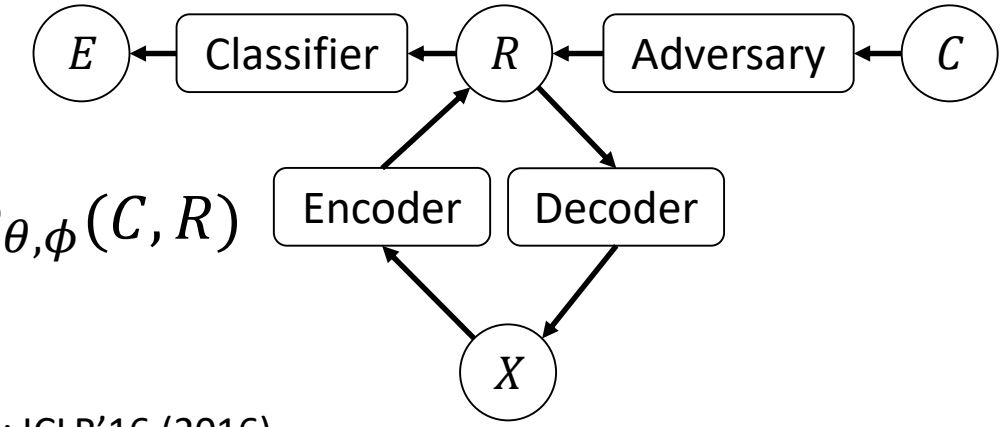
A_x, A_y, A_z are hyper-parameters

  - L_z captures the statistical parity of the representation.
  - L_x constrains the re-construction error.
  - L_y requires the accurate prediction.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning Fair Representations. In: ICML'13 (2013)

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

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

Edwards, H., Storkey, A.: Censoring representations with an adversary. In: ICLR'16 (2016)

51.Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: ICML'18 (2018)

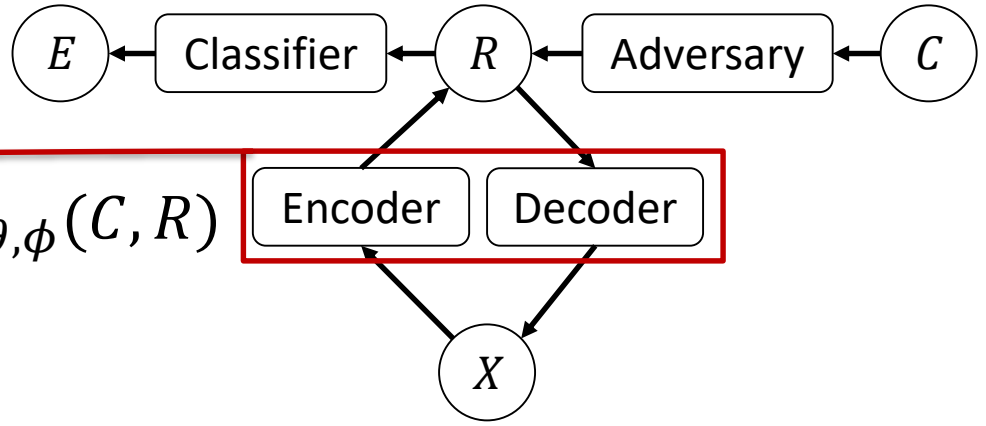
Fair Representations

- 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

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

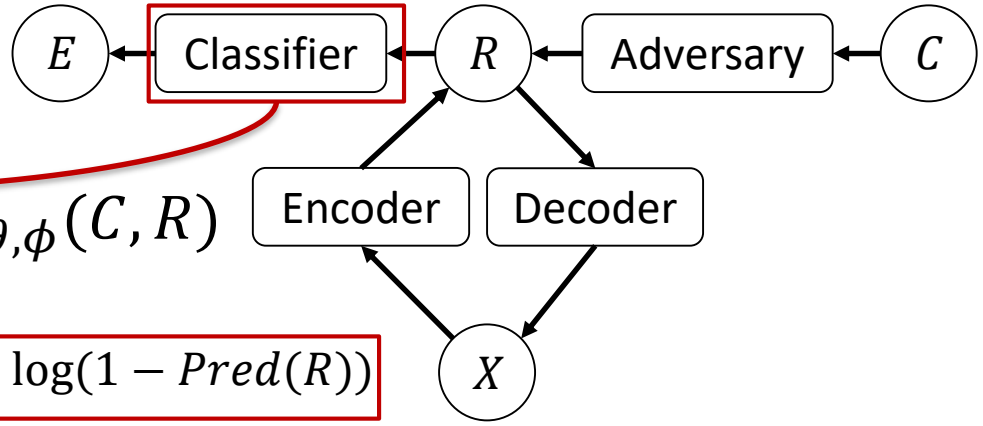
Auto-encoder loss: $\mathcal{C}_\theta(X,R) = \|X - Dec(R)\|_2^2$

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



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

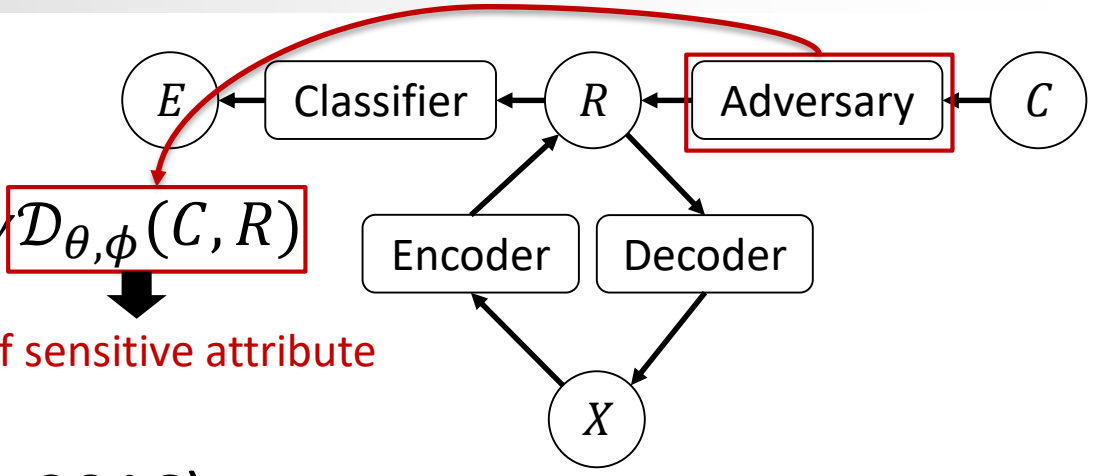
Classification loss: $\mathcal{B}_\eta(E,R) = -E \cdot \log\big(Pred(R)\big) - (1-E) \cdot \log(1 - Pred(R))$

Train prediction model

Fair Representations

- Loss function

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



Adversarial training to remove the information of sensitive attribute

  - Risk difference (Edwards, et al. 2016)

$$\min_{\theta}\max_{\phi}\mathcal{D}_{\theta,\phi}(C,R) = C\cdot\log\big(Adv(R)\big) + (1-C)\cdot\log(1-Adv(R))$$

  - Equalized odds (Madras, et al. 2018)

$$\min_{\theta}\max_{\phi}\mathcal{D}_{\theta,\phi}(C,R) = 2 - \sum_{(i,j)\in\{0,1\}^2}\frac{1}{|\boldsymbol{V}_i^j|}\sum_{(c,x)\in\boldsymbol{V}_i^j}|Adv(R)-R|$$

where $\boldsymbol{V}_i^j = \{(c,x,e)\in\boldsymbol{V}|c=i,e=j\}$

Discrimination Prevention

- Algorithm tweak (In-processing)
  - Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21(2), 277-292 (2010)
  - Kamishima, T., Akaho, S., and Sakuma J.: Fairness-aware learning through regularization approach. In: ICDMW'11 (2011)
  - Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P.: Fairness constraints: Mechanisms for fair classification. In: AISTAS'17 (2017)
  - Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW'17 (2017)
  - Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: SIGKDD'17 (2017)

Fair Regularization for Classification

- Objective functions:

$$-\ell(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Maximize Log-likelihood | Avoid discrimination | Avoid overfitting

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

$$\mathrm{PI} = \sum_{Y,S} \hat{\Pr}[Y,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

$$= \sum_{Y,X,S} \mathcal{M}[Y|X,S;\boldsymbol{\Theta}]\tilde{\Pr}[X,S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

$\widetilde{Pr}[\cdot]$: induced by the training sample
$\widehat{Pr}[\cdot]$: induced by the model

Kamishima, T., Akaho, S., and Sakuma J.: Fairness-aware learning through regularization approach. In: ICDMW'11 (2011)

Fairness Constraints for Classification

- Classification fairness is measured using risk ratio
  - Classifier $f(\boldsymbol{x})$ is learned by minimizing a loss function $L(\boldsymbol{\theta})$.
  - $f(\boldsymbol{x}_i) = 1$ if $d_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \geq 0$ and $f(\boldsymbol{x}_i) = -1$ otherwise.
- Use the covariance as the measure of fairness.

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

- Two formulations

Minimize $L(\boldsymbol{\theta})$

Subject to $\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \leq \tau$

$\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \geq -\tau$

Minimize $\left|\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right|$

Subject to $L(\boldsymbol{\theta}) \leq (1+\gamma)L(\boldsymbol{\theta}^*)$

Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P.: Fairness constraints: Mechanisms for fair classification. In: AISTAS'17 (2017)

Discrimination Prevention

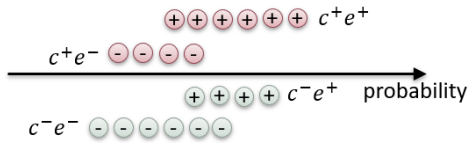
- Prediction changing (Post-processing)
  - Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: ICDM'12 (2012)
  - Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination-and privacy-aware patterns. *Data Min. Knowl. Discov.* 29(6), 1733-1782 (2015)
  - Hardt M., Price E., Srebro N.: Equality of opportunity in supervised learning. In: NIPS'16 (2016)

Post-processing: Manipulation

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

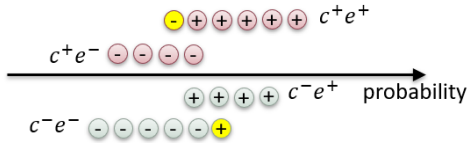
Massaging

- 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



$$RD = \frac{6}{10} - \frac{4}{10} = 0.2$$

  3. Flip the decision of individuals that close to the bottom/top



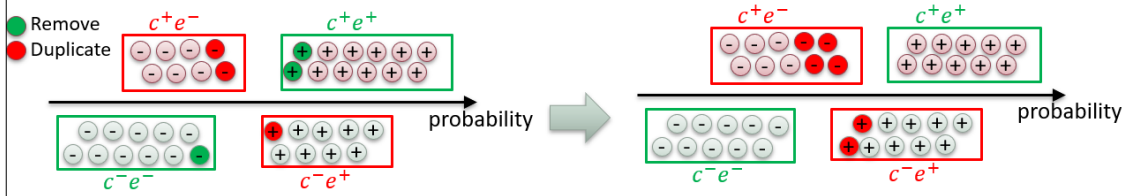
$$RD = \frac{5}{10} - \frac{5}{10} = 0.0$$

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

48

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



$$RD = \frac{12}{20} - \frac{9}{20} = 0.15$$

$$RD = \frac{10}{20} - \frac{10}{20} = 0$$

Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33(1), 1–33 (2012)

49

Decision Theory for Discrimination-aware Classification

- Hypothesis: discrimination decisions are made close to the decision boundary
  - Reject Option based Classification (ROC)
    - For probabilistic classifiers, $P(\hat{y} = +|\boldsymbol{x}) = 0.5 \Rightarrow$ discrimination.
    - Define $\max[P(\hat{y} = +|\boldsymbol{x}) - P(\hat{y} = -|\boldsymbol{x})] < \theta$ as the critical region.
    - Relabel the prediction of individuals in the critical regions.
  - Discrimination-Aware Ensemble (DAE)
    - For ensemble methods, larger disagreement of classifiers $\Rightarrow$ discrimination
    - Define disagreement $disagr_{\mathcal{D}} = \frac{|\{X_i|\exists(j,k)\ \mathcal{F}_j(X_i) \neq \mathcal{F}_k(X_i)\}|}{|\{X_i\}|}$
    - Relabel the prediction of individuals with large disagreement

Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: ICDM'12 (2012)

Construct Equalized Odds Predictor

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

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

A: Protected attribute
Y: Label

  - 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:

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

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

$$\forall_{y,a} 0 \leqslant p_{ya} \leqslant 1$$

Make $\tilde{Y}_p$ close to Y

Ensure equalized odds

Ensure that $\tilde{Y}_p$ is derived from $\hat{Y}$

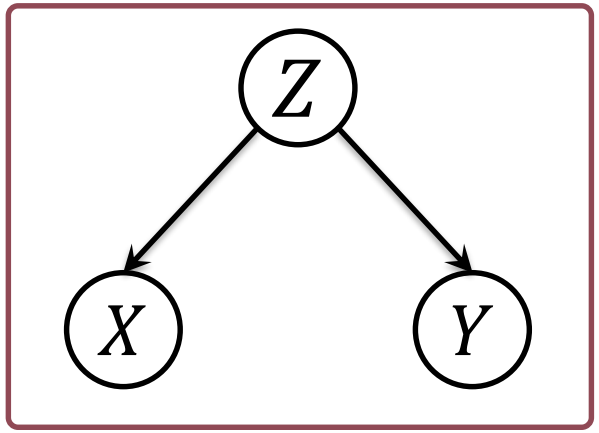
Hardt M., Price E., Srebro N.: Equality of opportunity in supervised learning. In: NIPS'16 (2016)

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

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

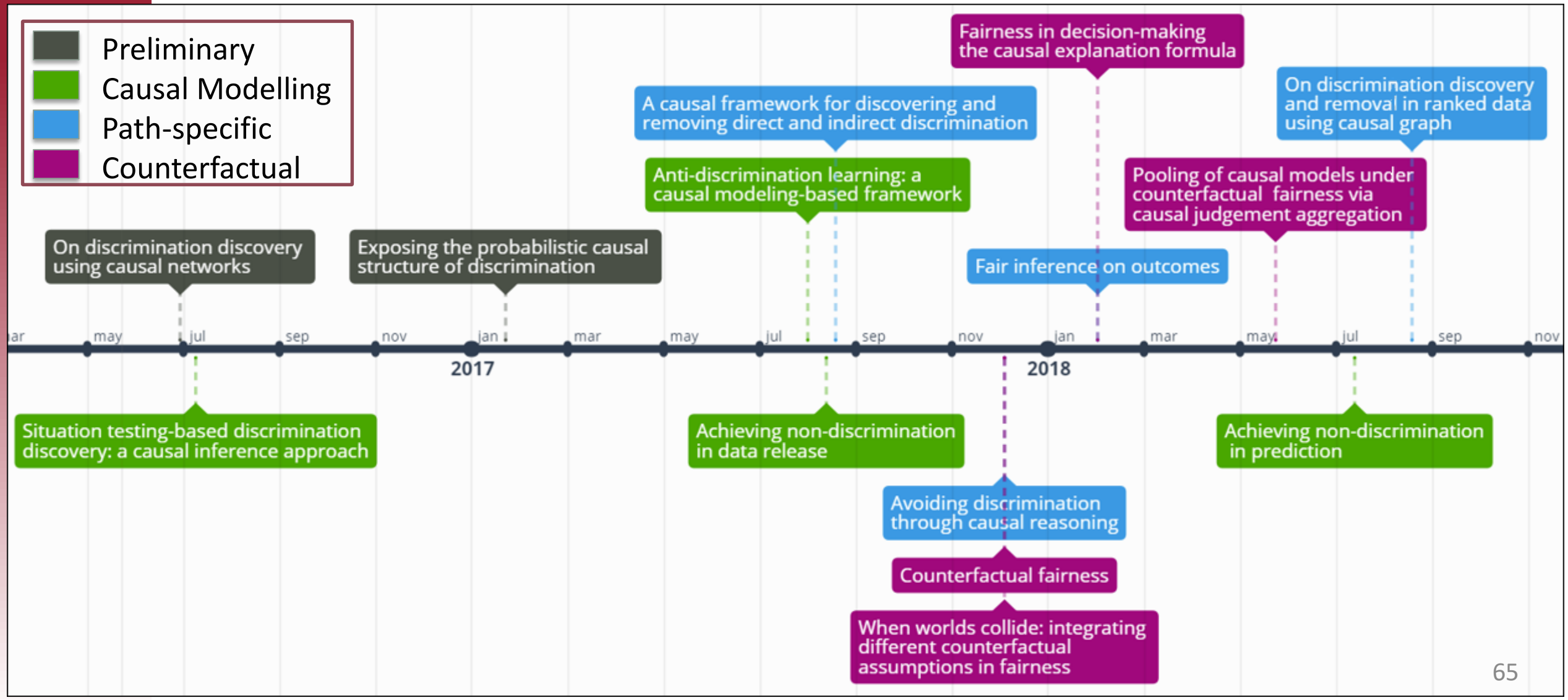


- In order to imply causation, a true experiment must be performed where subjects are randomly assigned to different conditions.

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



Causal based Discrimination Discovery

- Preliminary work
  - Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. *Int. J. Data Sci. Anal.* 3(1), 1–21 (2017)
  - Zhang, L., Wu, Y., Wu, X.: On discrimination discovery using causal networks. In: SBP-BRiMS 2016 (2016)
- Causal-modeling-based
  - Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)
  - Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)
  - Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. *Int. J. Data Sci. Anal.* 4(1), 1-16 (2017)
  - Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. In: IJCAI'18 (2018)

Causal based Discrimination Discovery

- Path-specific-effect-based
  - Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
  - Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NIPS'17 (2017)
  - Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI'18 (2018)
  - Wu, Y., Zhang, L., Wu, X.: On discrimination discovery and removal in ranked data using causal graph. In: SIGKDD'18 (2018)

Causal based Discrimination Discovery

- Counterfactual-based
  - Kusner, M.J., Loftus, J., Russell, C, Silva, R.: Counterfactual fairness. In: NIPS'17 (2017)
  - Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: NIPS'17 (2017)
  - Zhang, J., Bareinboim, E.: Fairness in decision-making – the causal explanation formula. In: AAAI'18 (2018)
  - Zennaro, F.M., Ivanovska, M.: Pooling of causal models under counterfactual fairness via causal judgement aggregation. Preprint (2018)

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

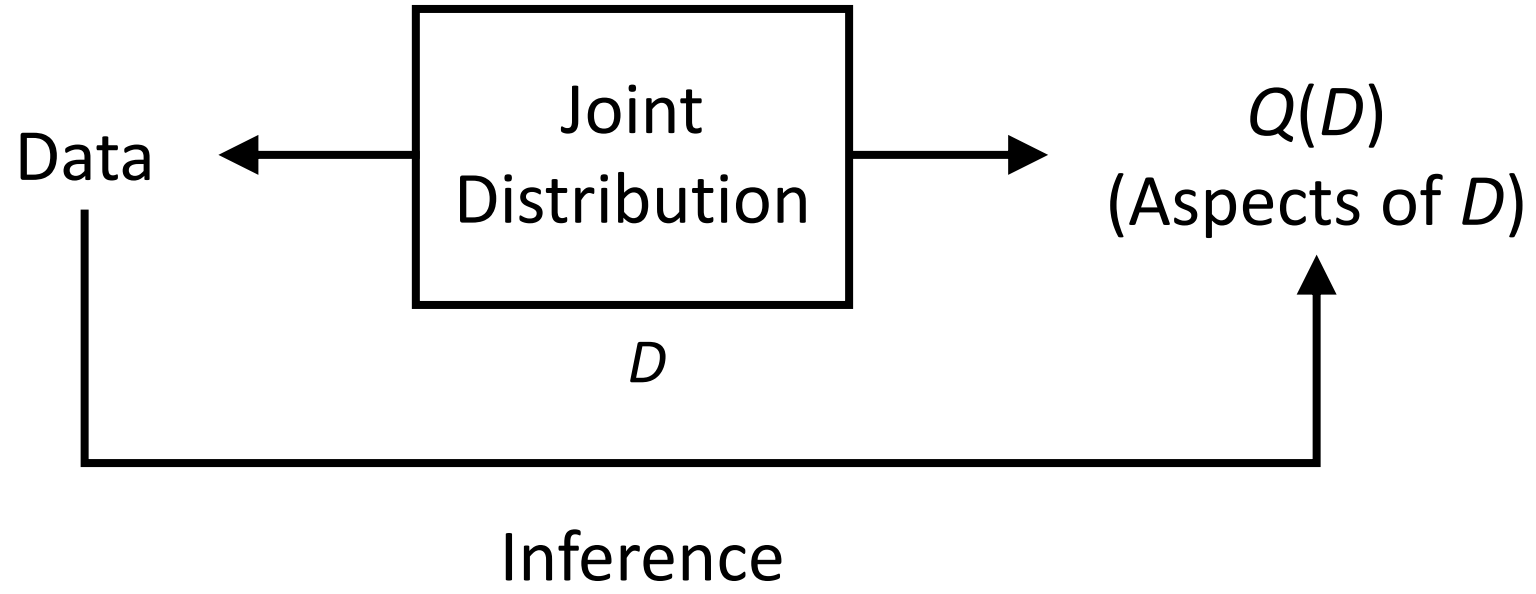
How to construct causal graph is omitted.

Lessons of Causal Inference (*Pearl*)

1. No cause in — no cause out
2. $\left.\begin{matrix}\text{Data} \\ \text{Causal assumptions/knowledge}\end{matrix}\right\} \Rightarrow$ causal conclusions
3. Causal assumptions/knowledge cannot be expressed in the mathematical language of standard statistics.
4. Need ways of encoding causal assumptions/knowledge mathematically and test their implications.

From Statistics to Causal Modeling

- Traditional statistical inference paradigm:



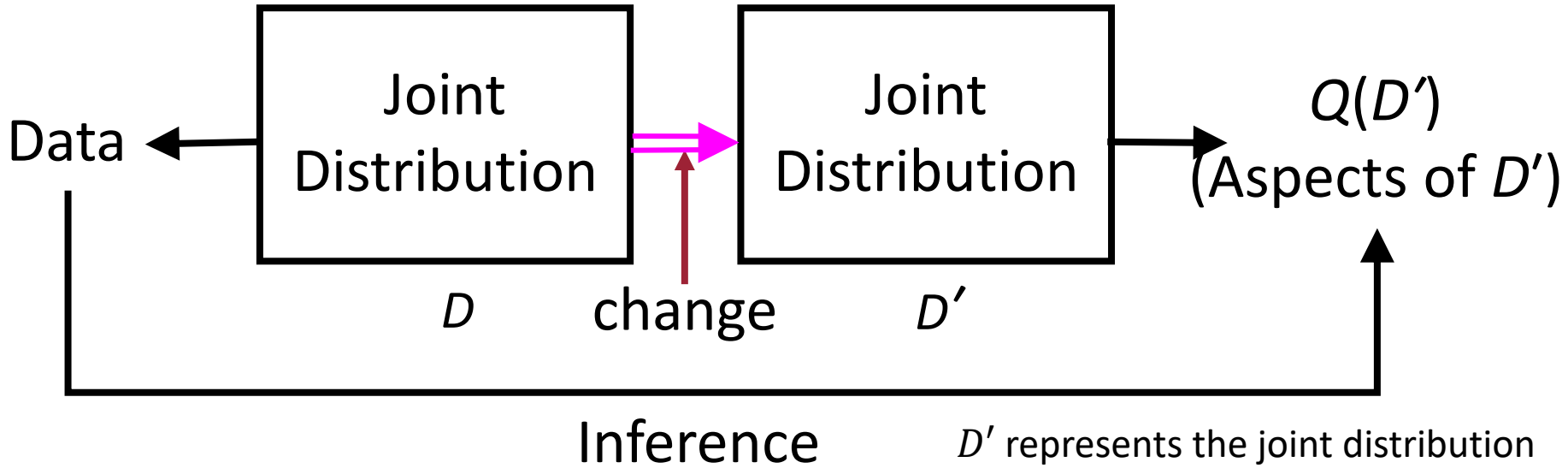
- What is the probability of getting Grade A for the students who study 1 hour each day?

Estimate $Q(D) = P_D(E = \text{'}A\text{'} \mid H = 1)$

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.

