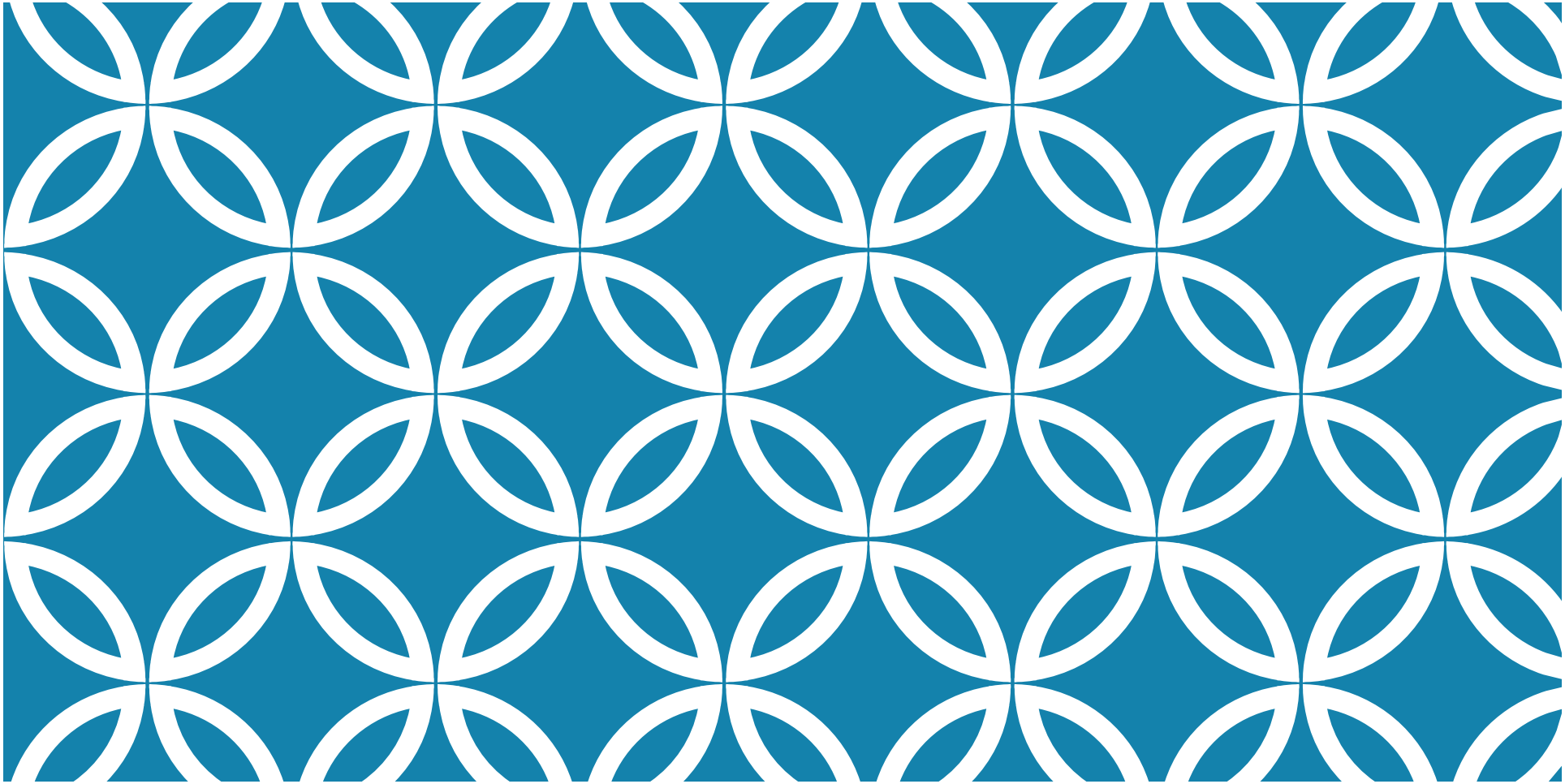


ON DISCRIMINATION DISCOVERY USING CAUSAL NETWORKS

Lu Zhang
Yongkai Wu
Xintao Wu
University of Arkansas

OUTLINE

Introduction

Related Work

Methods

Preliminary Results

Conclusions and Future Work

INTRODUCTION

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

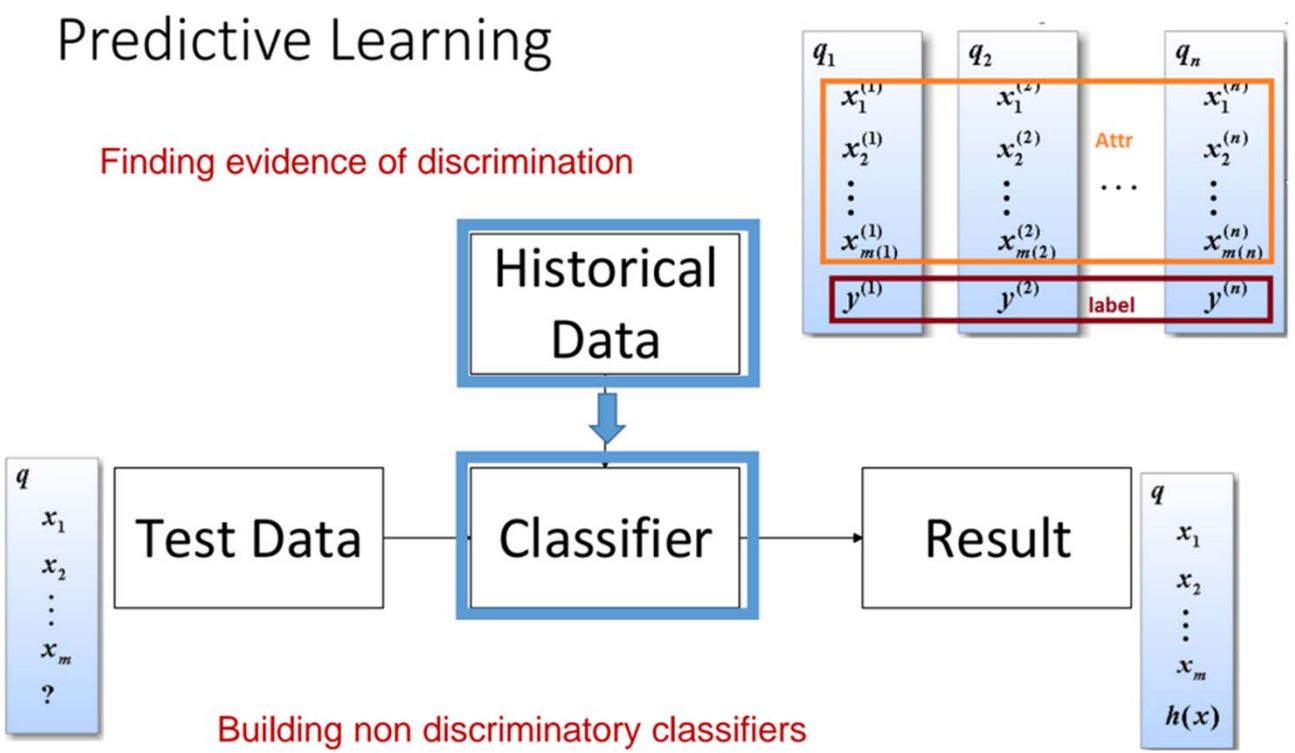
Federal Laws and regulations disallow discrimination on several grounds:

- Gender, Age, Marital Status, Sexual Orientation, Race, Religion or Belief, Disability or Illness ……
- These attributes are referred to as the protected attributes.

In 2014, US President Obama called for a 90-day review of big data collecting and analyzing practices.

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

INTRODUCTION



RELATED WORK

Several types of discrimination have proposed in the literature. Accordingly, different types discrimination discovery techniques have been developed.

- Group discrimination (association rules), individual discrimination (situation testing), direct discrimination (conditional discrimination analysis), and indirect discrimination (rule inference), etc.

The down side:

- Current approaches are mainly based on correlation or association. In discrimination discovery, it is critical to derive causal relationship, and not merely association relationship.
- We need a unifying framework for determining all types of discrimination rather than using different types of data mining techniques.

METHODS

For a quantitative measurement, we use risk difference to measure discrimination.

- The difference in the proportion of positive decisions between the protected group and the non-protected group.
- Formally, $\Delta P \mid s = P(e+ \mid c+, s) - P(e+ \mid c-, s)$.
  - E: decision attribute (admission); e+: positive decision (admitted); e-: negative decision (rejected).
  - C: protected attributes (gender); c+: non-protected group (male); c-: protected group (female).
  - S: A set of attributes that represents a partition; s: a subpopulation under the partition.
- Discriminatory effect presents if $\Delta P \mid s > \tau$, where τ is a threshold for discrimination depending on law (e.g., 5%).

The key to discrimination discovery: find meaningful partitions.

ILLUSTRATIVE EXAMPLES

In a university admission system: gender, major, test score, admission.

TABLE I: Summary statistics of Example 1.

test score	L				H			
gender	female		male		female		male	
major	CS	EE	CS	EE	CS	EE	CS	EE
No. applicants	450	150	150	450	300	100	100	300
admission rate	20%	40%	20%	40%	50%	70%	