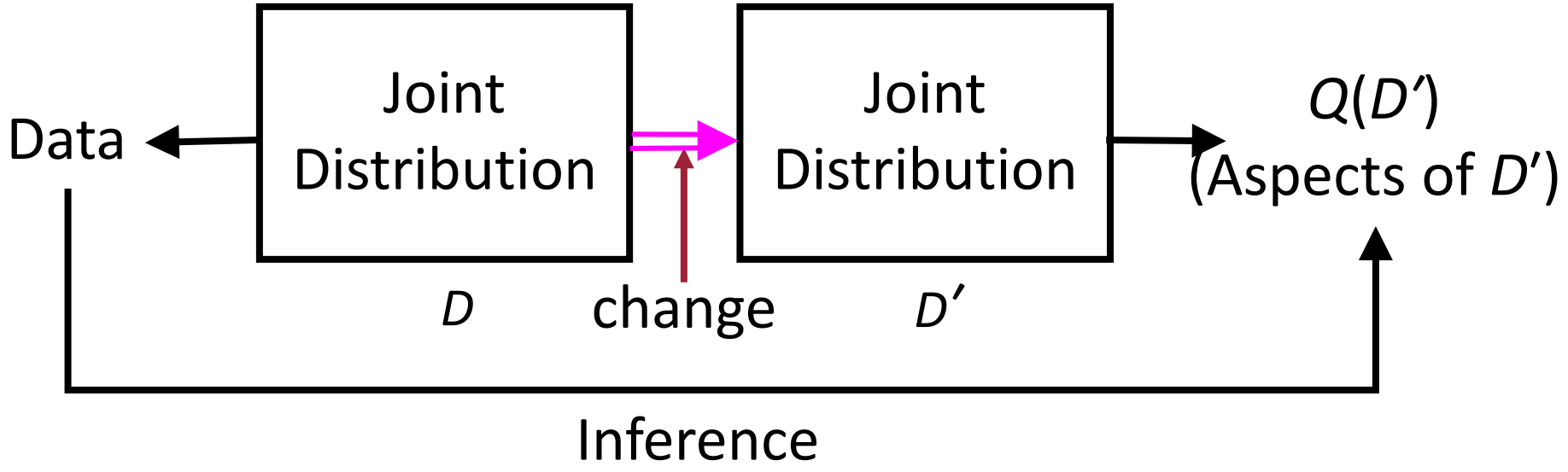


D' represents the joint distribution after adopting the new policy.

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

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.

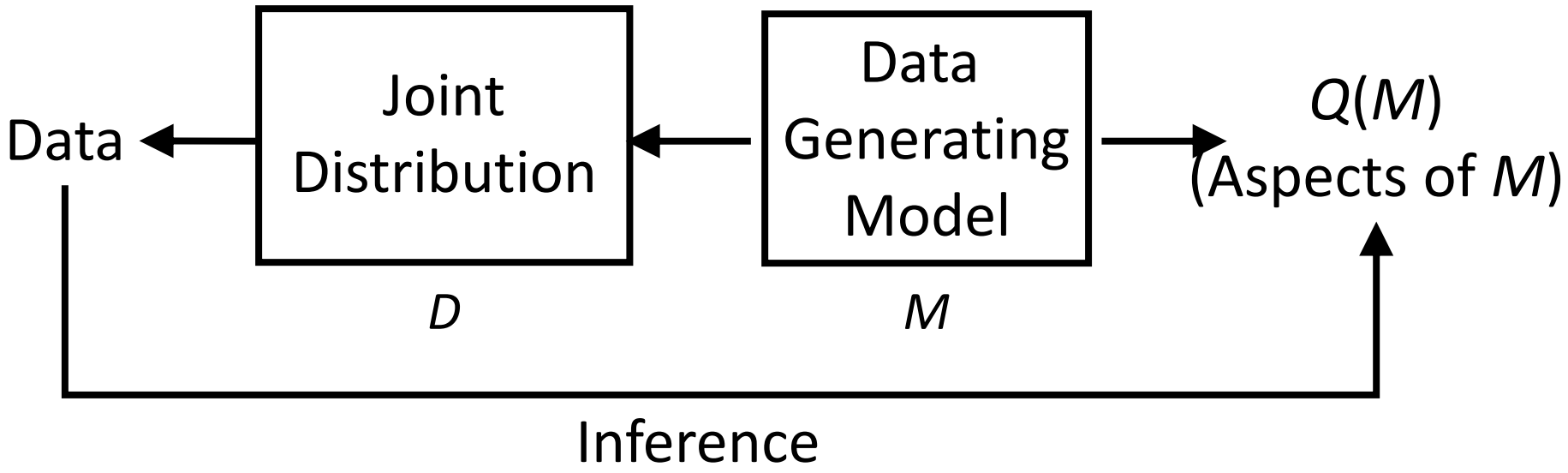


$$P_{D'}(E = \text{`}A\text{'}) \neq P_D(E = \text{`}A\text{'} \mid H = 2)$$

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

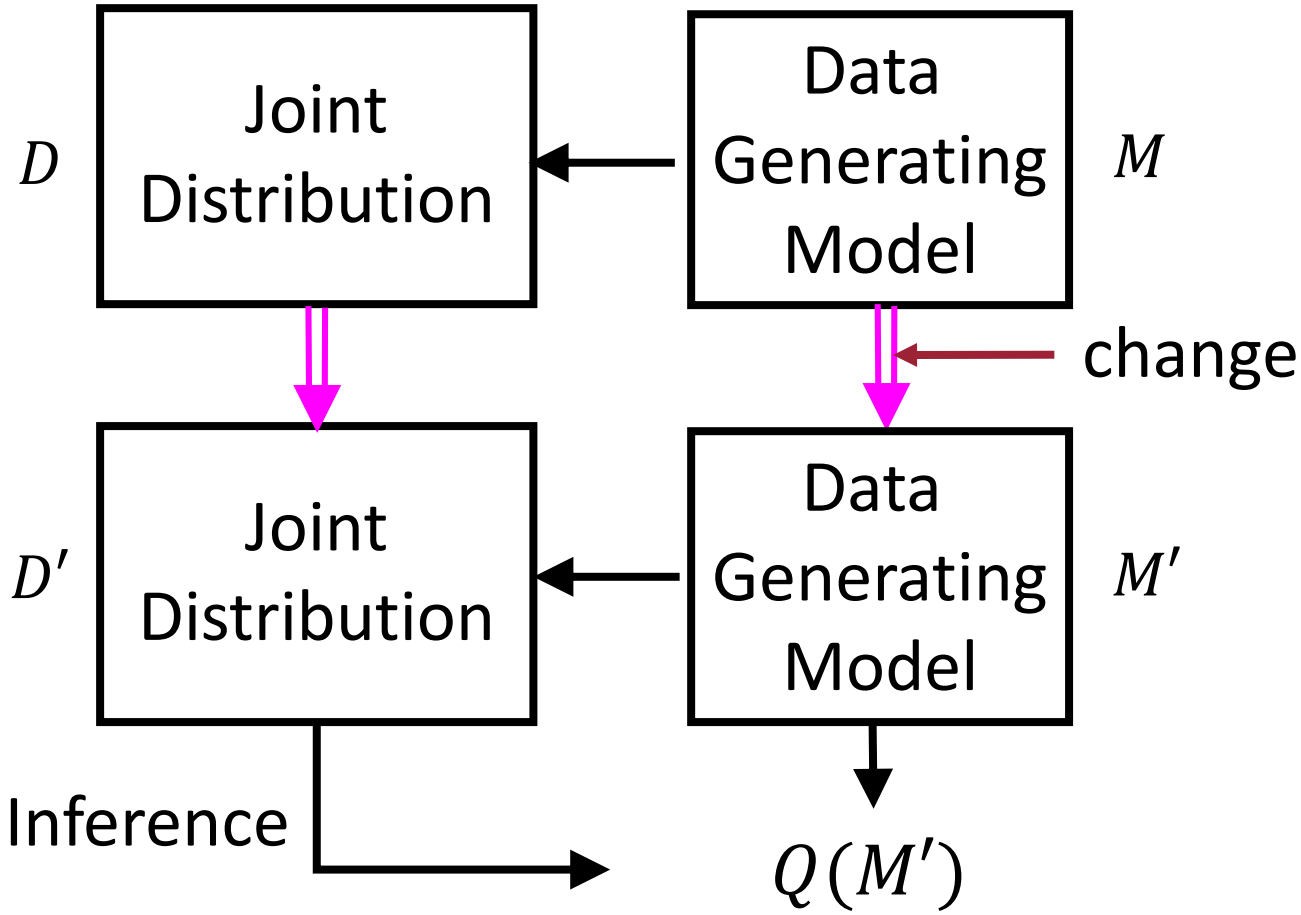


M – Data generation model that encodes the causal assumptions/knowledge.

D – model of data, M – model of reality

From Statistics to Causal Modeling

- Causal inference



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 $\mathcal{M} =< \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F} >$, where
  - $\boldsymbol{U}$ is a set of exogenous (hidden) variables whose values are determined by factors outside the model;
  - $\boldsymbol{V} = \{X_1, \cdots, X_i, \cdots\}$ is a set of endogenous (observed) variables whose values are determined by factors within the model;
  - $\boldsymbol{F} = \{f_1, \cdots, f_i, \cdots\}$ is a set of deterministic functions where each f_i is a mapping from $\boldsymbol{U} \times (\boldsymbol{V} \setminus X_i)$ to X_i. Symbolically, f_i can be written as

$$x_i = f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i)$$

where $\boldsymbol{pa}_i$ is a realization of X_i's parents in $\boldsymbol{V}$, i.e., $\boldsymbol{Pa}_i \subseteq \boldsymbol{V}$, and $\boldsymbol{u}_i$ is a realization of X_i's parents in $\boldsymbol{U}$, i.e., $\boldsymbol{U}_i \subseteq \boldsymbol{U}$.

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 represent the variables $\boldsymbol{U} \cup \boldsymbol{V}$ 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 $\boldsymbol{Pa}_i \cup \boldsymbol{U}_i$ 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 $\boldsymbol{V} = \{I, H, W, E\}$

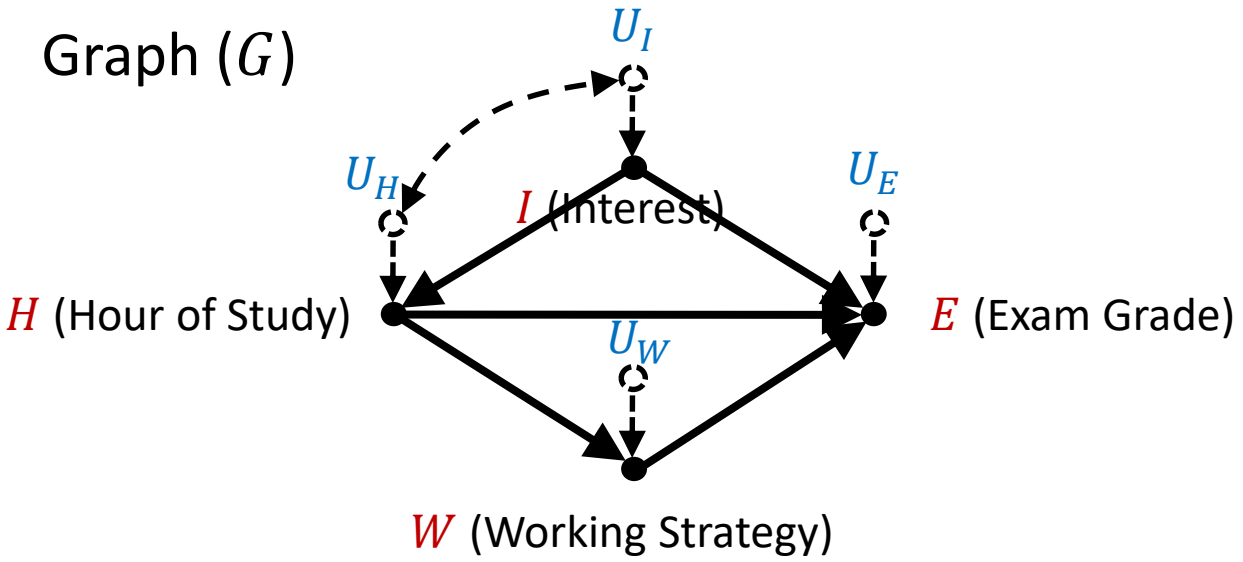
Hidden Variables $\boldsymbol{U} = \{U_I, U_H, U_W, U_E\}$

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.

Graph (G)



Markovian Model

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

↓ Equivalent expression

Each node X is conditionally independent of its non-descendants given its parents $\boldsymbol{Pa}_X$.

↓

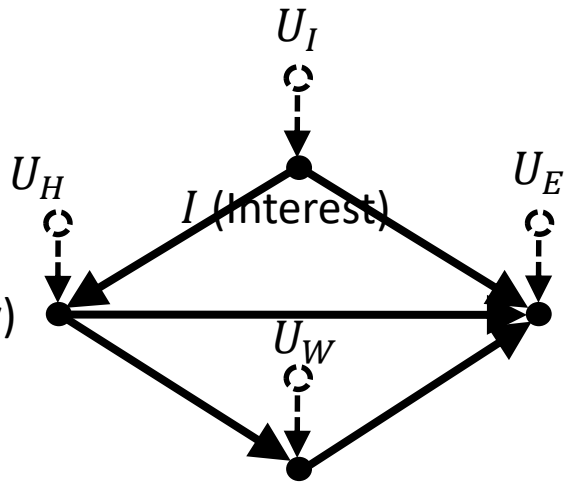
Known as the local Markov condition (e.g., in Bayesian network), or causal Markov condition in the context of causal modeling.

A Markovian Model and Its Graph

Model (M)

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

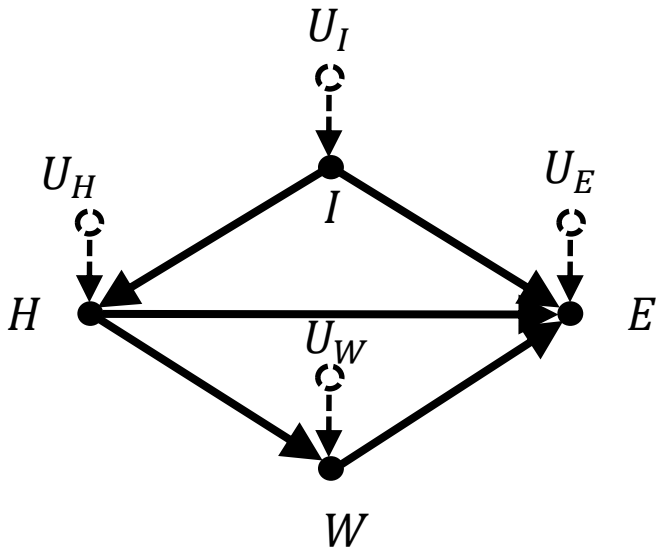
Graph (G)

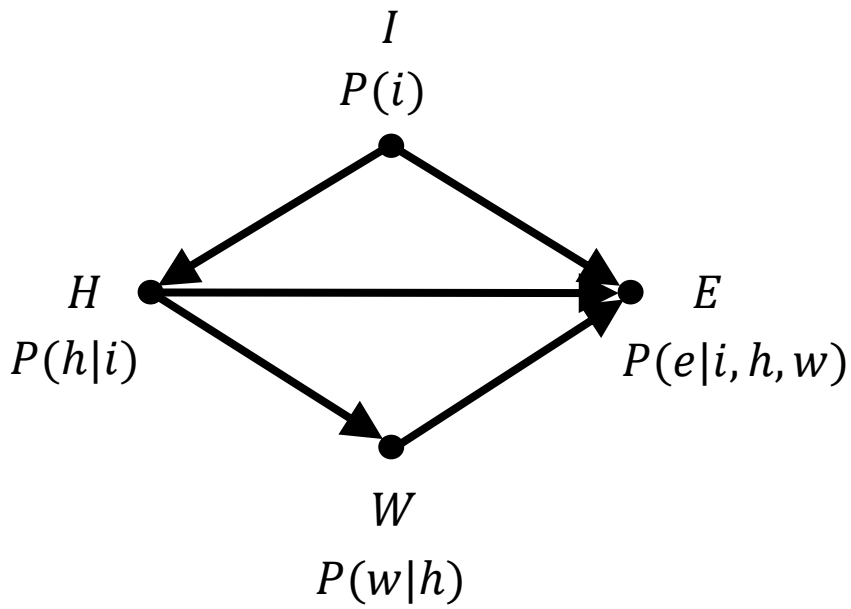


Assume U_I, U_H, U_W, U_E are mutually independent.

Causal Graph of Markovian Model

Each node is associated with a observable conditional probability table (CPT) $P(x_i|\boldsymbol{pa}_i)$





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 $(X \perp Y|Z)_D$

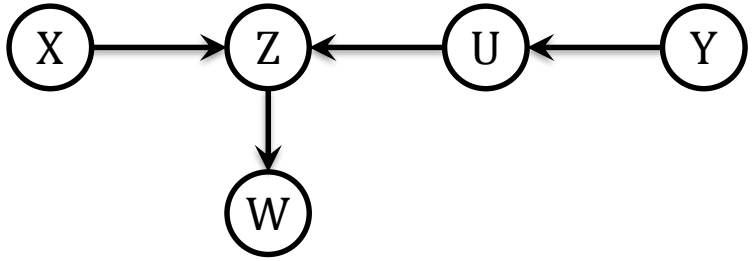
d-Separation

- Definition of *d*-separation
- A path q is said to be blocked by conditioning on a set $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$, or
  - q contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\boldsymbol{Z}$ and such that no descendant of m is in $\boldsymbol{Z}$.
- $\boldsymbol{Z}$ is said to *d*-separate X and Y if $\boldsymbol{Z}$ blocks every path from X to Y, denoted by $(X \perp Y|Z)_G$

d-Separation

- Example (blocking of paths)

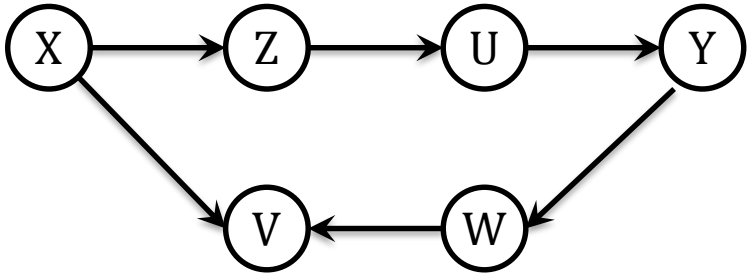
  - Path from X to Y is blocked by conditioning on $\{U\}$ or $\{Z\}$ or both $\{U, Z\}$
- Example (unblocking of paths)



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

d-Separation

- Example (*d*-separation)

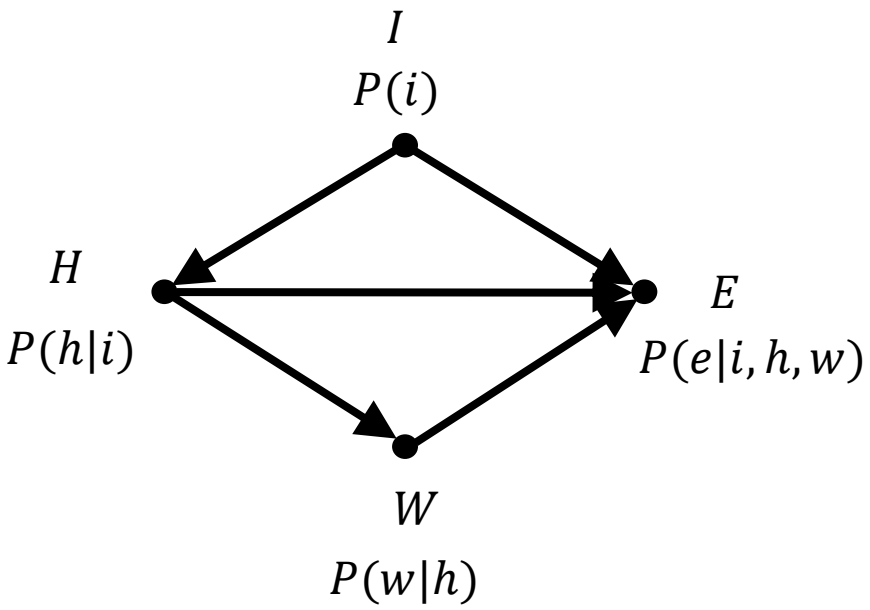


- We have following *d*-separation relations
  - $(X \perp Y|Z)_G, (X \perp Y|U)_G, (X \perp Y|ZU)_G$
  - $(X \perp Y|ZW)_G, (X \perp Y|UW)_G, (X \perp Y|ZUW)_G$
  - $(X \perp Y|VZUW)_G$
- However we do NOT have
  - $(X \perp Y|VZU)_G$

Factorization Formula

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

$$P(\boldsymbol{v}) = \prod_{X \in \boldsymbol{V}} P(x|\boldsymbol{pa}_X)$$



$$P(i,h,w,e) = P(i)P(h|i)P(w|h)P(e|i,h,w)$$

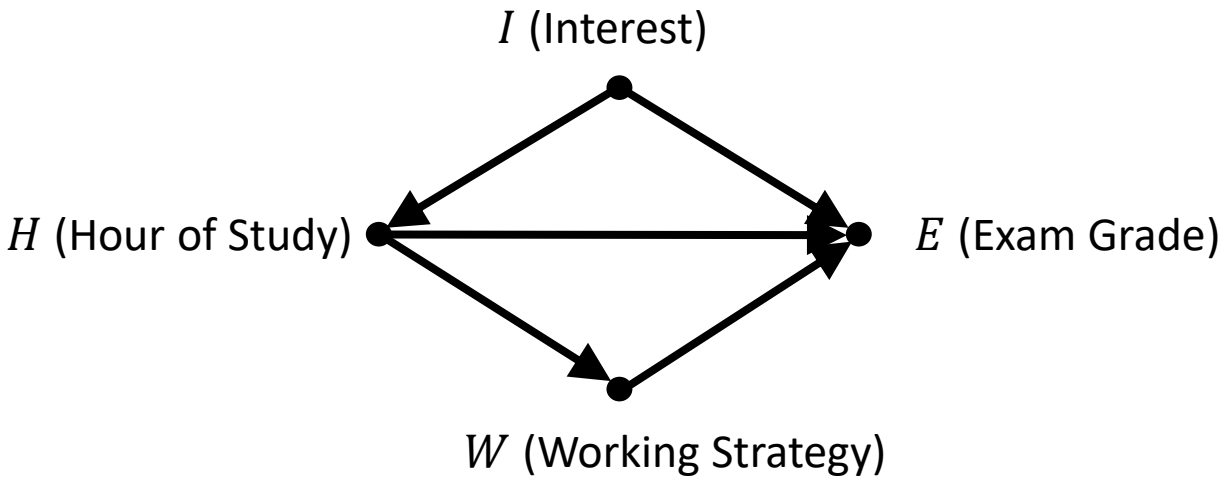
$$P(e) = \sum_{I,H,W} P(i)P(h|i)P(w|h)P(e|i,h,w)$$

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 *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 $P(E = \text{'}A\text{'} | H = 1)$

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 $P(E = \text{`}A\text{'} | H = 2)$, i.e., the conditional probability of getting grade A given study hour equals to 2.

Intervention and *do*-Operator

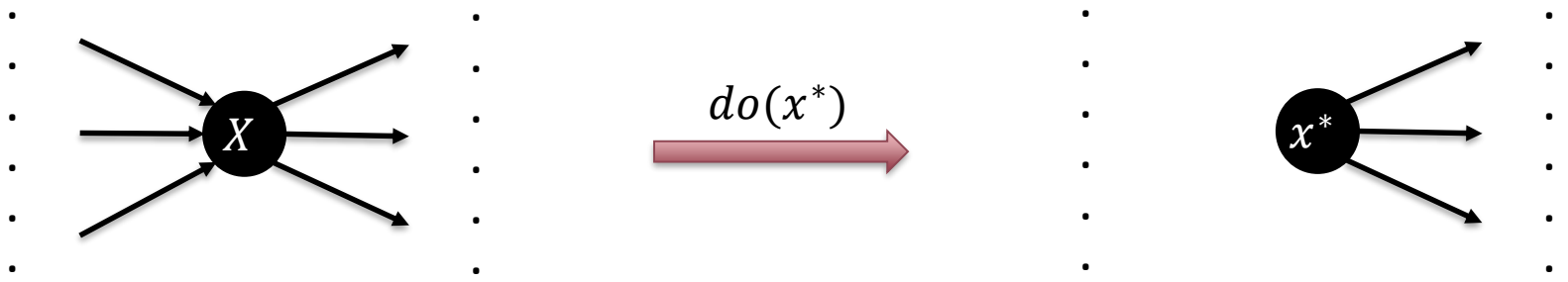
- The basic operation of manipulating a causal model.
  - Simulate the manipulation of the physical mechanisms by some physical interventions or hypothetical assumptions.
  - Forces some observed variables $\boldsymbol{X} \in \boldsymbol{V}$ to take certain constants $\boldsymbol{x}$.
- Mathematically formulated as $do(\boldsymbol{X} = \boldsymbol{x})$ or simply $do(\boldsymbol{x})$.
- For an observed variable Y disjoint with $\boldsymbol{X}$, its interventional variant under intervention $do(\boldsymbol{x})$ is denoted by $Y_{\boldsymbol{X} \leftarrow \boldsymbol{x}}$ or $Y_{\boldsymbol{x}}$.
- The effect of intervention on all other observed variables $\boldsymbol{Y} = \boldsymbol{V} \backslash \boldsymbol{X}$ is represented by the post-intervention distribution of $\boldsymbol{Y}$.
  - Denoted by $P(\boldsymbol{Y} = \boldsymbol{y} | do(\boldsymbol{X} = \boldsymbol{x}))$ or simply $P(\boldsymbol{y} | do(\boldsymbol{x}))$;
  - Or equivalently $P(\boldsymbol{Y}_{\boldsymbol{X} \leftarrow \boldsymbol{x}} = \boldsymbol{y})$ or simply $P(\boldsymbol{y}_{\boldsymbol{x}})$.

Intervention and *do*-Operator

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

$$\mathcal{M}: \quad x = f_X(\boldsymbol{pa}_X, \boldsymbol{u}_X) \xrightarrow{do(x^*)} \mathcal{M}_{x^*}: \quad x = x^*$$

- From the point of view of the causal graph, performing $do(x^*)$ 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 $P(\boldsymbol{y}|do(\boldsymbol{x}))$ can be calculated from the CPTs, known as the truncated factorization:

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

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

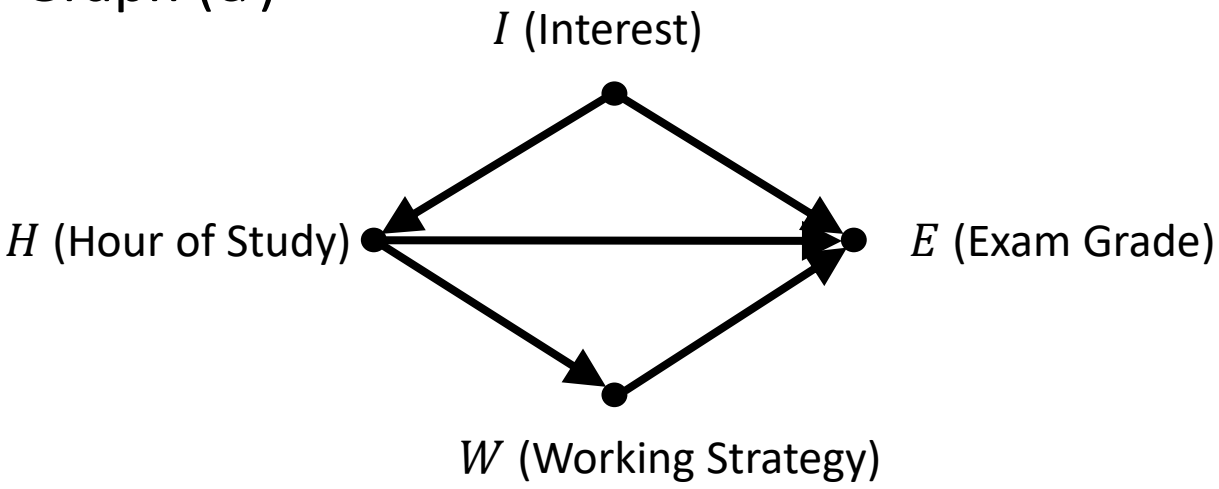
- Specifically, for a single attribute Y given an intervention on a single attribute X,

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V} \setminus \{X,Y\} \\ Y=y}} \prod_{V \in \boldsymbol{V} \setminus \{X\}} P(v|\boldsymbol{Pa}_V)\delta_{X \leftarrow x}$$

Intervention Example

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

Graph (G)



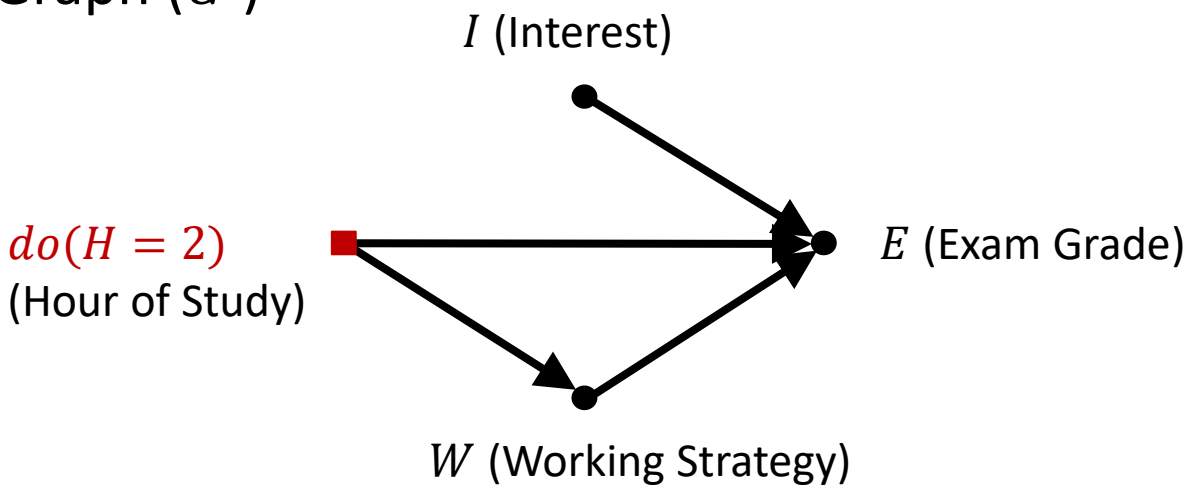
Model (M)

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

Intervention Example

- What is the probability of getting grade A if we change the study hour to 2, i.e., $do(H = 2)$?

Graph (G')



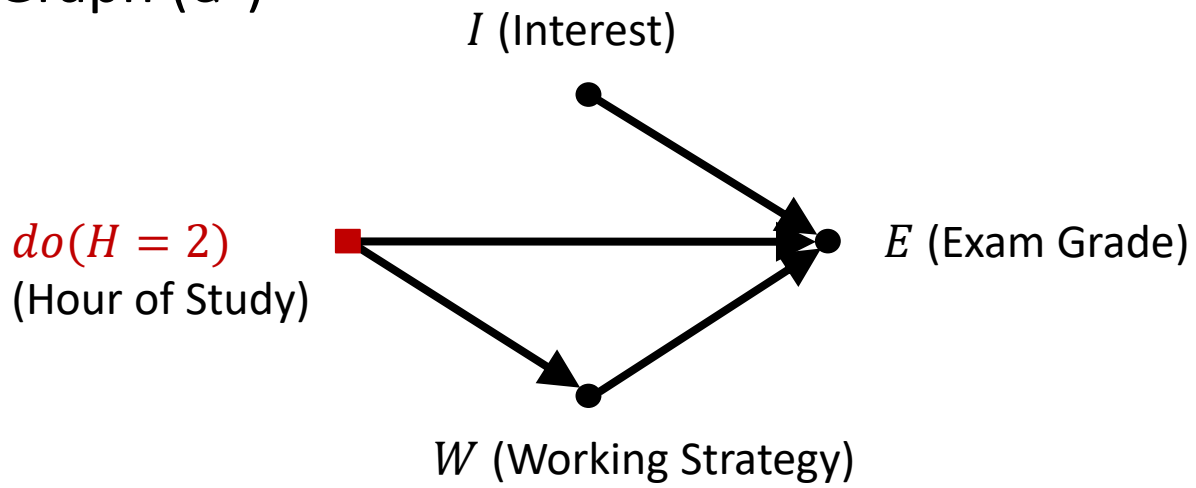
Model (M')

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

- Find $P(E = \text{'}A\text{'} | do(H = 2))$

Intervention Example

Graph (G')



Model (M')

$i = f_I(u_I)$

$h = 2$

$w = f_W(h, u_W)$

$e = f_E(i, h, k, u_E)$

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V}\setminus\{X,Y\} \\ Y=y}} \prod_{V\in\boldsymbol{V}\setminus\{X\}} P(v|\boldsymbol{Pa}_V)\delta_{X\leftarrow x}$$

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

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

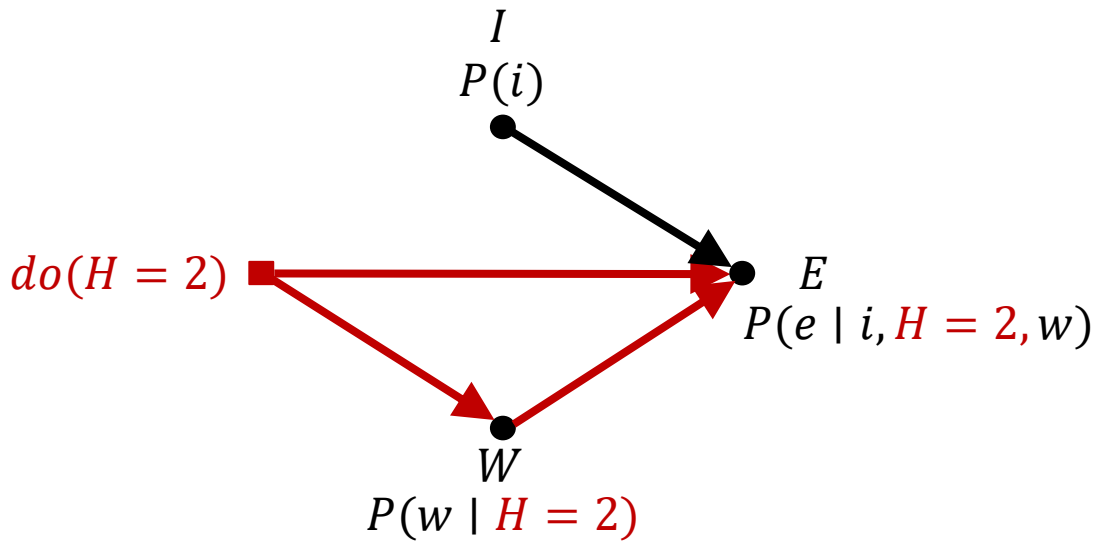
Total Causal Effect

- 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.

$P(e \mid do(H = 2))$



$$TE(H = 2, H = 1)$$
$$= P(E = \text{'A'}|do(H = 2)) - P(E = \text{'A'}|do(H = 1))$$
$$= \sum_{I,W} P(i)P(w|H = 2)P(E = \text{'A'}|i, H = 2, w) - \sum_{I,W} P(i)P(w|H = 1)P(E = \text{'A'}|i, H = 1, w)$$

Path-Specific Effect

- Path-specific effect measures the causal effect transmitted along certain paths.
- Given a subset of causal paths π, the causal effect of X on Y transmitted along π is denoted by

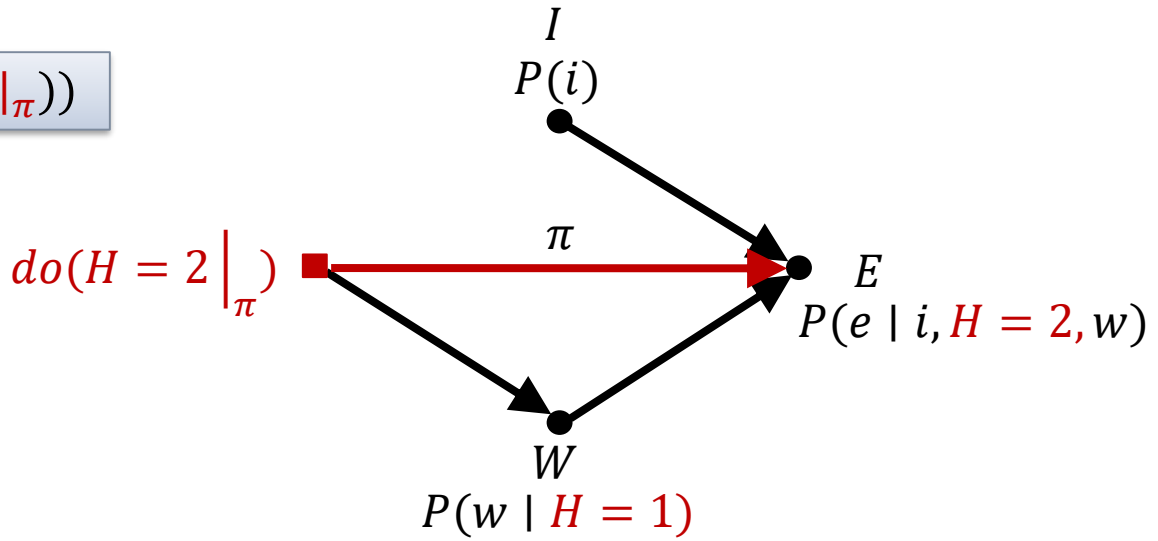
$$SE_{\pi}(x_2, x_1) = P(y|do(x_2|_{\pi})) - P(y|do(x_1))$$

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

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 (π).

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

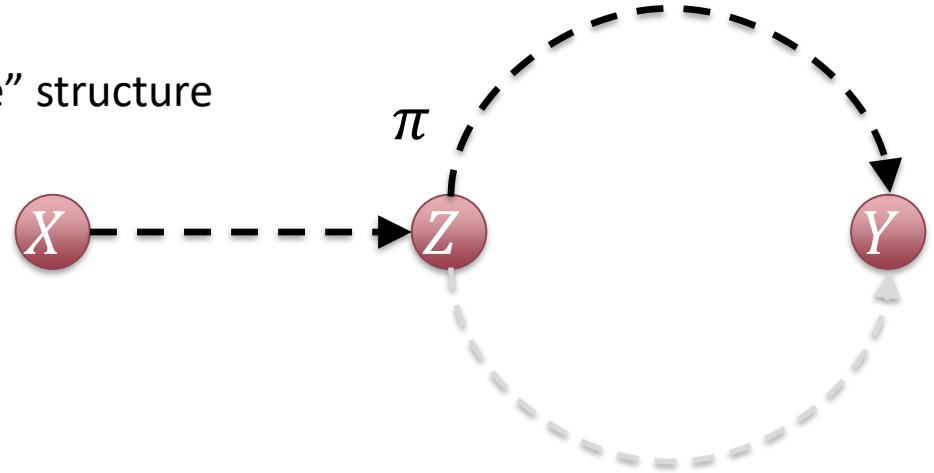


$$SE_{\pi}(H = 2, H = 1)$$
$$= P(E = \mathrm{A}|do(H = 2|_{\pi})) - P(E = \mathrm{A}|do(H = 1))$$
$$= \sum_{I,W} P(i)P(w|H = 2)P(E = A|i, H = 1, w) - \sum_{I,W} P(i)P(w|H = 1)P(E = A|i, H = 1, w)$$

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:

The "kite" structure



$P(y \mid do(x|_{\pi}))$ is non-identifiable in this graph

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

Avin, C., Shpitser, I., Pearl, J.: Identifiability of path-specific effects. In: IJCAI'05 (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 $\boldsymbol{e}$.
- General form of a counterfactual query: "what would we expect the value of Y had X been x, given that we observe $\boldsymbol{E} = \boldsymbol{e}$?"

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

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

$\mathcal{M}$

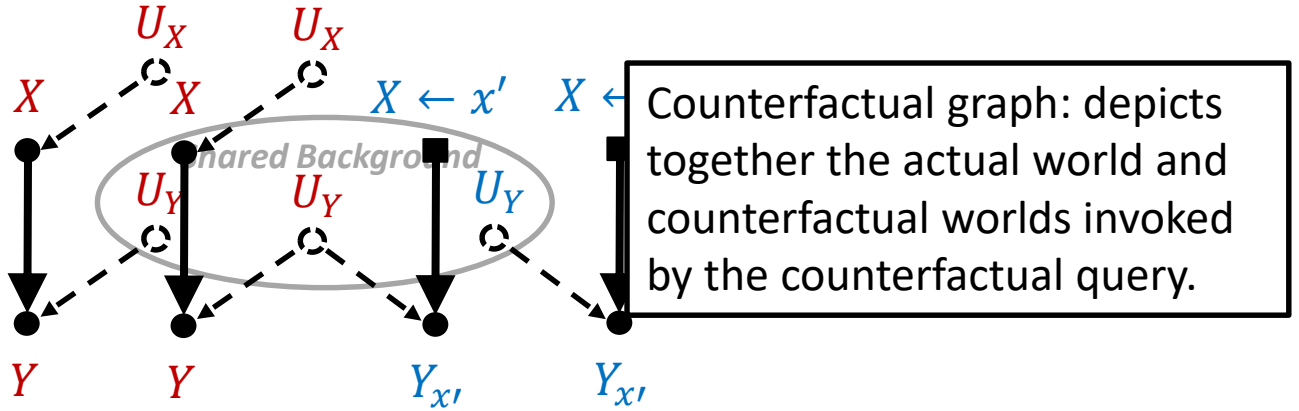
$x = f_X(u_X)$

$y = f_Y(x, u_Y)$

$\mathcal{M}_{x'}$

$x = x'$

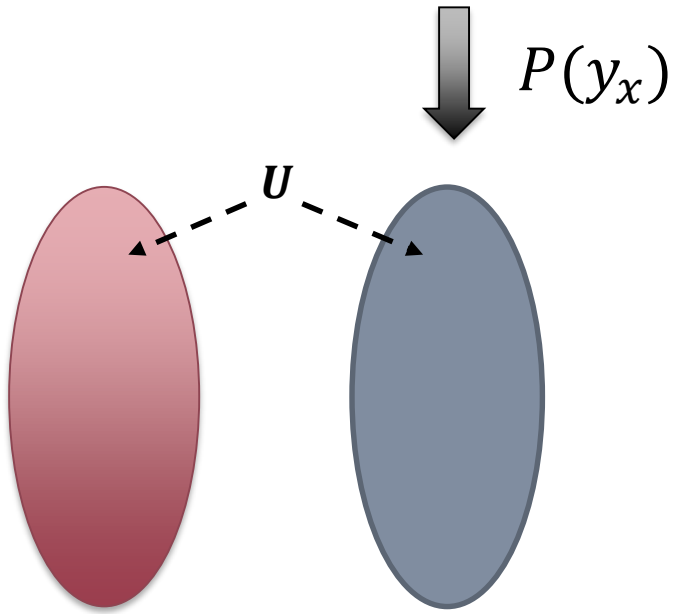
$y = f_Y(x, u_Y)$



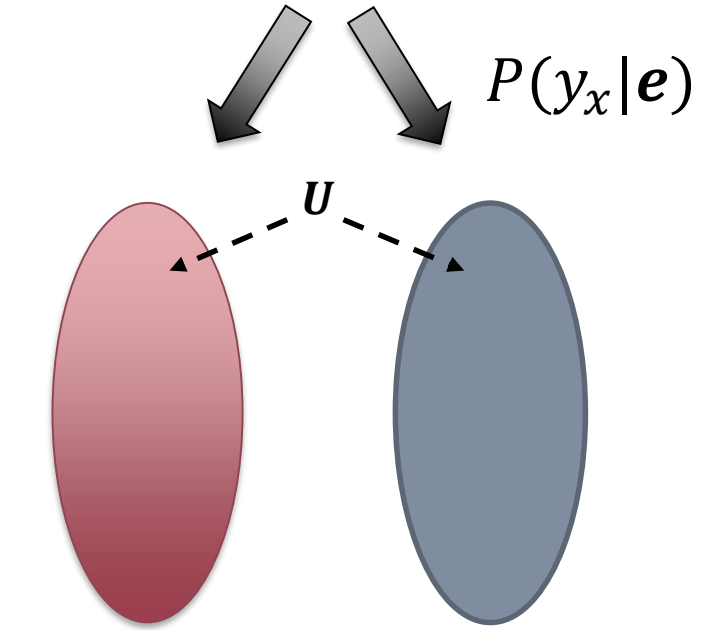
Counterfactual graph: depicts together the actual world and counterfactual worlds invoked by the counterfactual query.

Intervention vs. Counterfactual

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



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



Counterfactual Analysis

$$P(y_x \mid \boldsymbol{e}) = \sum_{\boldsymbol{u}} P(y_x \mid \boldsymbol{e}, \boldsymbol{u}) P(\boldsymbol{u} \mid \boldsymbol{e}) = \sum_{\boldsymbol{u}} P(y_x \mid \boldsymbol{u}) P(\boldsymbol{u} \mid \boldsymbol{e})$$

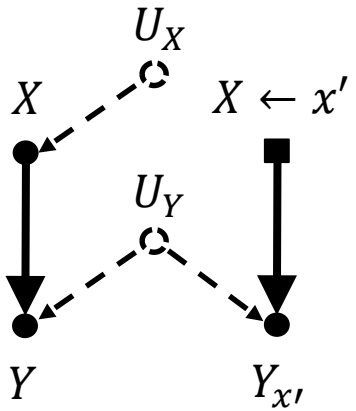
- Principled procedure for computing $P(y_x \mid \boldsymbol{e})$:
  - **Abduction**: Update $P(\boldsymbol{u})$ by the evidence $\boldsymbol{e}$ to obtain $P(\boldsymbol{u}|\boldsymbol{e})$;
  - **Action**: Perform intervention $do(x)$ on causal model $\mathcal{M}$ to obtain $\mathcal{M}_x$;
  - **Prediction**: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

- Usually don't know $P(\boldsymbol{u})$.

Identifiability of Counterfactual

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

- "W-graph": the simplest non-identifiable counterfactual graph structure.



$P(y'_{x'} \mid x, y)$ is non-identifiable for any causal model

Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* 9(Sep), 1941-1979 (2008)

Identifiability of Counterfactual

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

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

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

Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* 9(Sep), 1941-1979 (2008)
Pearl, J.: A linear "microscope" for interventions and counterfactuals. *Journal of Causal Inference*, *5*(1). (2017)

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 $\boldsymbol{V} = \{C, \overbrace{\cdots, X_i, \cdots}^{\boldsymbol{X}}, E\}$

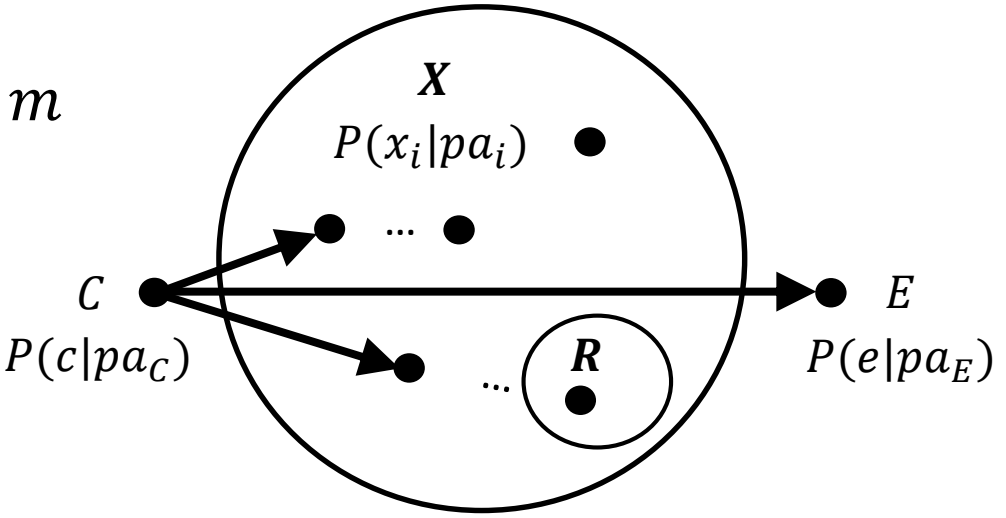
Hidden Variables $\boldsymbol{U}$

Causal Model M

$$c = f_C(pa_C, \boldsymbol{u}_C)$$
$$x_i = f_i(pa_i, \boldsymbol{u}_i), i = 1, \cdots, m$$
$$e = f_E(pa_E, \boldsymbol{u}_E)$$

Observed Variables Hidden Variables

Causal Graph G



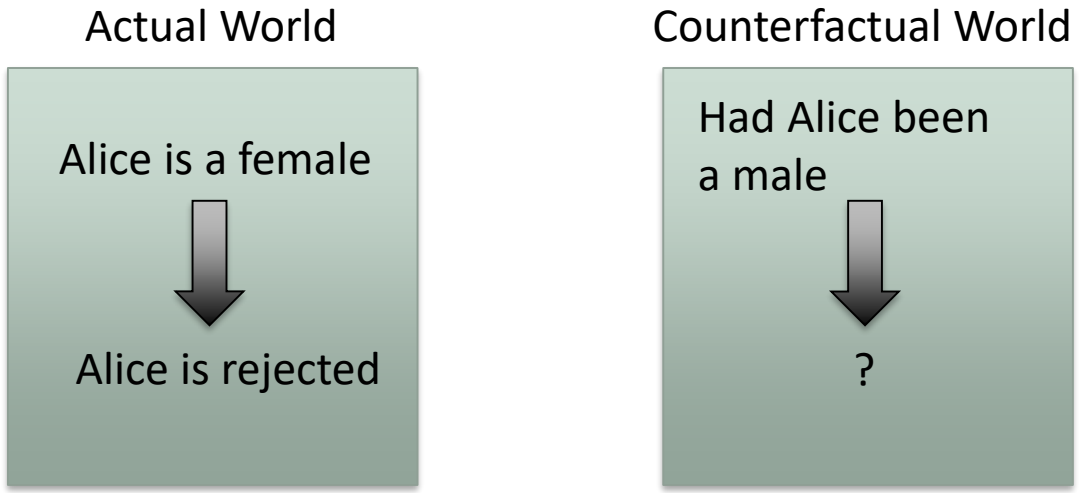
$\boldsymbol{U}_C, \cdots, \boldsymbol{U}_i, \cdots, \boldsymbol{U}_E$ are mutually independent
(Markovian Assumption)

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.

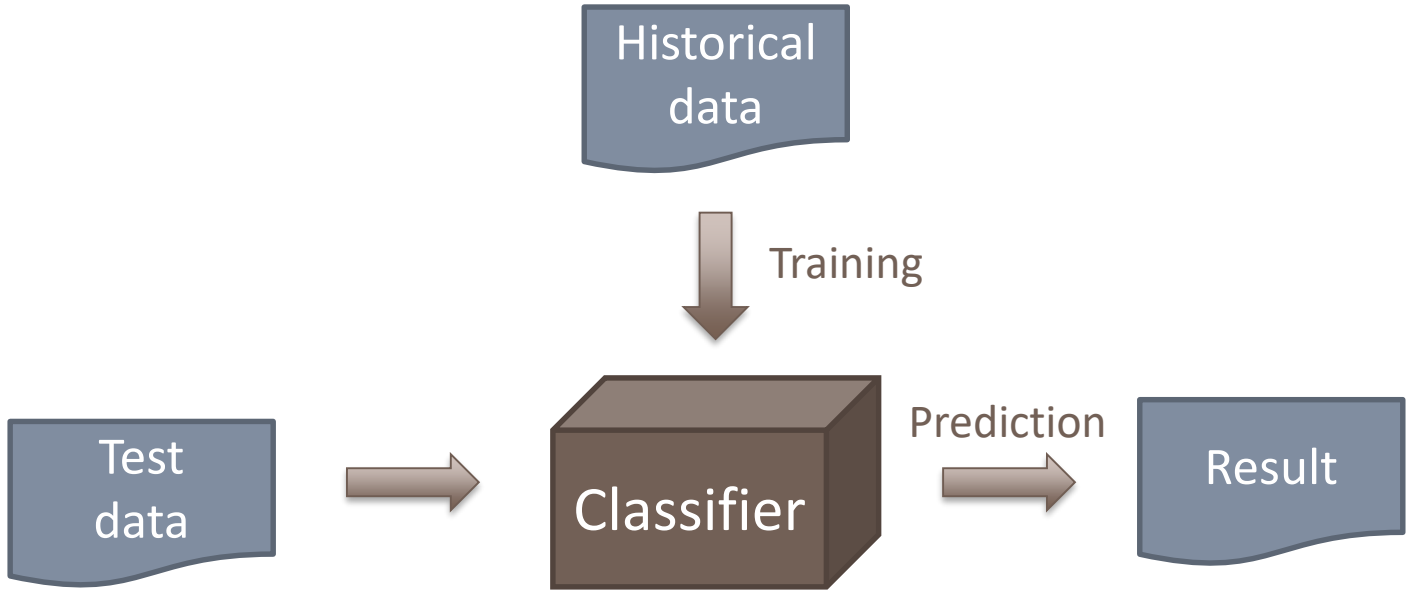
Motivating Examples (ME2)

- 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 extend this assumption holds?

Motivating Examples (ME4)

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

gender	female	male
admission (%)	37%	47%

$P(e^+|c^+) - P(e^+|c^-) = 0.1$

major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	20%	20%	50%	50%	40%	40%	70%	70%

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

gender	female	male
admission (%)	43%	43%

$P(e^+|c^+) - P(e^+|c^-) = 0$

major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	30%	36%	50%	40%	40%	45%	60%	50%

$P(e^+|c^+, \{\text{CS}, \text{L}\}) - P(e^+|c^-, \{\text{CS}, \text{L}\}) = 0.06$

Motivating Examples (ME5)

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

No.	gender	major	score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject

- 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 $\boldsymbol{R}$: non-protected attributes that can cause indirect discrimination.

Direct and Indirect Discrimination Discovery and Removal

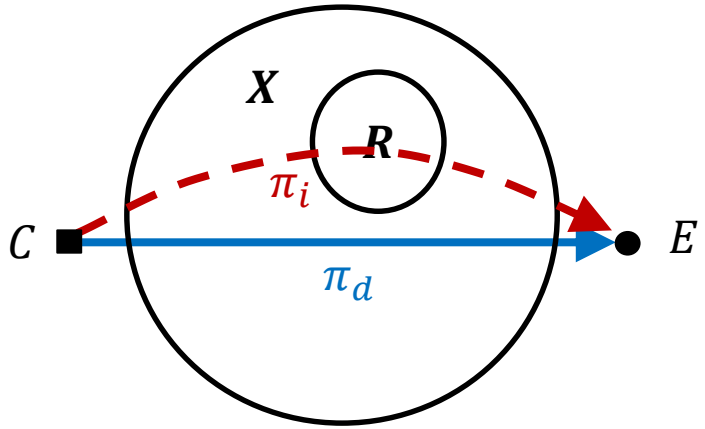
- How to deal with indirect discrimination due to redlining attributes?

- Modeling direct and indirect discrimination using the causal model.
- Quantitative discrimination measure and criterion.
- Algorithm for removing direct and indirect discrimination from a dataset.

Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI'18 (2018)
Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NIPS'17 (2017)

Modeling Discrimination as Path-Specific Effects

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



Quantitative Measuring

- π_d-specific effect:

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

$\boldsymbol{Q}$: E's parents except C

- π_i-specific effect:

$$SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{v}'} \Big(P(e^+|c^-, \mathbf{q}) \prod_{G \in \mathbf{S}_{\pi_i}} P(g|c^+, pa_G \backslash \{C\}) \prod_{H \in \bar{\mathbf{S}}_{\pi_i} \backslash \{E\}} P(h|c^-, pa_H \backslash \{C\}) \prod_{O \in \mathbf{V} \backslash Ch_C} P(o|pa_O) \Big) - P(e^+|c^-)$$

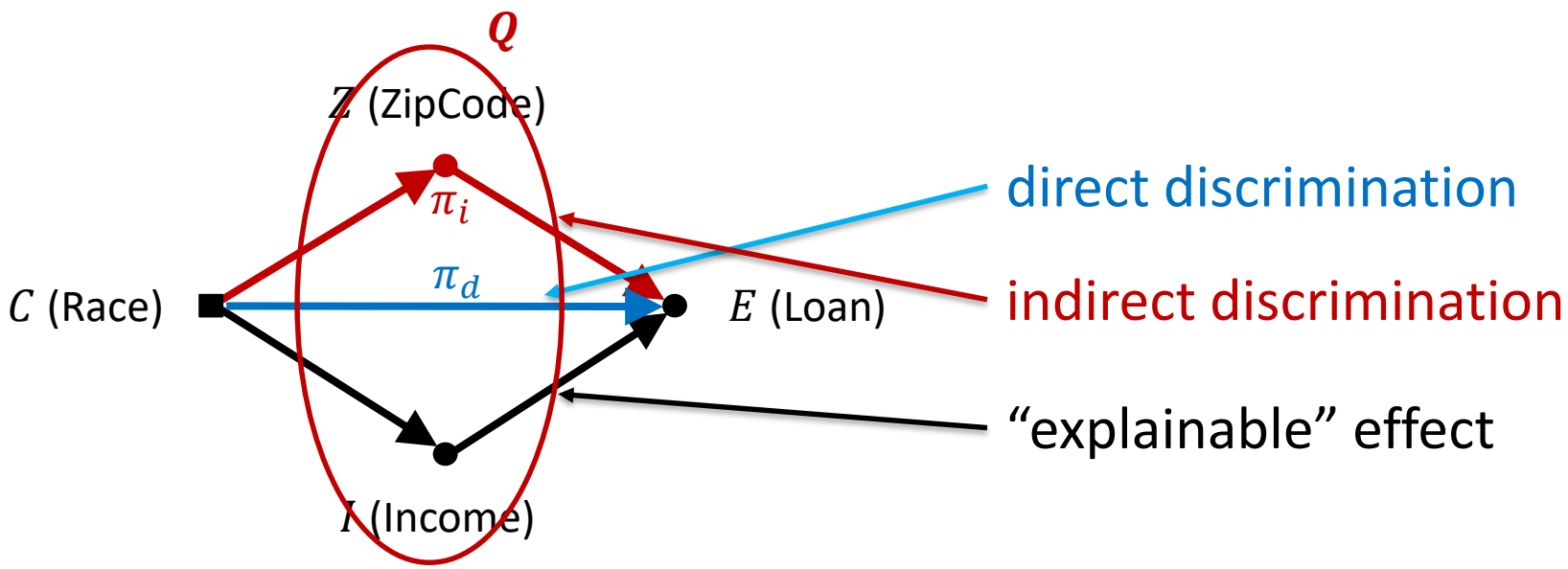
$\boldsymbol{S}_{\pi_i}$: C's children that lie on paths in π_i

$\overline{\boldsymbol{S}}_{\pi_i}$: C's children that don't lie on paths in π_i

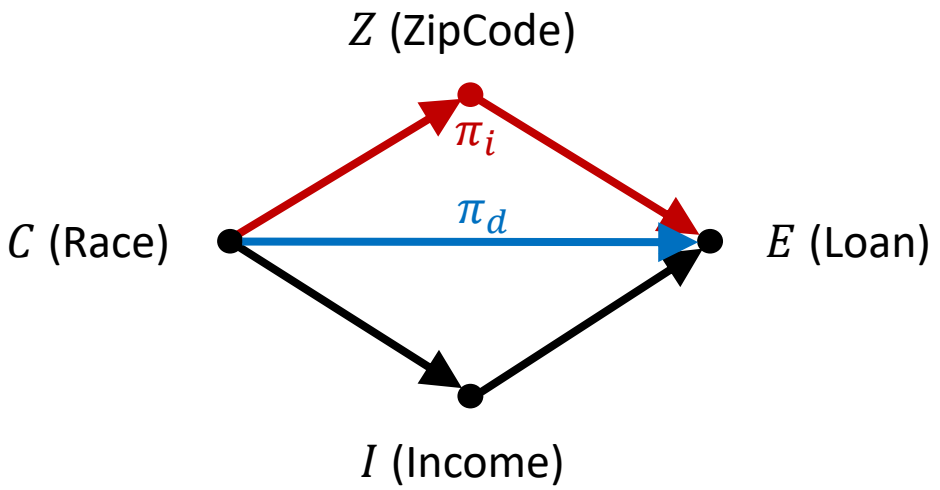
Illustrative Example

- 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



Illustrative Example



$$SE_{\pi_d}(c^+, c^-) = \sum_{Z,I} \big(P(e^+|c^+, z, i) - P(e^+|c^-, z, i)\big) P(z|c^-) P(i|c^-)$$

$$SE_{\pi_i}(c^+, c^-) = \sum_{Z,I} P(e^+|c^-, z, i) \big(P(z|c^+) - P(z|c^-)\big) P(i|c^-)$$

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

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

Total Causal Effect vs. Path-Specific Effect

- For any π_d and π_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 π_i contains all causal paths from C to E except π_d, then

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

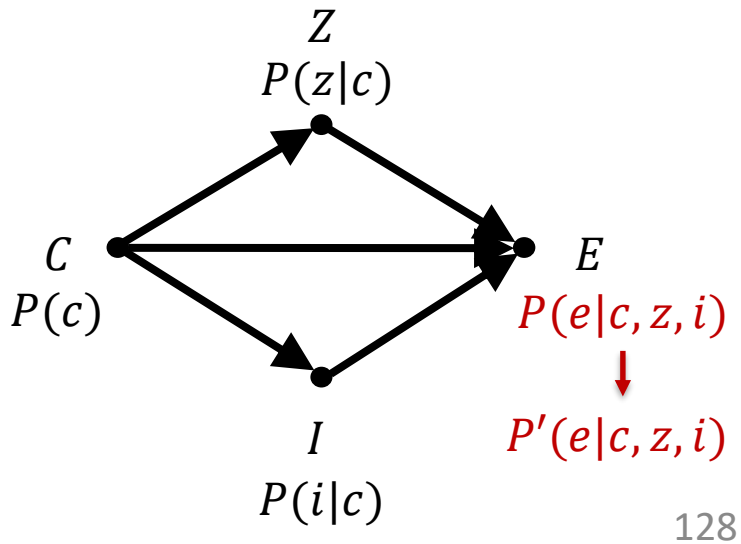
"reverse" π_i-specific effect

Discrimination Discovery and Removal Algorithms

- Path-Specific Effect based Discrimination Discovery (PSE-DD) algorithm
  - Build causal graph
  - Compute SE_{π_d} and SE_{π_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.

$$\text{minimize} \quad \sum_{\mathbf{V}} \left(P'(\mathbf{v}) - P(\mathbf{v})\right)^2$$

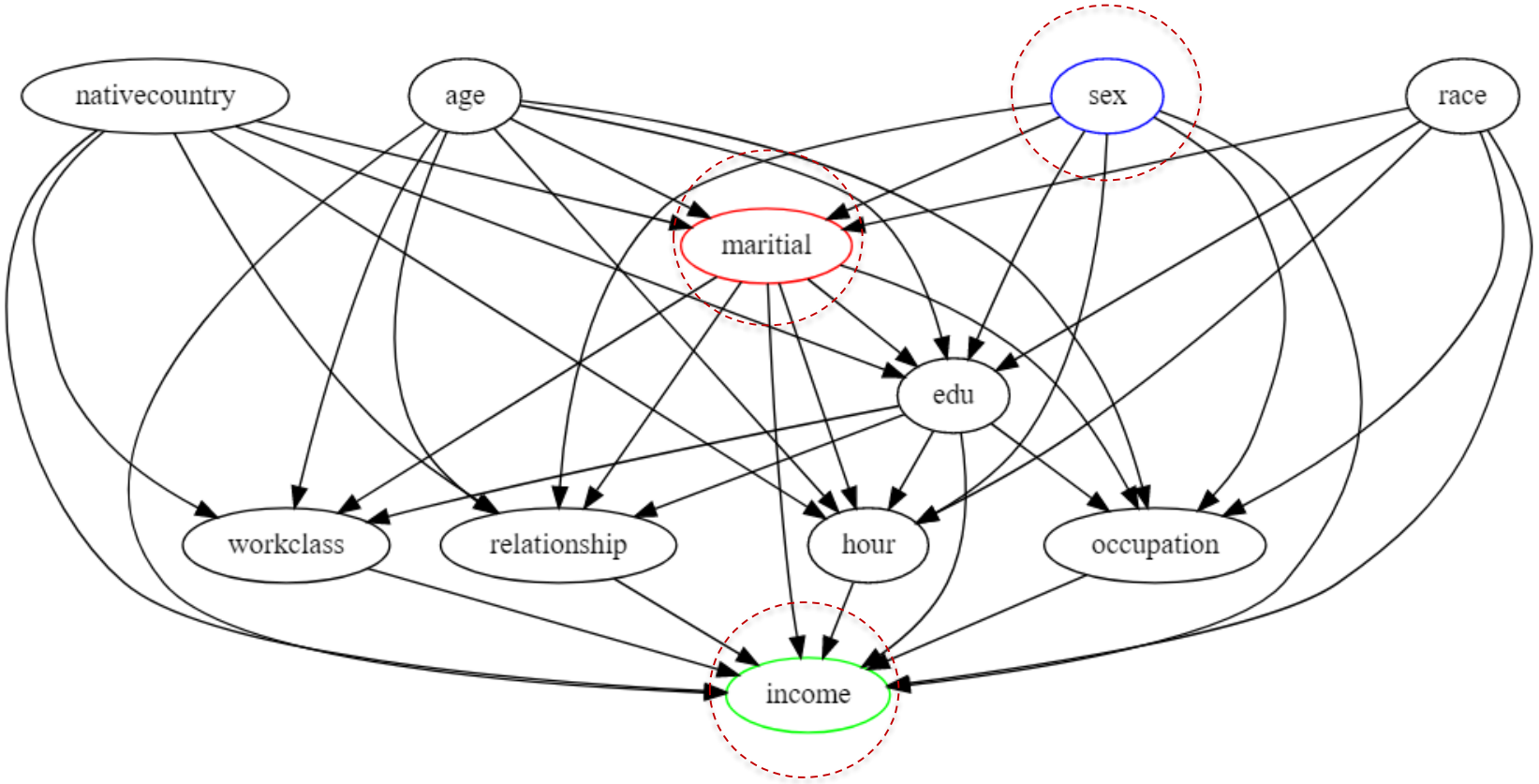
$$\begin{aligned}
\text{subject to} \quad & SE_{\pi_d}(c^+, c^-) \leq \tau, \quad SE_{\pi_d}(c^-, c^+) \leq \tau, \\
& SE_{\pi_i}(c^+, c^-) \leq \tau, \quad SE_{\pi_i}(c^-, c^+) \leq \tau, \\
& \forall Pa(E), \quad P'(e^-|Pa(E)) + P'(e^+|Pa(E)) = 1, \\
& \forall Pa(E), e, \quad Pr'(e|Pa(E)) \geq 0,
\end{aligned}$$



Empirical Evaluation

- Data: Adult dataset

protected attribute: sex decision: income redlining attribute: marital_status



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.

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

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.

	Remove Algorithm			
	PSE-DR	*DI*	*LMSG*	*LPS*
Direct	0.013	0.001	-0.142	-0.142
Indirect	0.049	0.050	0.288	0.174
$\chi^2(\times 10^4)$	1.038	4.964	1.924	1.292

Fair Inference on Outcomes

- Infer a fair distribution $P^*(C, \boldsymbol{X}, E)$ from a sample D drawn from the original distribution $P(C, \boldsymbol{X}, E)$.
- Approximate $P^*(C, \boldsymbol{X}, E)$ by solving a constrained maximum likelihood problem using path-specific effects

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

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

Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI'18 (2018)

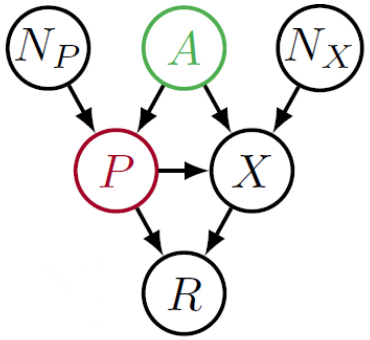
Variants of Indirect Discrimination

- 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

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NIPS'17 (2017)

Variants of Indirect Discrimination

- Develop procedures for avoiding discrimination in the predictor under linearity assumptions about the causal model.
- Example:

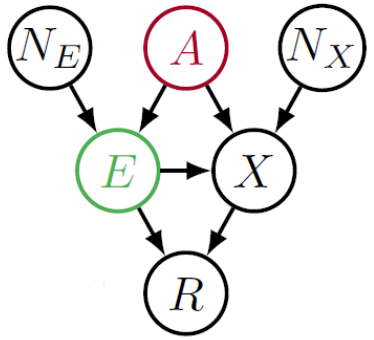


$$P = \alpha_P A + N_P$$
$$X = \alpha_X A + \beta P + N_X$$
$$R_\theta = \lambda_P P + \lambda_X X$$

Result: any predictor of the form

$$R_\theta = \lambda_X (X - \beta P)$$

with free parameter λ_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 λ_X, λ_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: $C \longrightarrow A$
- Non-protected attribute: $\boldsymbol{X} \longrightarrow \boldsymbol{X}$
- Decision attribute: $E \longrightarrow Y$
- Predictor: $\hat{E} \longrightarrow \hat{Y} = f(\boldsymbol{x}, a)$

Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: NIPS'17 (2017)
Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: NIPS'17 (2017)

Counterfactual Fairness

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

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

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

(ϵ, δ)-approximate counterfactual fair if

$$P(|f(\boldsymbol{x}_{A\leftarrow a}, a) - f(\boldsymbol{x}_{A\leftarrow a'}, a')| \leq \epsilon \mid \boldsymbol{x}, a) \geq 1 - \delta$$

- Also works for a dataset:

Counterfactual fair if for any context $\boldsymbol{U} = \boldsymbol{u}$, $\boldsymbol{X} = \boldsymbol{x}$ and $A = a$,

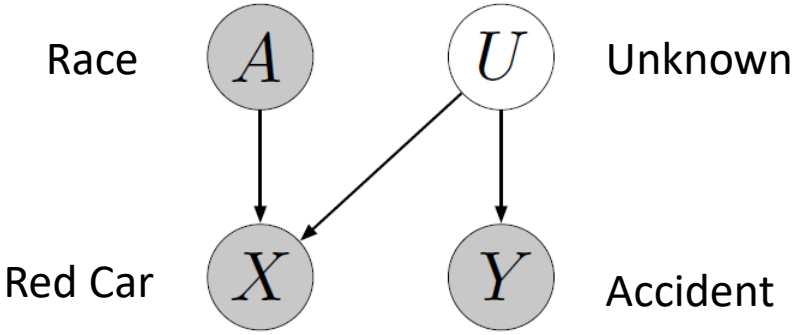
$$P(Y_a(\boldsymbol{u}) = y \mid \boldsymbol{X} = \boldsymbol{x}, A = a) = P(Y_{a'}(\boldsymbol{u}) = y \mid \boldsymbol{X} = \boldsymbol{x}, A = a),$$

for all y and a'.

Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: NIPS'17 (2017)

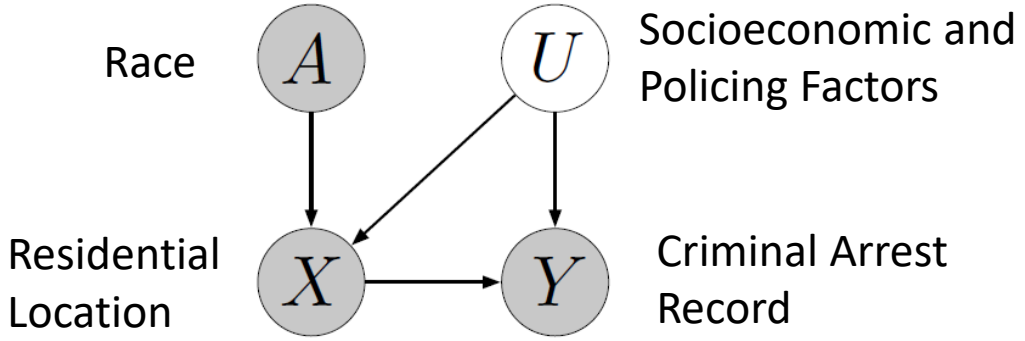
Toy Examples

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 $\boldsymbol{U}$ whose distribution, i.e., $P(\boldsymbol{u}|x, a)$, is known based on explicit domain knowledge;
  3. $\hat{Y}$ is built on latent variables $\boldsymbol{U}$ where the causal model is postulated, e.g., $x_i = f_i(pa_i) + \boldsymbol{u}_i$ with given types of function $f_i(\cdot)$.

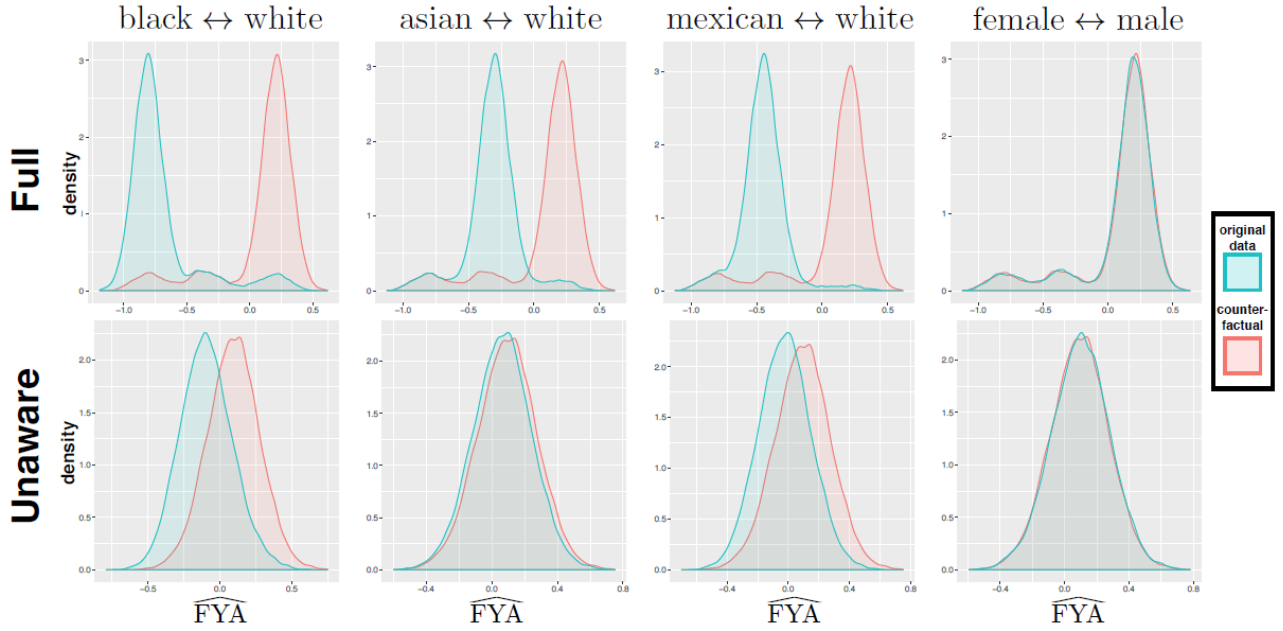
Illustrative Example

Protected attribute

- 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). Label
- 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.

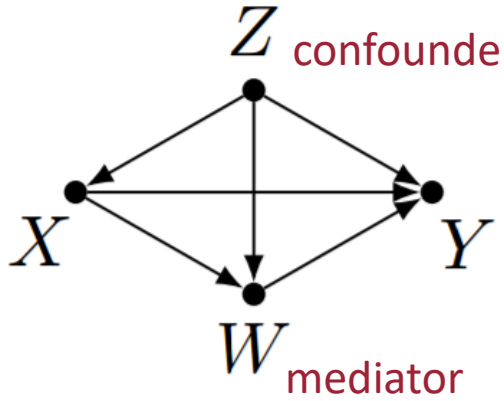


- Accuracy

	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918

Counterfactual Direct/Indirect Discrimination

- Protected attribute: $C \longrightarrow X$
- Non-protected attribute: $\boldsymbol{X} \longrightarrow Z, W$
- Decision attribute: $E \longrightarrow Y$



- 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 $X = x_0$.

Zhang, J., Bareinboim, E.: Fairness in decision-making – the causal explanation formula. In: AAAI'18 (2018)

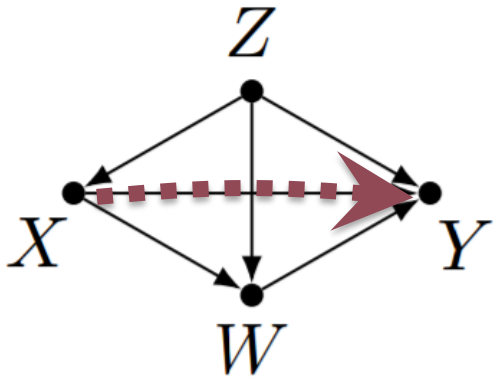
Counterfactual Direct/Indirect Discrimination

- 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 $X = x$

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

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):

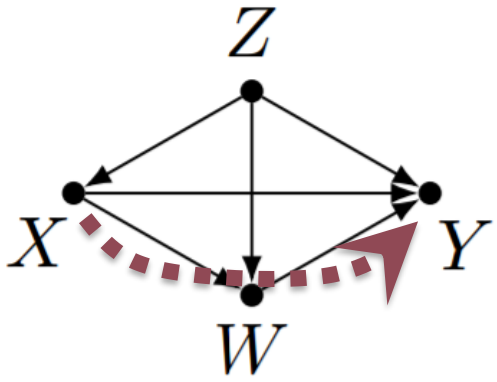
$$DE_{x_0,x_1}(y|x) = P\left(y_{x_1,W_{x_0}} \mid x\right) - P(y_{x_0} \mid x)$$

  - Counterfactual indirect effect (Ctf-IE): Indirect effect of intervention $X = x_1$ on Y (with baseline x_0) conditioned on $X = x$

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

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



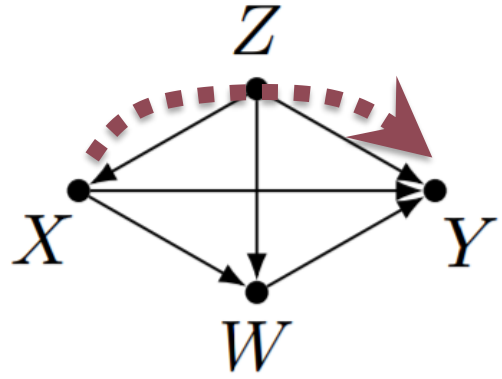
Counterfactual Direct/Indirect Discrimination

- 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

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

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*

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

$$TV_{x_0,x_1}(y) = DE_{x_0,x_1}(y|x_0) - SE_{x_1,x_0}(y) - IE_{x_1,x_0}(y|x_0) \quad (10)$$

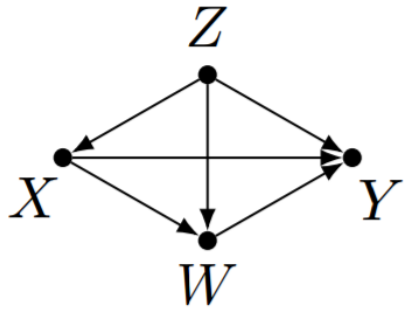
- Summary of different discrimination measures

Fairness	Measures	Discrimination		
		Direct	Indirect	Spurious
Outcome	TV	✓	✓	✓
	TE	✓	✓	✗
	ETT	✓	✓	✗
	NIE	✗	✓	✗
Procedure	NDE	✓	✗	✗
	QII	✓	✗	✗
	CDE	✓	✗	✗
Our Approach	Ctf-DE	✓	✗	✗
	Ctf-IE	✗	✓	✗
	Ctf-SE	✗	✗	✓

name	meaning	defination
TV	total variation	$P(y\|x_1) - P(y\|x_0)$
TE	total effect	$P(y_{x_1}) - P(y_{x_0})$
ETT	effect of the treatment on the treated	$P(y_{x_1}\|x_0) - P(y\|x_0)$
NIE	natural indirect effect	$P(y_{x_0}, w_{x_1}) - P(y_{x_0})$
NDE	natural direct effect	$P(y_{x_1}, w_{x_0}) - P(y_{x_0})$
QII	quantitative input influence	$E[Y] - E[Y_{X\sim P(x),Z,W\sim P(z,w)}]$
CDE	controlled direct effect	$P(y_{x_1,z,w}) - P(y_{x_0,z,w})$

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

$$\sum_{z,w}(P(y|x_1,w,z)-P(y|x_0,w,z))P(w|x_0,z)P(z|x),$$

$$\sum_{z,w}P(y|x_0,w,z)(P(w|x_1,z)-P(w|x_0,z))P(z|x),$$

$$\sum_{z,w}P(y|x_0,w,z)P(w|x_0,z)(P(z|x_1)-P(z|x_0)).$$

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

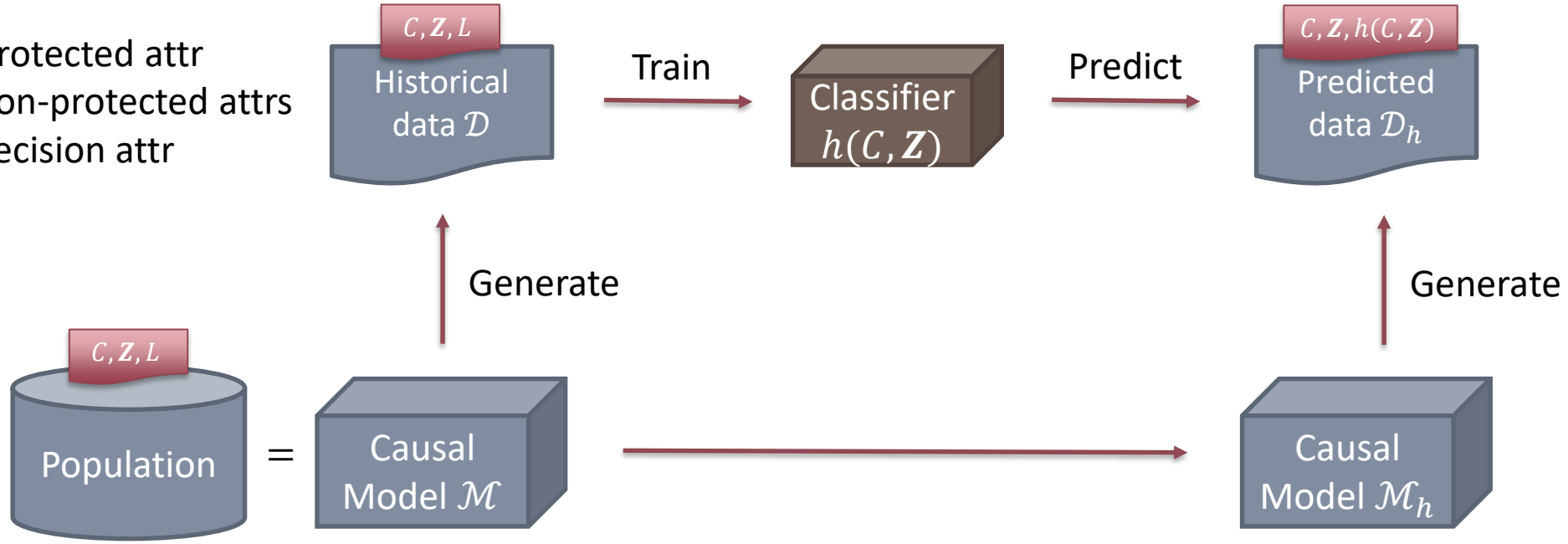
- Will a classifier learned from a discrimination-free training data also be discrimination-free?

- 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.

Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. In: IJCAI'18 (2018)

Causal Modeling-Based Anti-Discrimination Framework

C: protected attr
$\boldsymbol{Z}$: non-protected attrs
L: decision attr



Causal Model $\mathcal{M}$

$$c = f_C(\boldsymbol{pa}_C, \boldsymbol{u}_C)$$
$$\forall z_i \in \boldsymbol{Z},\ z_i = f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i)$$
$$l = f_L(\boldsymbol{pa}_L, \boldsymbol{u}_L)$$

Causal Model $\mathcal{M}_h$

$$c = f_C(\boldsymbol{pa}_C, \boldsymbol{u}_C)$$
$$\forall z_i \in \boldsymbol{Z},\ z_i = f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i)$$
$$\boxed{l = h(c, \boldsymbol{z})}$$

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 $\boldsymbol{u}$, his/her label under intervention $do(c^+)$: $L_{c^+}(\boldsymbol{u})$
- Expectation of differences in labels under $do(c^+)$ and $do(c^-)$:

$$\mathbb{E}[L_{c^+}(\boldsymbol{u}) - L_{c^-}(\boldsymbol{u})]$$

- Derived causal measures of discrimination:
  - $DE_{\mathcal{M}} = \mathbb{E}[L_{c^+}(\boldsymbol{u}) - L_{c^-}(\boldsymbol{u})] = P(l^+|c^+) - P(l^+|c^-)$
  - $DE_{\mathcal{D}} = \hat{P}(l^+|c^+) - \hat{P}(l^+|c^-)$
  - $DE_{\mathcal{M}_h} = \mathbb{E}\left[h\left(c^+, \boldsymbol{Z}_{c^+}(\boldsymbol{u})\right) - h\left(c^-, \boldsymbol{Z}_{c^-}(\boldsymbol{u})\right)\right] = P\left(\tilde{l}^+\middle|c^+\right) - P(\tilde{l}^+|c^-)$
  - $DE_{\mathcal{D}_h} = \hat{P}\left(\tilde{l}^+\middle|c^+\right) - \hat{P}\left(\tilde{l}^+\middle|c^-\right)$

$$= \sum_{\boldsymbol{z}} \mathbb{I}_{[h(c^+,\boldsymbol{z})=l^+]} \hat{P}(\boldsymbol{z}|c^+) - \sum_{\boldsymbol{z}} \mathbb{I}_{[h(c^-,\boldsymbol{z})=l^+]} \hat{P}(\boldsymbol{z}|c^-)$$

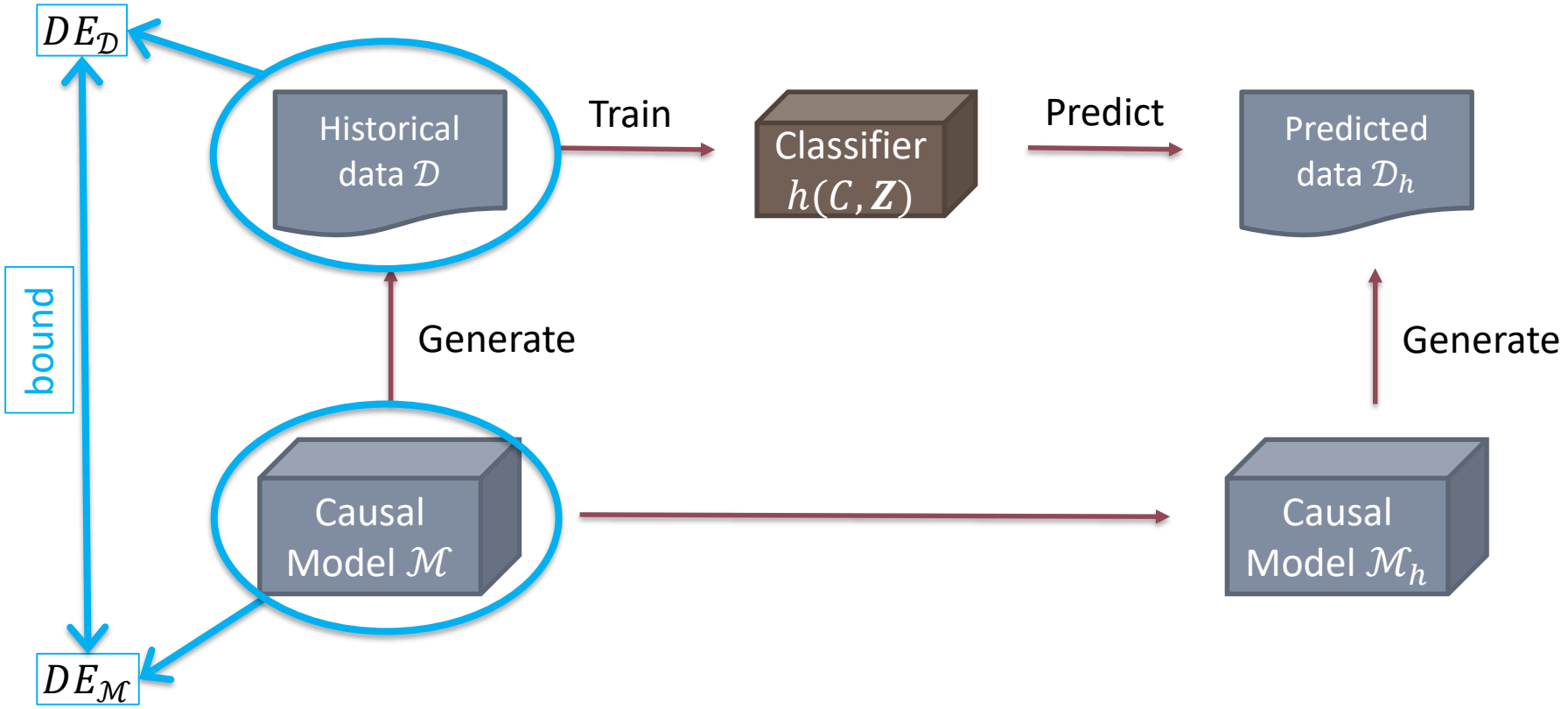
Coincident with risk difference

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 τ.

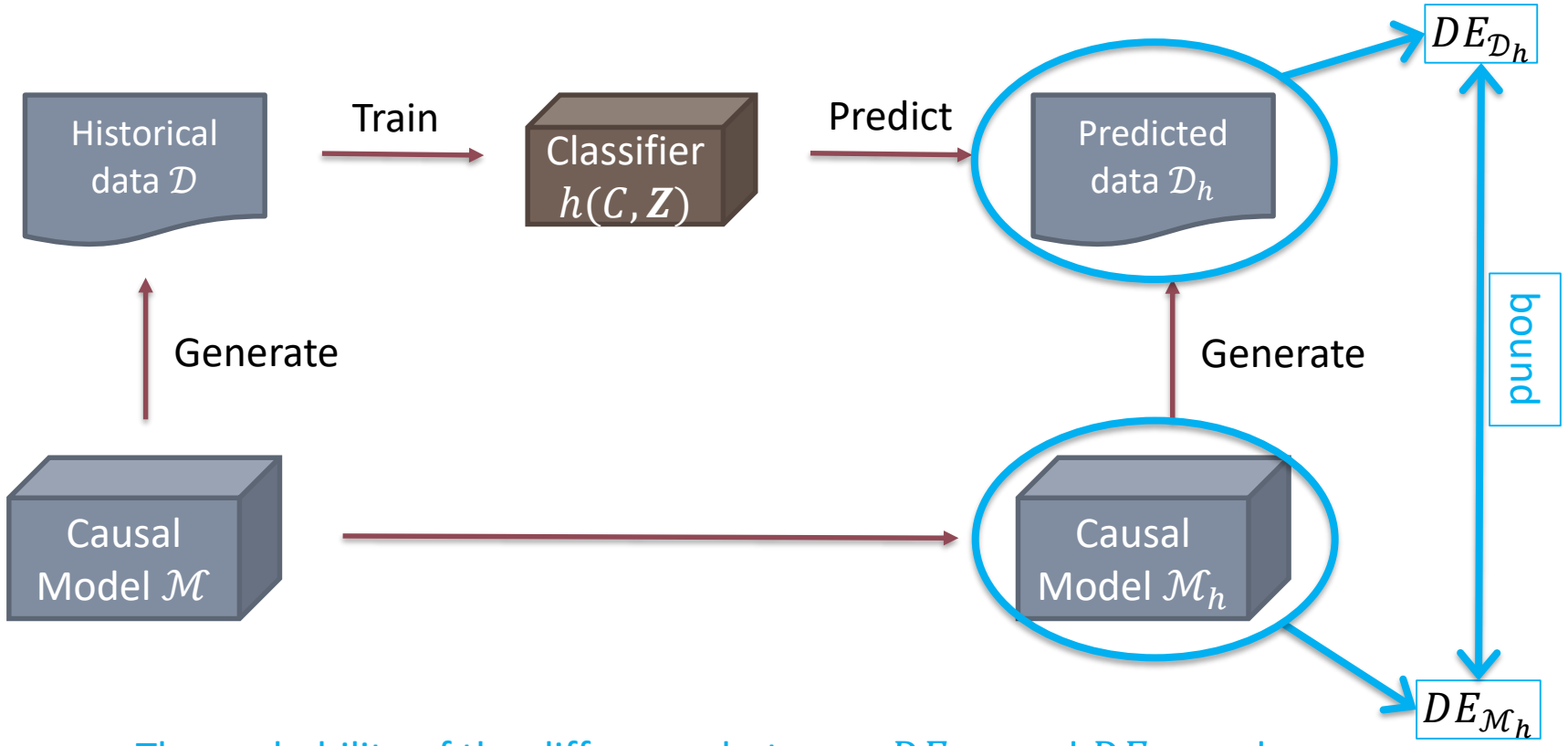
Discover (Bound) Discrimination in Prediction



The probability of the difference between $DE_{\mathcal{M}}$ and $DE_{\mathcal{D}}$ no larger than t is bounded by

$$P(|DE_{\mathcal{M}} - DE_{\mathcal{D}}| \leq t) > 1 - 4e^{-\frac{n^+ n^-}{n} t^2}$$

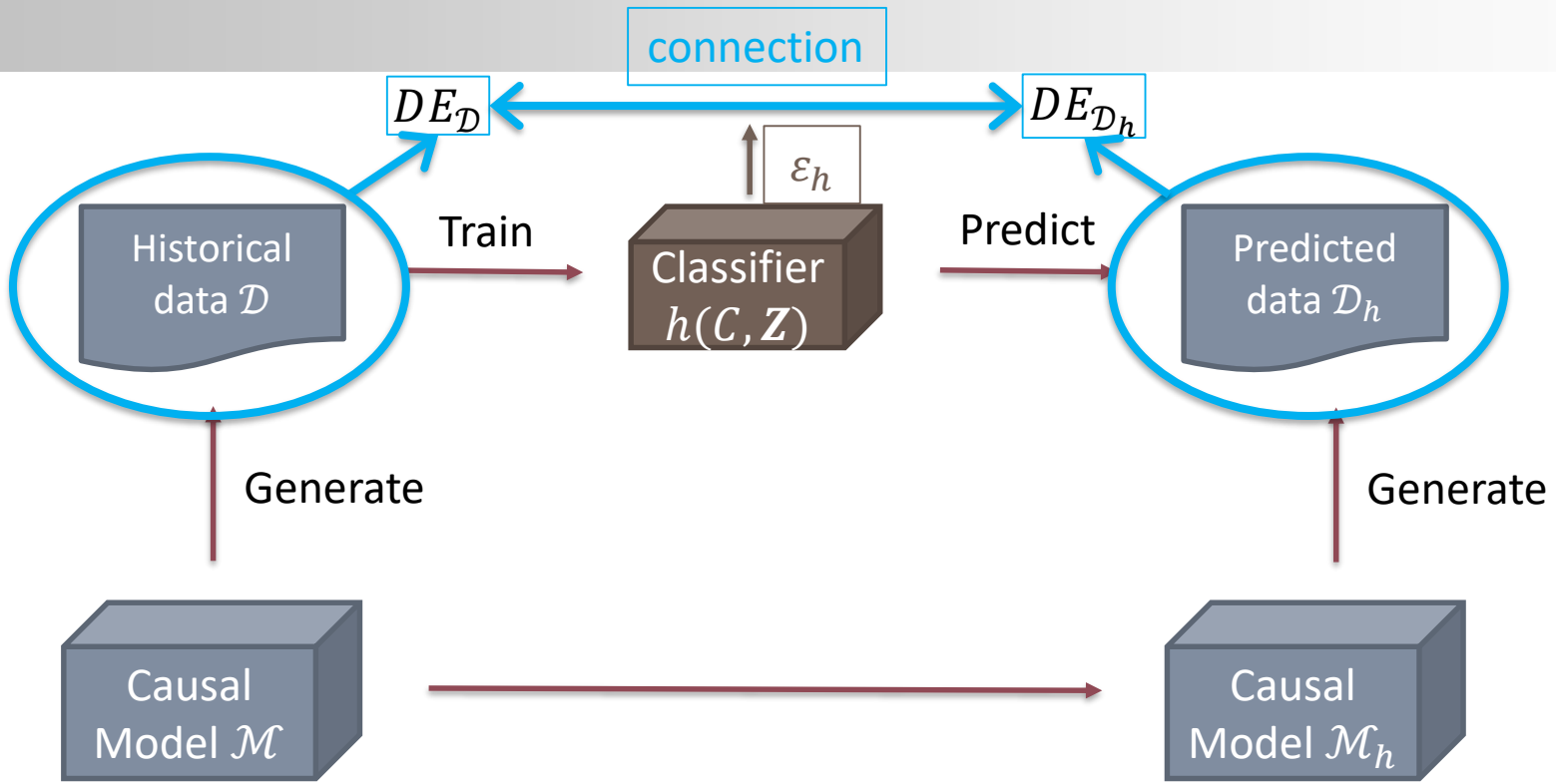
Bound Discrimination in Prediction



The probability of the difference between $DE_{\mathcal{M}_h}$ and $DE_{\mathcal{D}_h}$ no larger than t is bounded by

$$P\left(\left|DE_{\mathcal{M}_h} - DE_{\mathcal{D}_h}\right| \leq t\right) > 1 - \delta(t) \quad \text{where} \quad \delta(t) = \begin{cases} 4|\mathcal{H}|^2 e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is finite} \\ 4\dfrac{(2en^+)^d + (2en^-)^d}{d^d} e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is infinite} \end{cases}$$

Bound Discrimination in Prediction



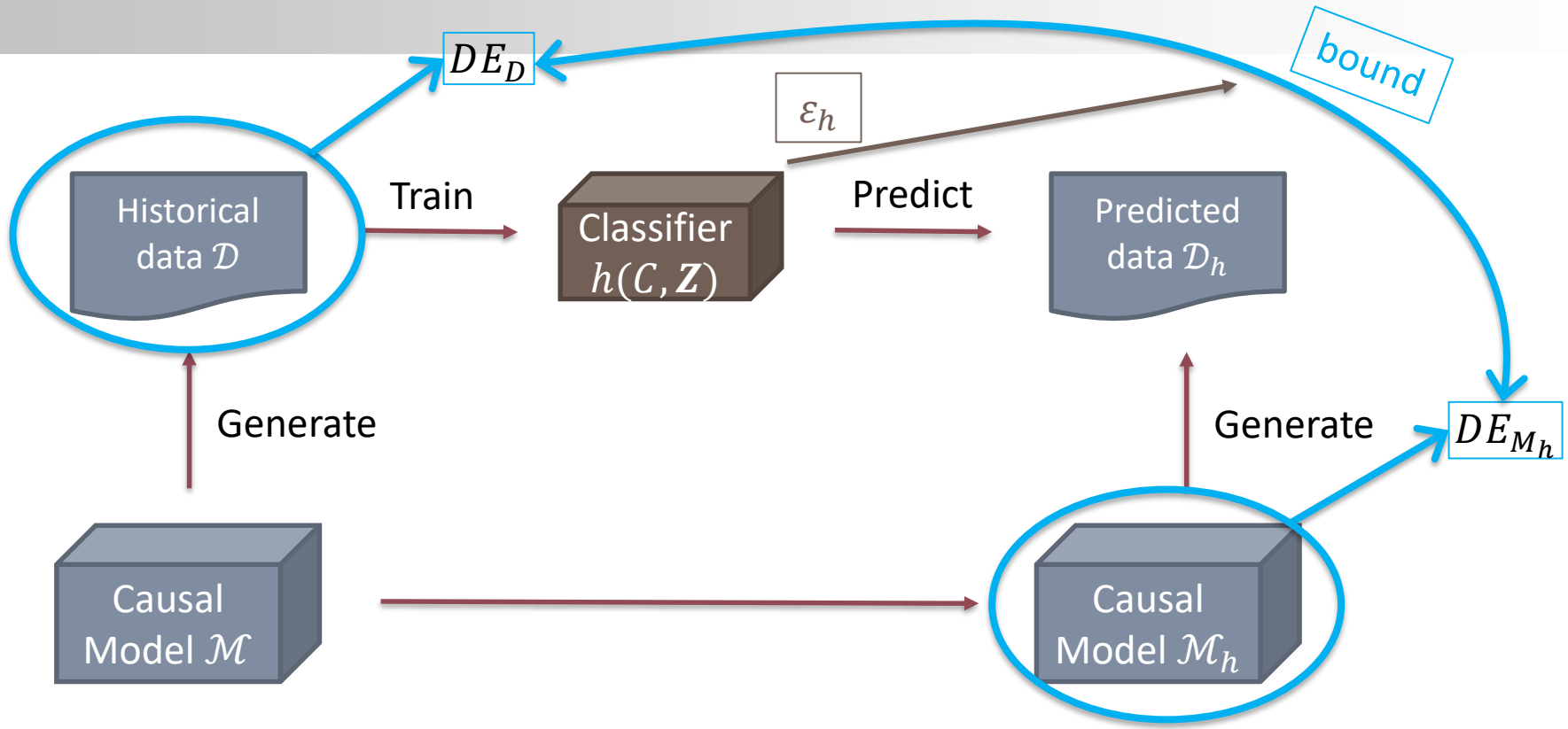
$$DE_{\mathcal{D}_h} - DE_{\mathcal{D}} = \varepsilon_{h,\mathcal{D}}$$

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^-

Bound Discrimination in Prediction



Theorem 1. *Given a causal model $\mathcal{M}$, a sample dataset $\mathcal{D}$ and a classifier h trained on $\mathcal{D}$, $\mathrm{DE}_{\mathcal{M}_h}$ is bounded by*

$$P\left(\left|\mathrm{DE}_{\mathcal{M}_h}\right| \leq \left|\mathrm{DE}_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}\right| + t\right) \geq 1 - \delta(t)$$

Remove Discrimination in Prediction

- Removing discrimination from training data ONLY is NOT enough as discrimination in prediction depends on $DE_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}$.
- Two-phase framework for non-discrimination in prediction:
  1. (Data modification) Modify training dataset $\mathcal{D}$ to obtain a modified dataset $\mathcal{D}^*$ such that $|DE_{\mathcal{D}^*}| \leq \tau$;
  2. (Classifier tweaking) Train a classifier h^* on $\mathcal{D}^*$ (and tweak it) such that $\left|DE_{\mathcal{D}^*} + \varepsilon_{h^*,\mathcal{D}^*}\right| \leq \tau$.
- 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

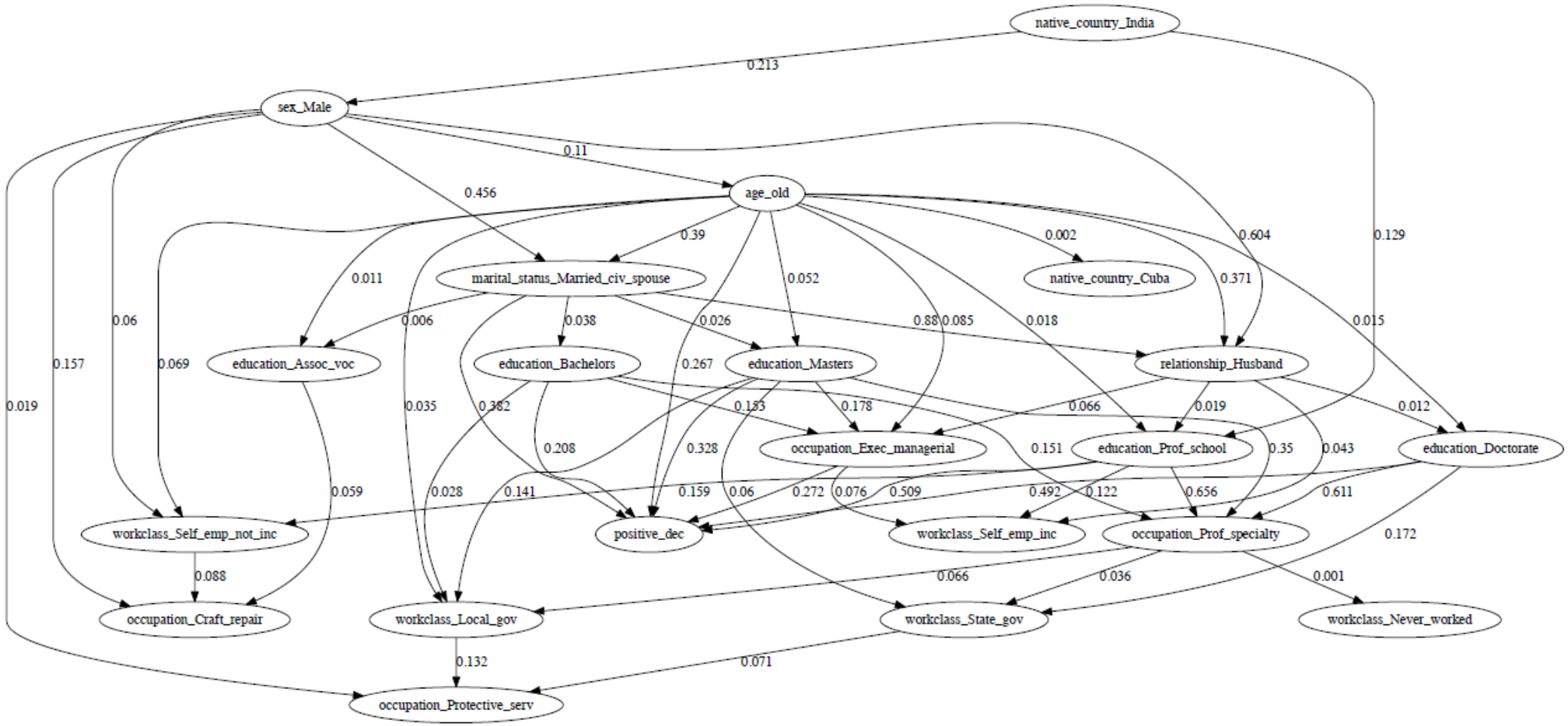
- Part I: Introduction
- Part II: Correlation based Anti-Discrimination Learning
- Part III: Causal Modeling Background
- Part IV: Causal Modeling-Based Anti-Discrimination Learning
  - Direct and Indirect Discrimination
  - Counterfactual Fairness
  - Data discrimination vs. model discrimination
  - 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 $p(u|v) - p(u|\neg v)$

Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. *Int. J. Data Sci. Anal.* 3(1), 1–21 (2017)

A SBCN Example



Discrimination Score using SBCN

- Discrimination score
  - $ds^{-}(v) = \frac{rw_{v \to e^{-}}}{n}$
  - v is a node of SBCN (e.g. female), e^{-} is the node of negative decision, $rw_{v->e^{-}}$ 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
  - $gds^{-}(v_1, \dots, v_n) = \frac{ppr(e^{-}|v_1, \dots v_n)}{ppr(e^{-}|v_1, \dots v_n) + ppr\left(e^{+}\middle|v_1, \dots v_n\right)}$
  - $ppr(e^{-}|v_1, \dots v_n)$ 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
  - $\Delta P|_{\boldsymbol{s}} = P(e^+|c^+, \boldsymbol{s}) - P(e^+|c^-, \boldsymbol{s})$
  - τ: an user-defined threshold for discrimination detection depending on laws and regulations (e.g., 0.05).
  - If $\Delta P|_{\boldsymbol{s}} < \tau$ holds across all possible partitions and their values $\boldsymbol{s}$, then no discrimination.

Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)

Achieving Non-Discrimination in Data Release

- Achieve a non-discrimination guarantee

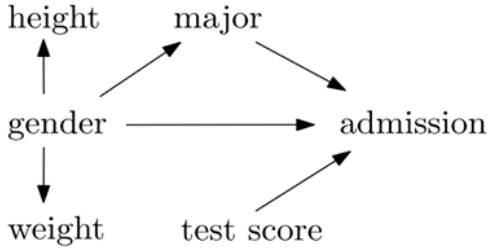
  against all possible lawsuits → for all meaningful subgroups

No.	...
gender	...
major	...
score	...
height	...
weight	...
admission	...

partition

gender	Female	Male
admission (%)	43%	43%

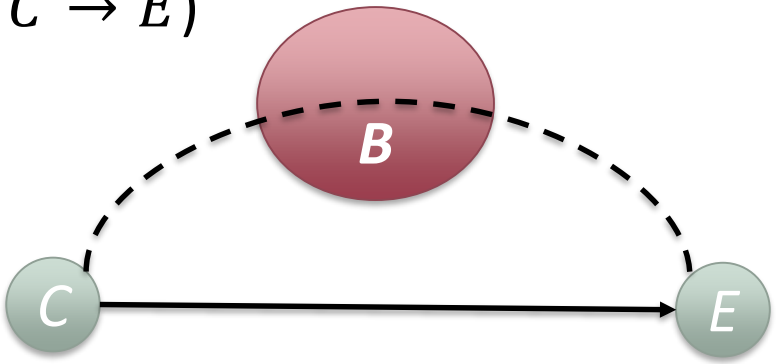
major	CS		EE	
gender	Female	Male	Female	Male
admission (%)	38%	38%	47%	47%



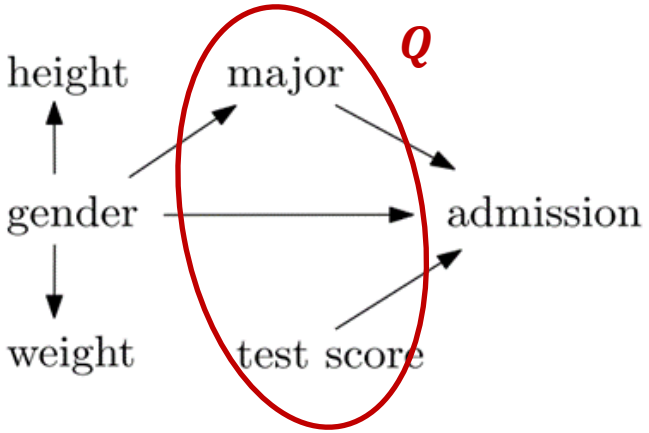
major	CS				EE			
test score	Low		High		Low		High	
gender	Female	Male	Female	Male	Female	Male	Female	Male
admission (%)	30%	36%	50%	40%	40%	45%	60%	50%

Achieving Non-Discrimination in Data Release

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



- Ensure $|\Delta P|_{\boldsymbol{b}}| < \tau$ for each $\boldsymbol{b}$ of each $\boldsymbol{B}$.
  - $\Delta P|_{\boldsymbol{b}} = P(e^{+}|c^{+}, \boldsymbol{b}) - P(e^{+}|c^{-}, \boldsymbol{b})$
- Let $\boldsymbol{Q} = Pa(E)\backslash\{C\}$, if $\left|\Delta P|_{\boldsymbol{q}}\right| < \tau$ holds, it is guaranteed $|\Delta P|_{\boldsymbol{b}}| < \tau$ 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 $\Delta P|_{\boldsymbol{q}} \geq \tau$, randomly select a number of individuals from the $\{c^- e^-\}$ group and change decision from e^- to e^+.
  - If $\Delta P|_{\boldsymbol{q}} \leq -\tau$, do the similar modification.
  - As a result, ensure that $|\Delta P|_{\boldsymbol{q}}| \leq \tau$ holds for each $\boldsymbol{q}$.

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.

Adult	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	1.18	10.27	39.35	38.65	35.60	60.65
n_T	1114	4122	29944	16048	16366	44582
χ^2	153	8470	18428	26900	10819	99770
Disc.	0	0	64	104	77	128
Dutch	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	5.68	6.75	13.91	18.00	15.48	14.10
n_T	10422	8838	32516	29288	24648	35728
χ^2	2832	4825	14014	10555	5039	19684
Disc.	0	0	1	9	4	12

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.

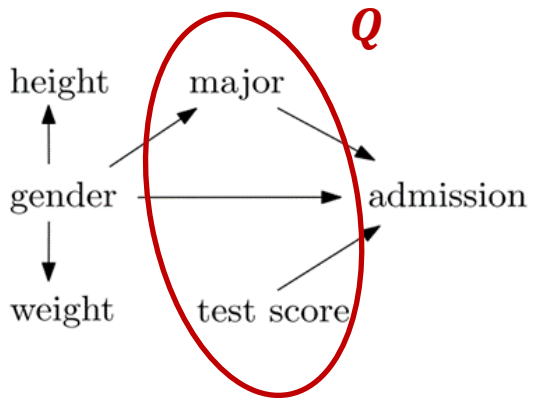
- How to find similar individuals for situation testing?

Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)

Individual Discrimination Discovery

- Situation testing: find similar individuals for the target.

No.	gender	major	score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject
⋮	⋮	⋮	⋮	⋮	⋮	⋮



Individual Discrimination Discovery

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

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

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

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

- $VD(q_k, q'_k)$ measures the difference between two values q_k and q'_k of each attribute $Q_k \in \boldsymbol{Q}$.

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

Empirical Evaluation

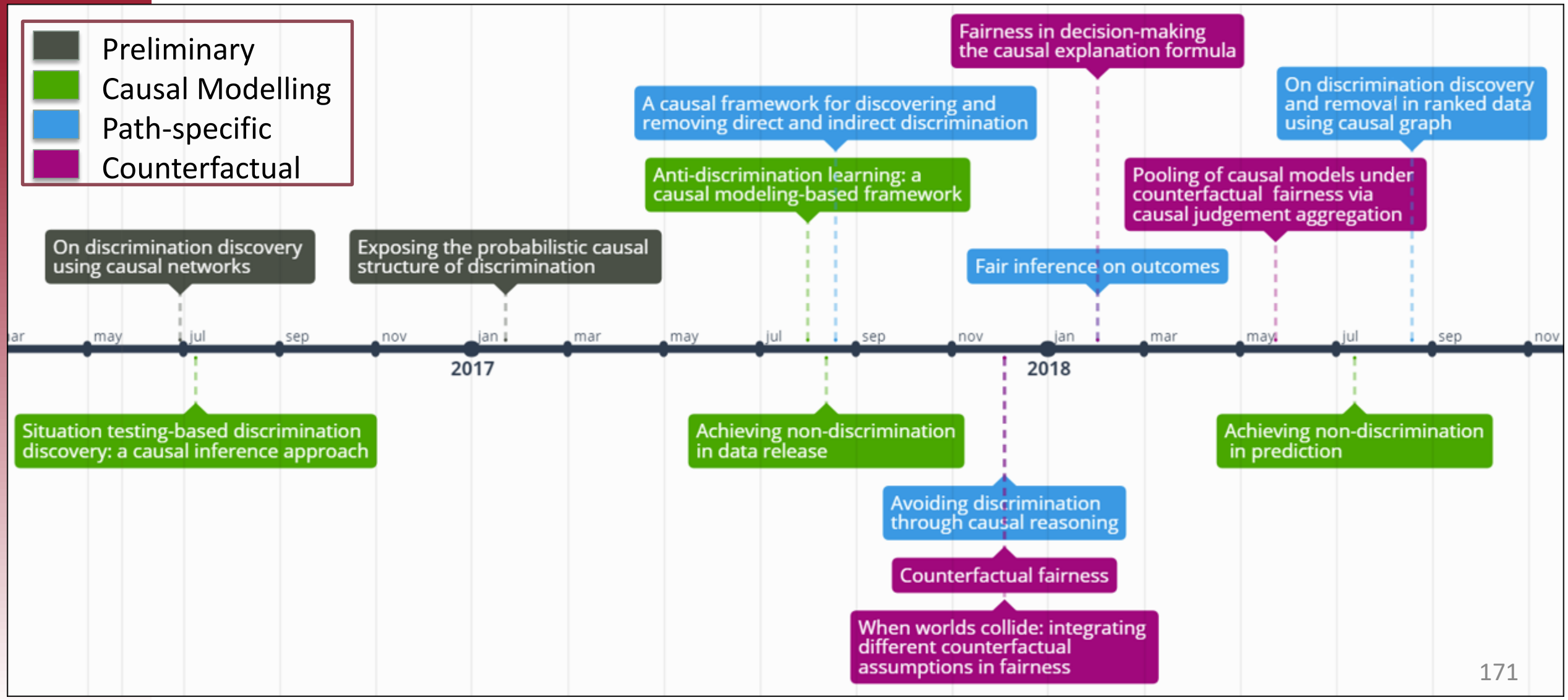
- Data: Dutch Census of 2001
- Comparison of Different Methods
  - CBN-based situation testing (CBN-DD) (Zhang et al. IJCAI 2017)
  - KNN-based situation testing (KNN-DD) (Luong et al. SIGKDD 2011)
- Result:
  - KNN-DD and CBN-DD are significantly different.
  - CBN-DD outperforms KNN-DD over the synthetic data.

- 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.

Accuracy

K	*CBN-DD*		*KNN-DD*	
	TP	TN	TP	TN
10	73.3	63.1	46	66.2
50	85.3	77.6	42.2	76.2
90	81.5	83.9	38.1	81.2

Summary


Preliminary
Causal Modelling
Path-specific
Counterfactual
Fairness in decision-making the causal explanation formula
A causal framework for discovering and removing direct and indirect discrimination
On discrimination discovery and removal in ranked data using causal graph
Anti-discrimination learning: a causal modeling-based framework
Pooling of causal models under counterfactual fairness via causal judgement aggregation
On discrimination discovery using causal networks
Exposing the probabilistic causal structure of discrimination
Fair inference on outcomes
may
jul
sep
nov
jan
mar
may
jul
sep
nov
jan
mar
may
jul
sep
nov
2017
2018
Situation testing-based discrimination discovery: a causal inference approach
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

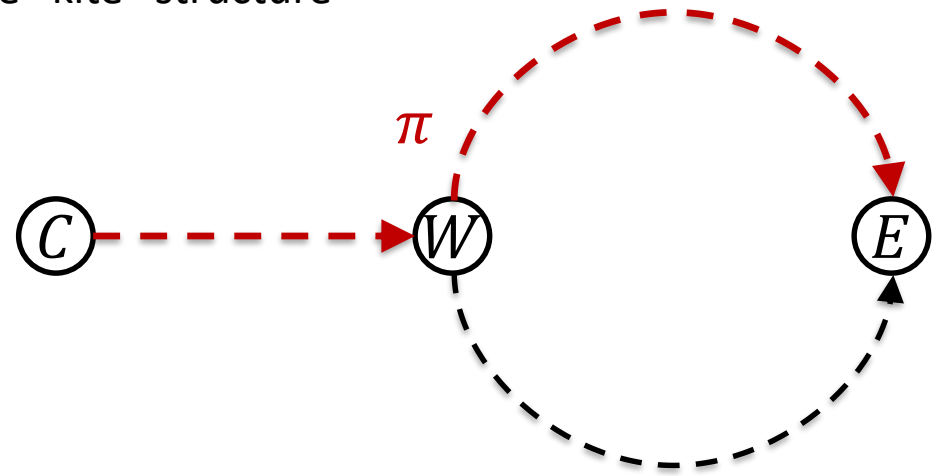
Challenges

- 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

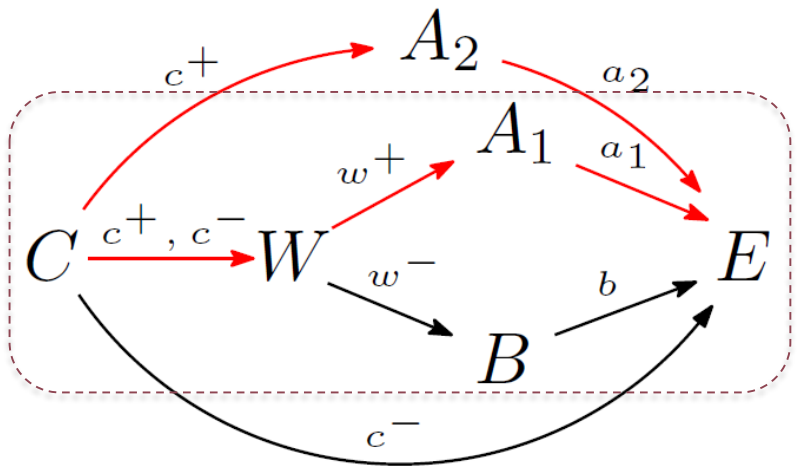


Unidentifiable Situation

- When the recanting witness criterion is satisfied, indirect discrimination SE_{π_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

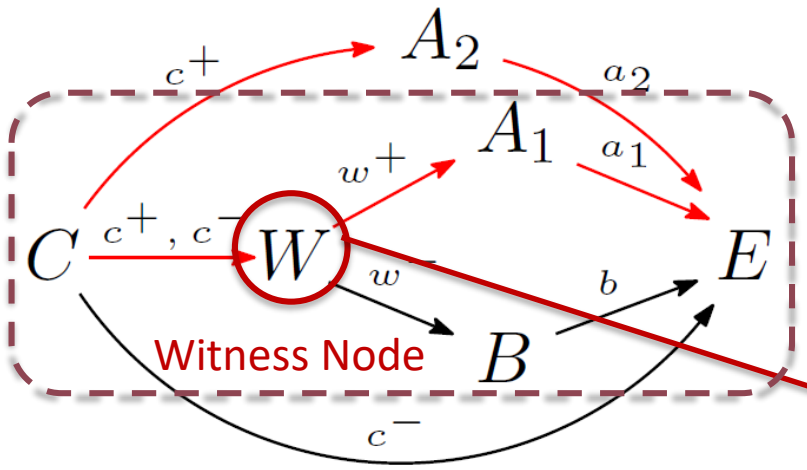
- Principled approaches for dealing with non-identifiable path-specific effects (Nabi et al., 2018)
  - Measure hidden variables $\boldsymbol{U}$ or obtain reliable proxies for them, if possible.
  - Consider a path-specific effect that is identifiable, which includes the paths of interest and some other 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.

Zhang, L., Wu, Y., Wu, X.: Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. In: TKDE, under review (2018)
Tian, J., Pearl, J.: Probabilities of causation: Bounds and identification. In: UAI'00 (2000)

Bounding π_i-specific Effect

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

The "kite" structure



Zhang, L., Wu, Y., Wu, X.: Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. In: TKDE, under review (2018)

$$P\left(e^+ \middle| do\left(c^+|_{\pi_i}\right)\right) = \sum_{A_1,A_2,B,W^+,W^-} P(e^+|c^-, a_1, a_2, b)P(a_1|w^+)P(a_2|c^+)P(b|w^-)P(w^+_{c^+}, w^-_{c^-})$$

- 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 $\sum_{w^-} P\left(w^+_{c^+}, w^-_{c^-}\right) = P(w^+_{c^+})$

Bounding π_i-specific Effect

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

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|pa_A^+) \prod_{B\in\mathbf{B}} P(b|pa_B^-) \prod_{W\in\mathbf{W}} P(w^-|pa_W^-)$$

- Lower bound

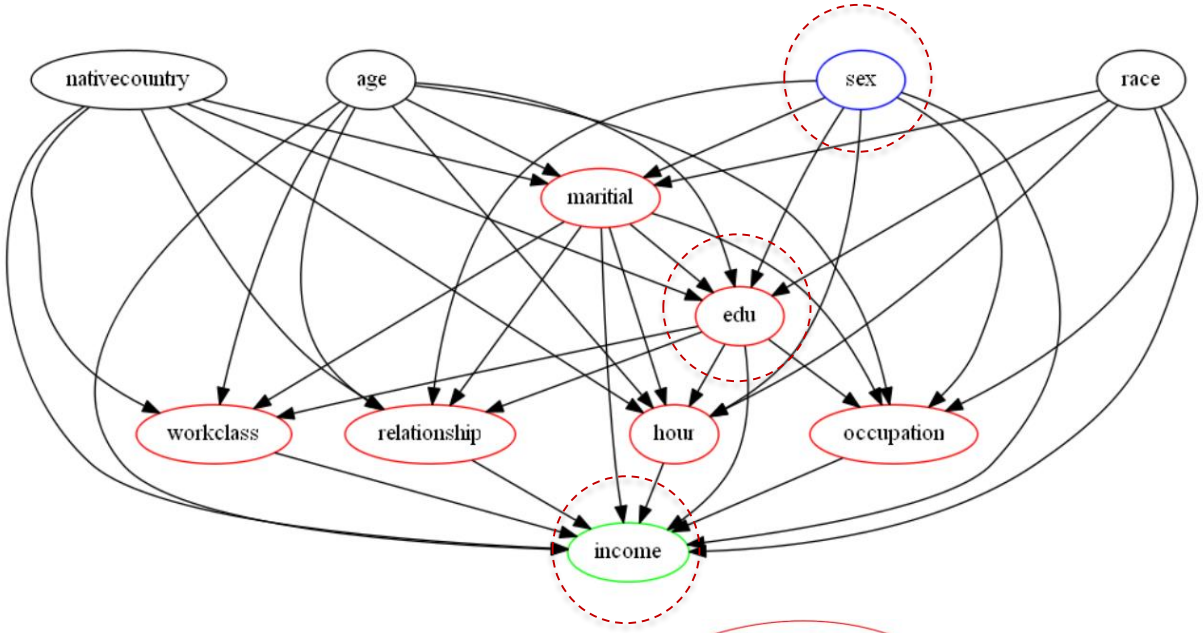
- Notations:
  - $\boldsymbol{W}$: witness nodes
  - $\boldsymbol{A}_1$: nodes in π_i not in $\boldsymbol{W}$ but involved in “kite pattern”
  - $\boldsymbol{A}_2$: nodes in π_i not in $\boldsymbol{W}$ and not involved in “kite pattern”
  - $\boldsymbol{B}$: nodes not in π_i

Using Bounds for Discrimination Discovery and Removal

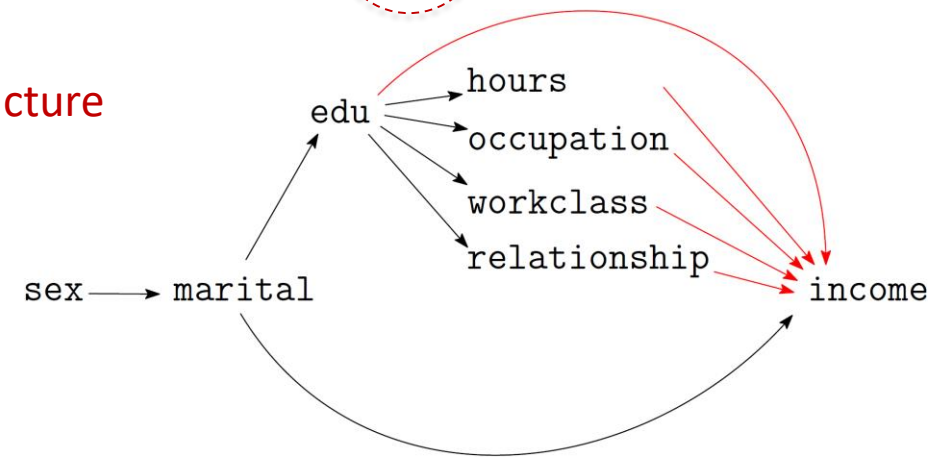
- Utilize lower and upper bounds for identifying indirect discrimination.
  - If $upper\ bound < threshold$, non-discrimination for certain.
  - If $lower\ bound \geq threshold$, discrimination for certain.
  - If $lower\ bound < threshold \leq upper\ bound$, uncertain.
- For removal, replace $SE_{\pi_i}(c^+, c^-)$ 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



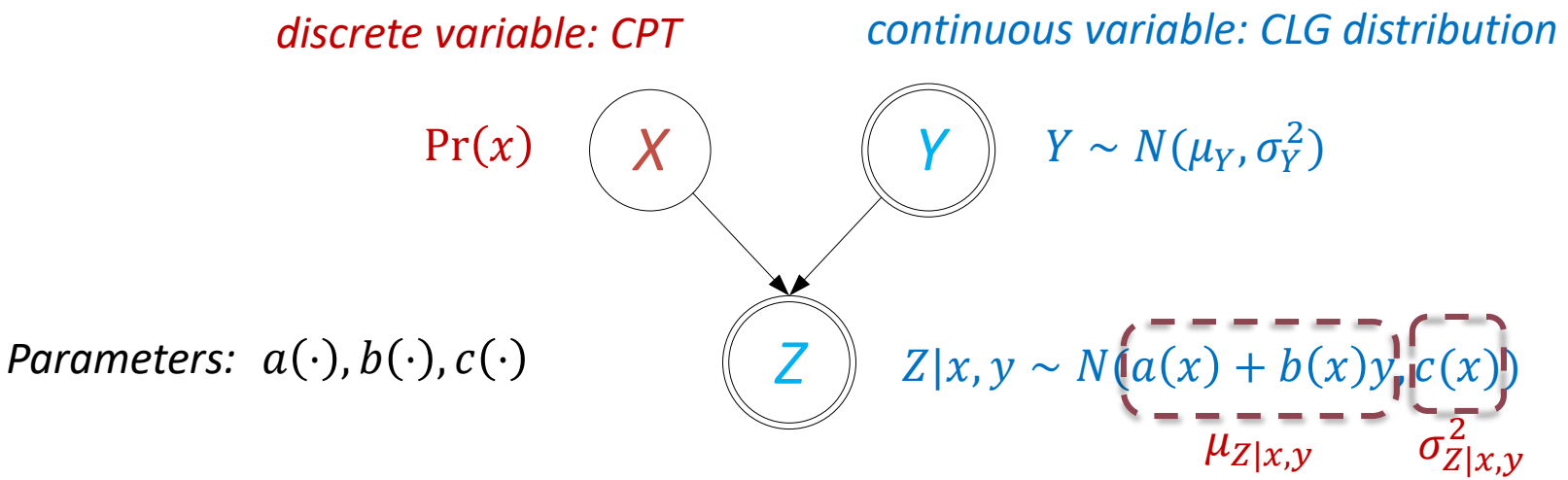
	PSE-DR	*PSE-DR**
Direct	0.038	0.033
Indirect (*ub*)	0	0.050
$\chi^2(\times 10^4)$	1.499	1.106

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

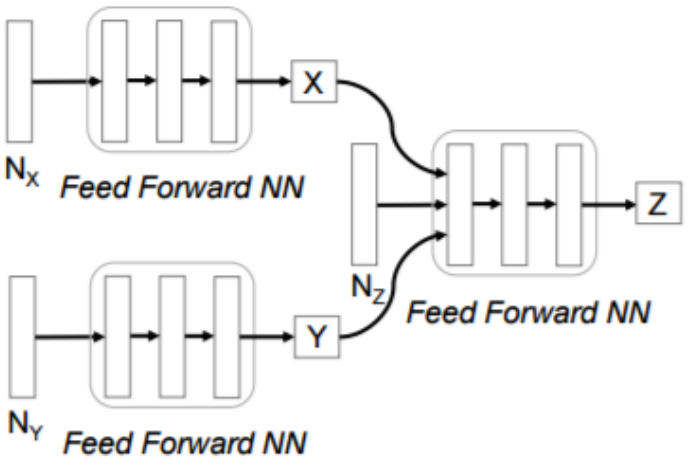


- Limitation: discrete variables can only have discrete parents.

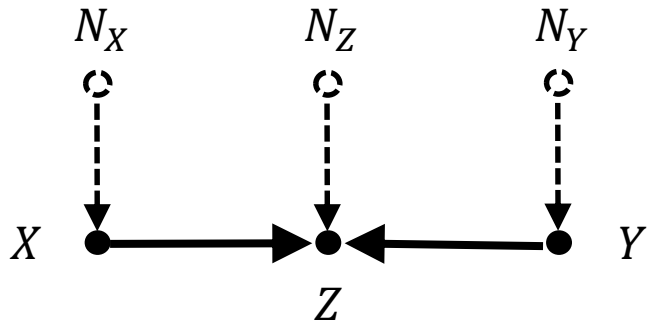
Madsen, A.L.: Belief update in CLG Bayesian networks with lazy propagation. *Int. J. Approx. Reason. 49*(2),503-521 (2008)

Causal Modeling for Mixed-type Variables

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



Neural network architecture that represents:



Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: CausalGAN: Learning causal implicit generative models with adversarial training. In: ICLR'18 (2018)

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 $\boldsymbol{U}$ are mutually independent.

Relaxing Markovian Assumption

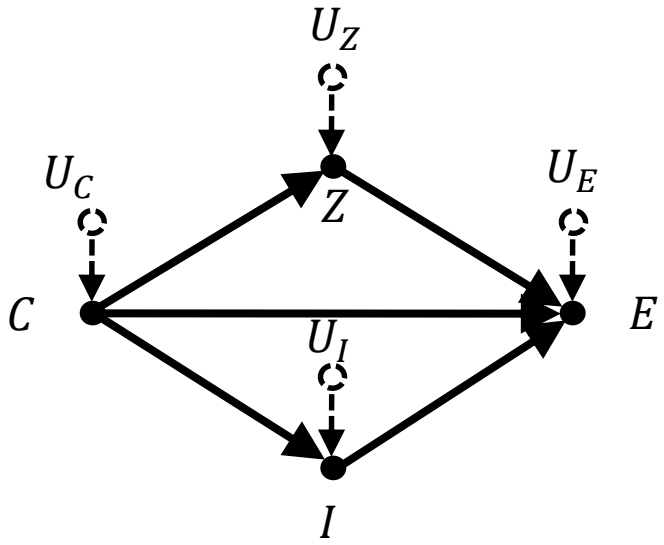
- A causal model is semi-Markovian if
  1. The causal graph is acyclic;
  2. All variables in $\boldsymbol{U}$ are NOT mutually independent.

- Hidden confounders are known to exist in the system.

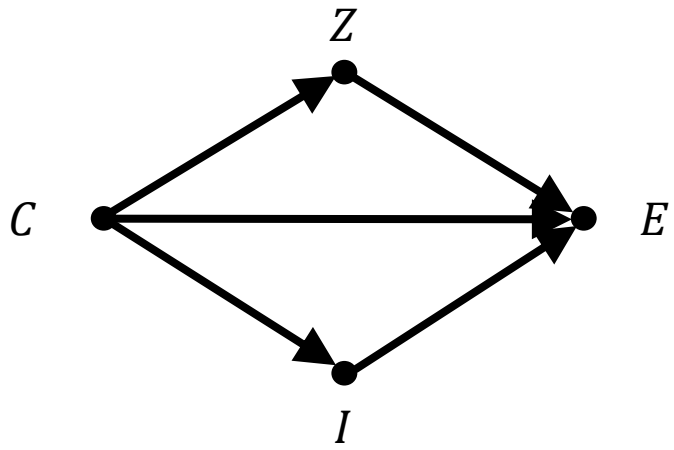
- The causal graph of the semi-Markovian model is commonly represented by the acyclic directed mixed graph (ADMG).
  - The bidirected arrow $\leftrightarrow$ implies the presence of unobserved confounder(s) between variables.

Relaxing Markovian Assumption

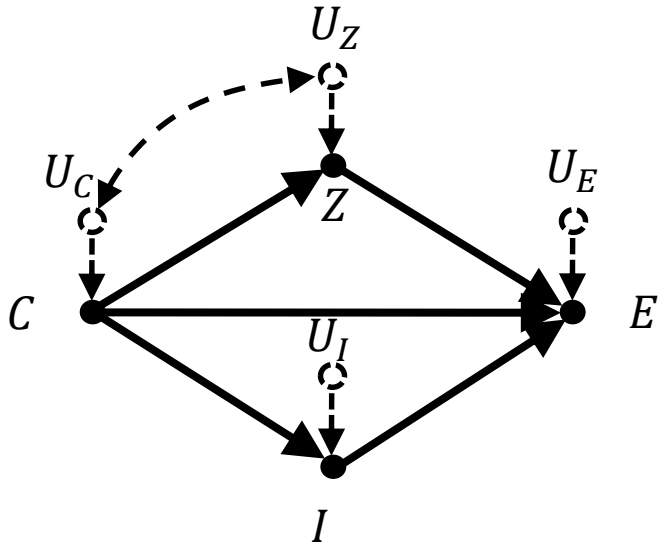
Markovian model



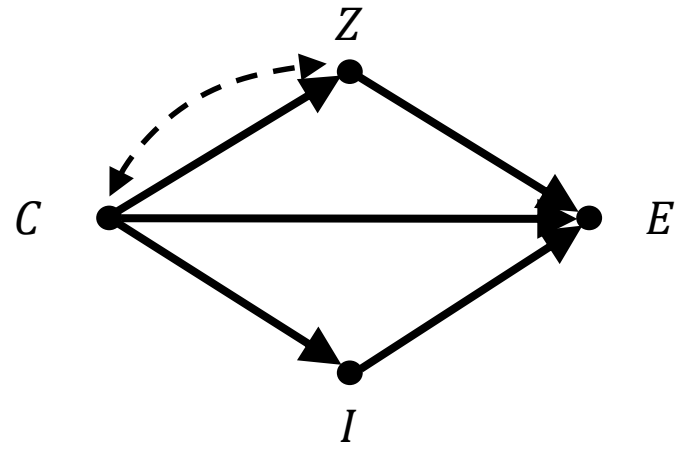
Directed Acyclic Graph (DAG)



Semi-Markovian model

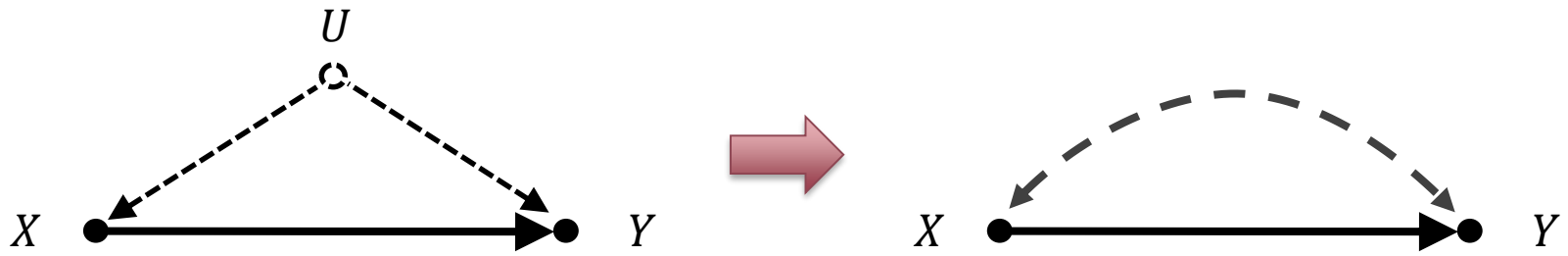


Acyclic Directed Mixed Graph (ADMG)



Intervention in Semi-Markovian Model

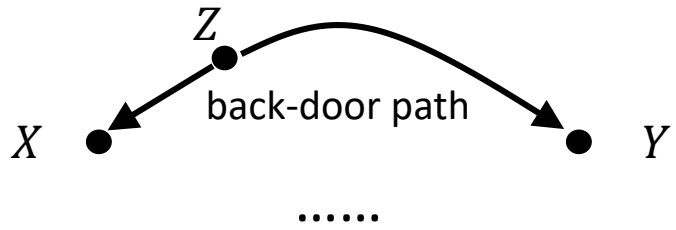
- Intervention also applies to semi-Markovian model.
- Unlike in the Markovian model, *do*-operations may not be able to be calculated from observational data (i.e., identifiable) due to unobserved confounders.
- "Bow-arc graph": the simplest non-identifiable graph structure.



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

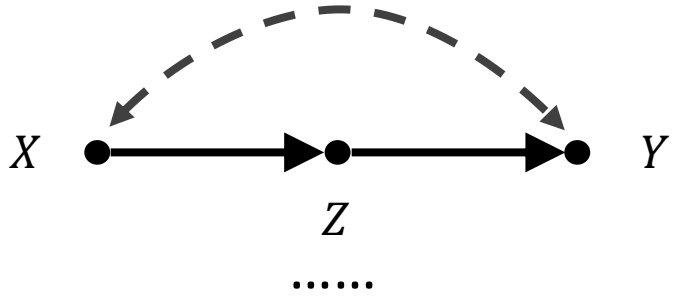
Intervention in Semi-Markovian Model

- Graphical criterion of identification:
  - Sufficient condition: *back-door criterion*



$P(y \mid do(x))$ is identifiable if exists a set of observed variables $\boldsymbol{Z}$ that blocks all back-door paths from X to Y

  - Sufficient condition: *front-door criterion*



$P(y \mid do(x))$ is identifiable if exists a set of observed variables $\boldsymbol{Z}$ such that:

- $\boldsymbol{Z}$ blocks all causal paths from X to Y;
- There is no back-door path from X to $\boldsymbol{Z}$;
- All back-door paths from $\boldsymbol{Z}$ to Y are blocked by X.

  - Complete criterion: *hedge criterion* (Shpitser et al., 2008)

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

Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* 9(Sep), 1941-1979 (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

Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: NIPS'17 (2017)
Zennaro, F.M., Ivanovska, M.: Pooling of causal models under counterfactual fairness via causal judgement aggregation. Preprint (2018)

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).

Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
Zhang, J., Bareinboim, E.: Fairness in decision-making – the causal explanation formula. In: AAAI'18 (2018)

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.,

$$\textit{Path-specific counterfactual quantity } (Y_{x|_\pi} \mid \boldsymbol{e})$$

- 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

$$\min_{h \in \mathcal{H}} \mathbb{L}(h)$$
$$\text{s.t. } \mathbb{C}(h) \leq \tau$$

Minimize the loss function

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?

Fairness Constraints for Classification

- Classification fairness is measured using risk difference

$$\mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}|S=s^+}[\mathbb{1}_{f(\mathbf{x})=1}] - \mathbb{E}_{\mathbf{X}|S=s^-}[\mathbb{1}_{f(\mathbf{x})=1}]$$

- Learn a classifier with fairness constraints

$$\min_{h\in\mathcal{H}} \quad \mathbb{L}_\phi(h)$$
$$\textit{subject to} \quad \mathbb{RD}_\kappa(h) \le c_1, \quad -\mathbb{RD}_\delta(h)$$

Minimize the loss function

Subject to fairness constraints

  - For computational feasibility, the loss function, fairness constraints are surrogated by convex/concave functions ϕ, κ, δ.

- Bounding fairness constraints with surrogate function

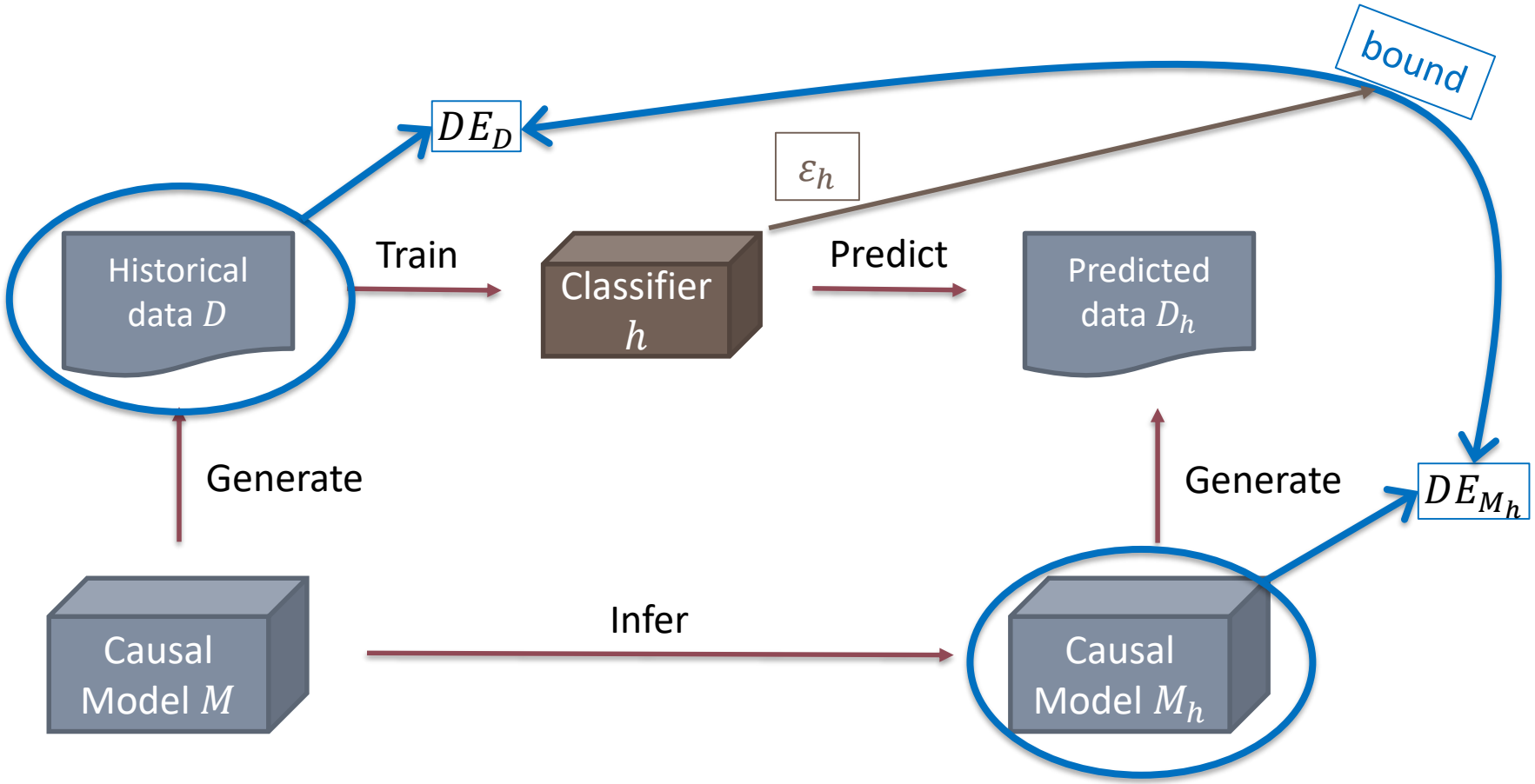
$$\min_{h\in\mathcal{H}} \quad \mathbb{L}_\phi(h)$$
$$\textit{subject to} \quad \mathbb{RD}_\kappa(h) \le \psi_\kappa(\tau - \mathbb{RD}^-) + \mathbb{RD}_\kappa^-$$
$$-\mathbb{RD}_\delta(h) \le \psi_\delta(\tau + \mathbb{RD}^+) + \mathbb{RD}_\delta^+$$

$-\tau \le RD(h^*) \le \tau$ is guaranteed

Wu, Y., Zhang, L., Wu, X.: Fairness-aware classification: Criterion, convexity, and bounds. Preprint (2018)

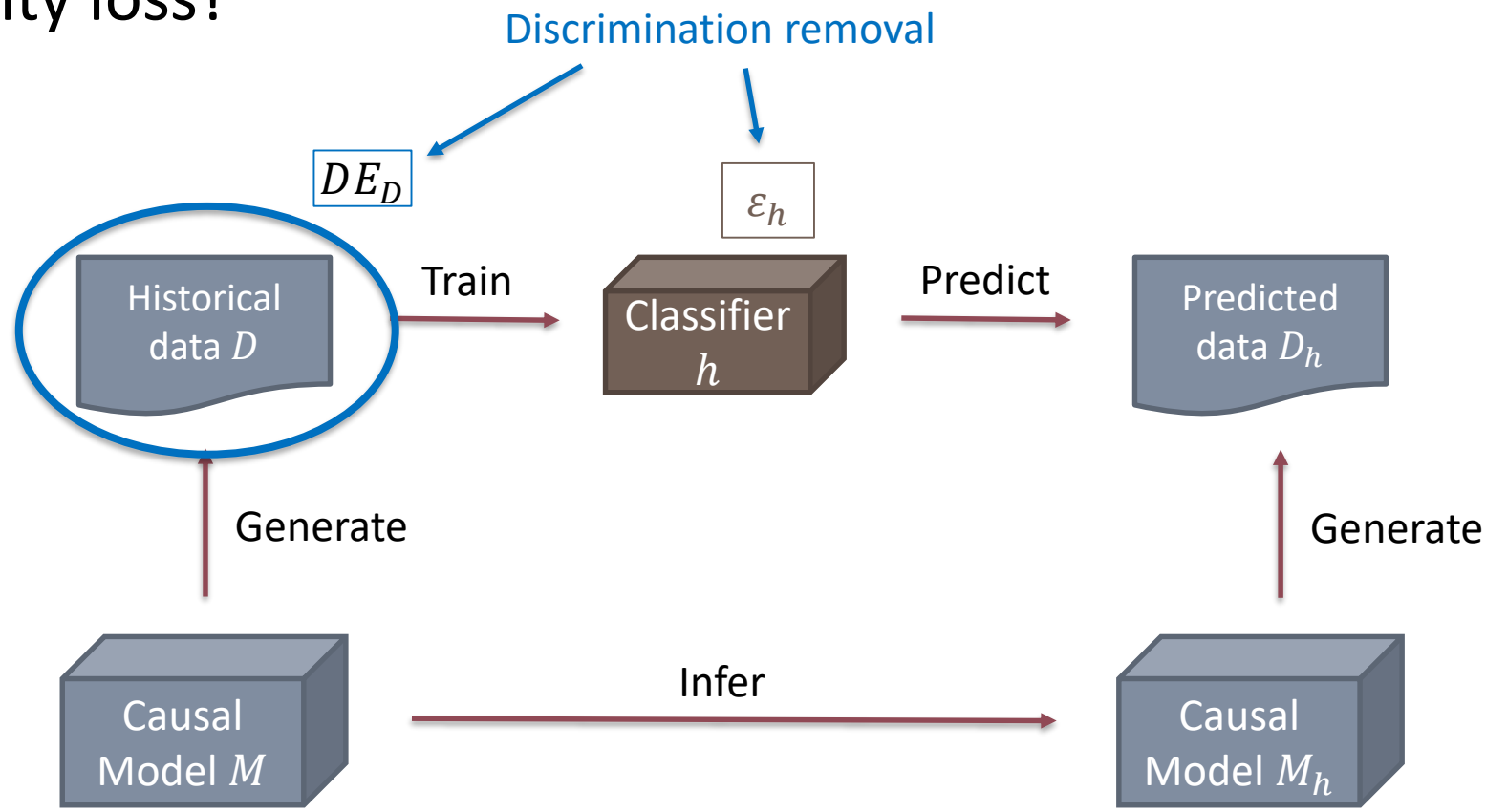
Direct/Indirect Discrimination: Data vs. Model

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



Trade-Off

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



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
  - Serbos, D., Qi, S., Mamoulis, N., Pitoura, E., Tsaparas, P.: Fairness in package-to-group recommendations. In: WWW'17 (2017)
  - Lin, X., Zhang, M., Zhang, Y., Gu, Z., Liu, Y., Ma, S.: Fairness-aware group recommendation with pareto-efficiency. In: RecSys '17 (2017)
  - Yao, S., Huang, B.: Beyond Parity: Fairness objectives for collaborative filtering. In: NIPS'17 (2017)
  - Burke, R., Sonboli, N. Ordonez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: FAT*'18 (2018)
- No causal modeling based method

Fairness-aware Ranking

- Fairness-aware ranking
  - Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: A fair top-k algorithm. In: CIKM ’17 (2017)
  - Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: SSDBM ’17 (2017)
  - Celis, L. E., Straszak, D., Vishnoí, N. K.: Ranking with fairness constraints. In: ICALP’18 (2018)
  - Singh, A., Joachims, T.: Fairness of exposure in rankings. In: SIGKDD’18 (2018)
  - Asudeh, A., Jagadish, H. V, Stoyanovich, J., Das, G.: Designing fair ranking schemes. Preprint (2018)
  - Wu, Y., Zhang, L., Wu, X.: On discrimination discovery and removal in ranked data using causal graph. In: SIGKDD’18 (2018)

Fair Ranking

- Decisions are given in permutation rather than binary decisions.

ID	Race	Zip Code	Interview	Education	Rank
U1	W	72701	1	1	10
U2	W	72701	2	2	8
U3	W	72701	2	1	9
U4	W	72701	4	2	7
U5	W	72701	2	4	6
U6	B	72701	5	5	1
U7	B	72702	4	4	3
U8	B	72702	4	5	2
U9	B	72701	3	3	5
U10	B	72702	2	5	4

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.

Wu, Y., Zhang, L., Wu, X.: On discrimination discovery and removal in ranked data using causal graph. In: SIGKDD'18 (2018)

Fair Ranking

ID	Race	Zip Code	Interview	Education	Rank
U1	W	72701	1	1	10
U2	W	72701	2	2	8
U3	W	72701	2	1	9
U4	W	72701	4	2	7
U5	W	72701	2	4	6
U6	B	72701	5	5	1
U7	B	72702	4	4	3
U8	B	72702	4	5	2
U9	B	72701	3	3	5
U10	B	72702	2	5	4

Bradley-Terry Model

Score
0.00
25.45
14.13
35.62
45.26
100.00
74.54
85.86
54.74
64.38

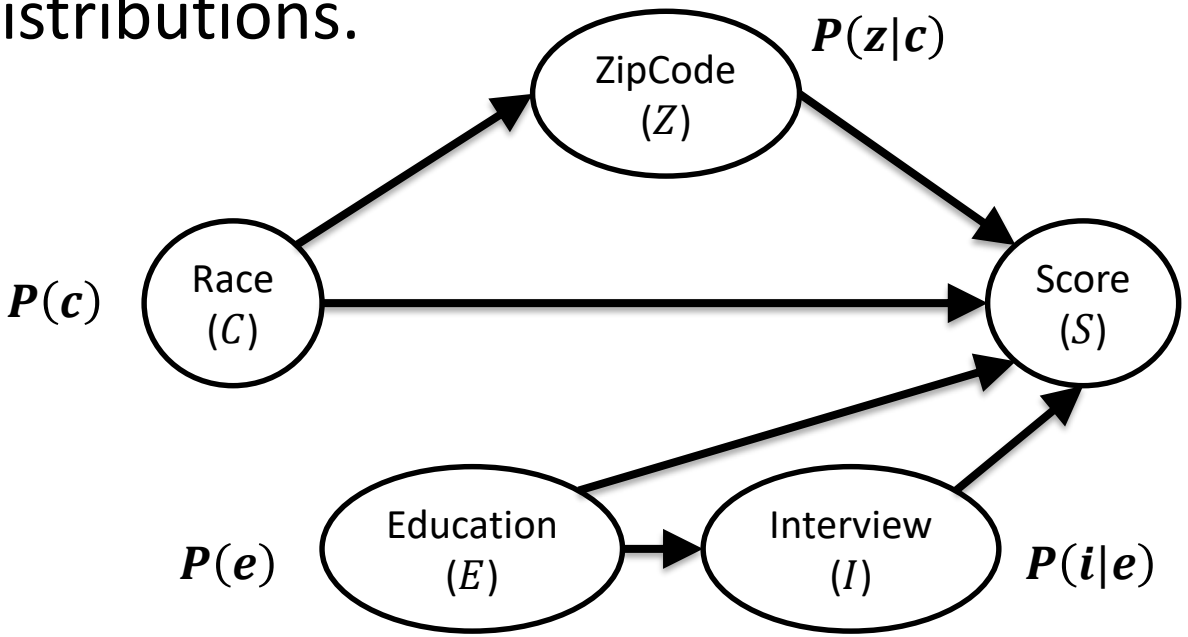
Fair Ranking

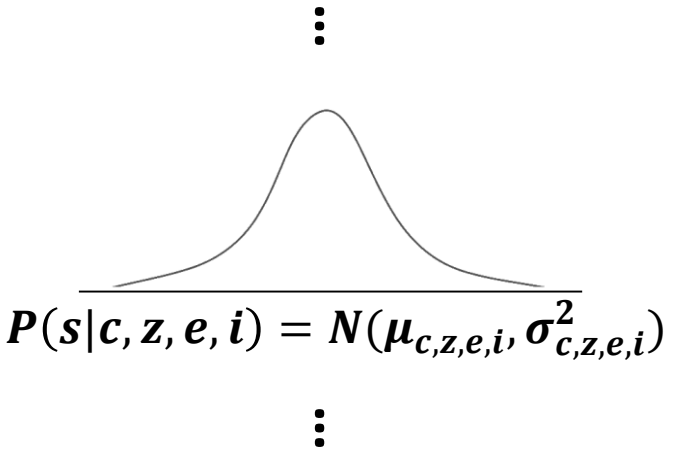
- Map ranking positions to continuous score using Bradley-Terry Model.

$$s_i - s_j = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right),$$

$$P(\omega|\mathcal{M}) \propto \sum_{(i,j):\omega_i<\omega_j} p_{ij}$$

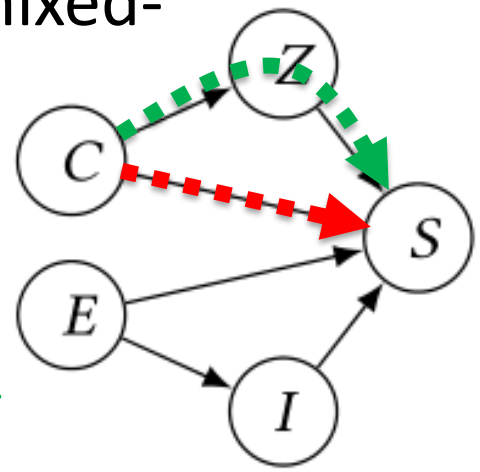
- Build a mixed-variable causal graph using conditional Gaussian distributions.





Fair Ranking

- Derive direct and indirect discrimination measure in mixed-variable causal graph.



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

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

- Identify the relationship between discrimination in ranking and discrimination in binary decision.
  - Assume that the decision is made based on a cut-off point θ of the score. If then $\boldsymbol{\theta \geq \mu_{c^+,q} \geq \mu_{c^-,q}}$,

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

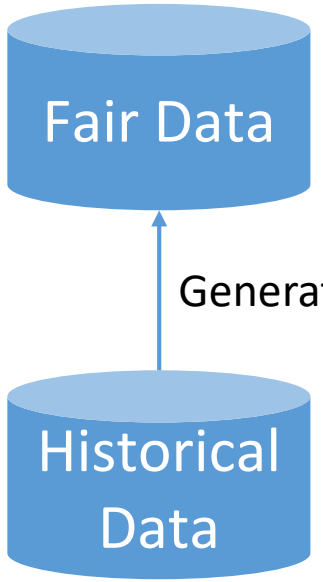
$$SE^{Bin}_{\pi_i} \leq \tau \text{ implies } SE_{\pi_i} \leq \frac{2\sqrt{2}(\tau-c)\sigma}{\alpha}$$

Fair Generative Adversarial Networks

- Fair Generative Adversarial Networks
  - Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: fairness-aware generative adversarial networks. Preprint (2018)
  - Sattigeri, Prasanna, Hoffman, Samuel C., Chenthamarakshan, Vijil, Varshney, Kush R.: Fair GAN. Preprint (2018)

Fairness-aware Generative Adversarial Networks (FairGAN)

- 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.



Directly generate fair data with no risk difference

Xu, D., et al.: FairGAN: Fairness-aware generative adversarial networks. Preprint (2018)

Fair Data Generation

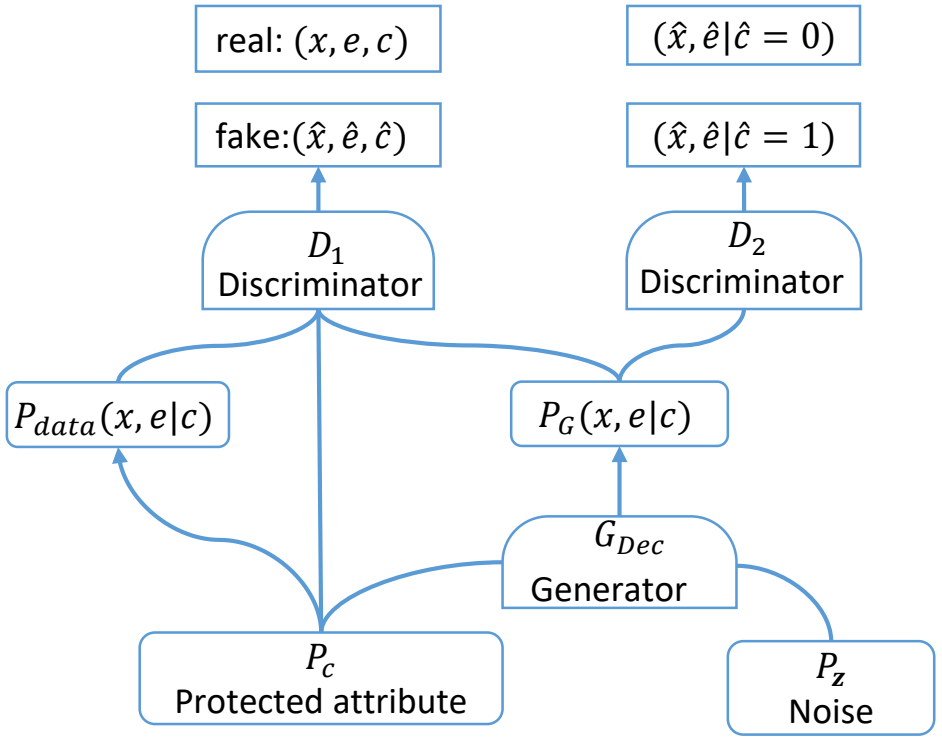
- The first minimax game ensures the generated data close to the real data
- The second minimax game ensures fairness by removing the correlation between unprotected attributes, decision and the protected attributes

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

where

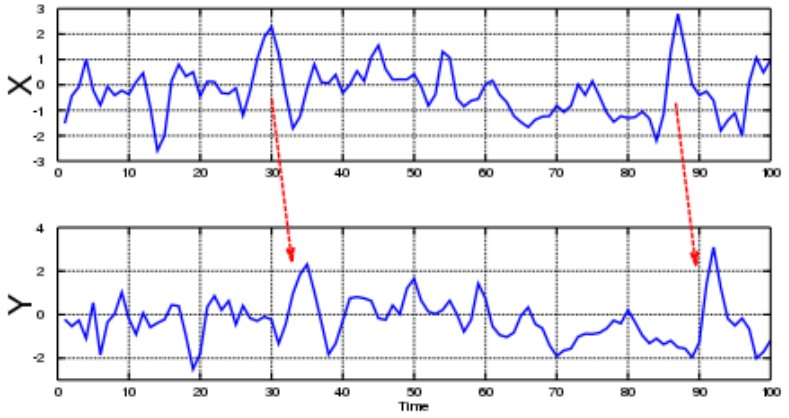
$$V_1(G_{Dec}, D_1) = \mathbb{E}_{c \sim P_{data}(c),(x,e) \sim P_{data}(x,e|c)}[\log D_1(x, e, c)] + \mathbb{E}_{\hat{s} \sim P_G(s),(\hat{x},\hat{e}) \sim P_G(x,e|c)}[\log(1 - D_1(\hat{x}, \hat{e}, \hat{c}))]$$

$$V_2(G_{Dec}, D_2) = \mathbb{E}_{(\hat{x},\hat{e}) \sim P_G(x,e|c=1)}[\log D_2(\hat{x}, \hat{e})] + \mathbb{E}_{(\hat{x},\hat{e}) \sim P_G(x,e|c=0)}[\log(1 - D_2(\hat{x}, \hat{e}))]$$



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**

UNIVERSITY OF ARKANSAS

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

1. Tian, J., Pearl, J.: Probabilities of causation: Bounds and identification. In: UAI'00 (2000)
2. Avin, C., Shpitser, I., Pearl, J.: Identifiability of path-specific effects. In: IJCAI'05 (2005)
3. Madsen, A.L.: Belief update in CLG Bayesian networks with lazy propagation. Int. J. Approx. Reason. 49(2),503-521 (2008)
4. Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. J. Mach. Learn. Res. 9(Sep), 1941-1979 (2008)
5. Kamiran, F., Calders, T.: Classifying without discriminating. IC4'09 (2009)
6. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21(2), 277-292 (2010)
7. Kamishima, T., Akaho, S., and Sakuma J.: Fairness-aware learning through regularization approach. In: ICDMW'11 (2011)
8. Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: SIGKDD'11 (2011)
9. Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM'11 (2011)
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS'12 (2012)
11. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33(1), 1–33 (2012)
12. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: ICDM'12 (2012)
13. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. 25(7), 1445–1459 (2013)
14. Magnani, L., Board, E., Longo, G., Sinha, C., Thagard, P.: Discrimination and privacy in the information society. Springer (2013)
15. Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML'13 (2013)
16. Mancuhan, K., Clifton, C.: Combating discrimination using Bayesian networks. Artif. Intell. Law 22(2), 211–238 (2014)
17. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowl. Eng. Rev. 29(05), 582–638 (2014)

References

18. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15 (2015)
19. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination-and privacy-aware patterns. Data Min. Knowl. Discov. 29(6), 1733-1782 (2015)
20. Edwards, H., Storkey, A.: Censoring representations with an adversary. In: ICLR'16 (2016)
21. Hardt M., Price E., Srebro N.: Equality of opportunity in supervised learning. In: NIPS'16 (2016)
22. Wu,Y., Wu,X.: Using loglinear model for discrimination discovery and prevention. In: DSAA'16 (2016)
23. Zhang, L., Wu, Y., Wu, X.: On discrimination discovery using causal networks. In: SBP-BRiMS 2016 (2016)
24. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)
25. Barocas, S., Hardt, M.: Fairness in machine learning. Tutorial. In: NIPS'17 (2017)
26. Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. Int. J. Data Sci. Anal. 3(1), 1–21 (2017)
27. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: SIGKDD'17 (2017)
28. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: NIPS'17 (2017)
29. Kleinberg, J., Mullainathan, S., & Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: ITCS'17 (2017)
30. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: NIPS'17 (2017)
31. Lin, X., Zhang, M., Zhang, Y., Gu, Z., Liu, Y., Ma, S.: Fairness-aware group recommendation with pareto-efficiency. In: RecSys '17 (2017)
32. Pearl, J.: A linear "microscope" for interventions and counterfactuals. Journal of Causal Inference, 5(1). (2017)

References

33. Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: NIPS'17 (2017)
34. Serbos, D., Qi, S., Mamoulis, N., Pitoura, E., Tsaparas, P.: Fairness in package-to-group recommendations. In: WWW'17 (2017)
35. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: SSDBM '17 (2017)
36. Yao, S., Huang, B.: Beyond Parity: Fairness objectives for collaborative filtering. In: NIPS'17 (2017)
37. Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW'17 (2017)
38. Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P.: Fairness constraints: Mechanisms for fair classification. In: AISTAS'17 (2017)
39. Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., Weller, A. : From parity to preference-based notations of fairness in classification. In: NIPS'17 (2017)
40. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: A fair top-k algorithm. In: CIKM '17 (2017)
41. Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. 4(1), 1-16 (2017)
42. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
43. Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)
44. Burke, R., Sonboli, N. Ordonez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: FAT*'18 (2018)
45. Celis, L. E., Straszak, D., Vishnoí, N. K.: Ranking with fairness constraints. In: ICALP'18 (2018)
46. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: CausalGAN: Learning causal implicit generative models with adversarial training. In: ICLR'18 (2018)

References

47. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: ICML'18 (2018)
48. Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI'18 (2018)
49. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: SIGKDD'18 (2018)
50. Wu, Y., Zhang, L., Wu, X.: On discrimination discovery and removal in ranked data using causal graph. In: SIGKDD'18 (2018)
51. Zhang, J., Bareinboim, E.: Fairness in decision-making – the causal explanation formula. In: AAAI'18 (2018)
52. Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. In: IJCAI'18 (2018)
53. Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S.: Causal discrimination discovery through propensity score analysis. Preprint (2016)
54. Asudeh, A., Jagadish, H. V, Stoyanovich, J., Das, G.: Designing fair ranking schemes. Preprint (2018)
55. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. Preprint (2018)
56. Sattigeri, Prasanna, Hoffman, Samuel C., Chenthamarakshan, Vijil, Varshney, Kush R.: Fair Gan. Preprint (2018)
57. Wu, Y., Zhang, L., Wu, X.: Fairness-aware classification: Criterion, convexity, and bounds. Preprint (2018)
58. Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: fairness-aware generative adversarial networks. Preprint (2018)
59. Zennaro, F.M., Ivanovska, M.: Pooling of causal models under counterfactual fairness via causal judgement aggregation. Preprint (2018)
60. Zhang, L., Wu, Y., Wu, X.: Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. IEEE Trans. Knowl. Data Eng., under review (2018)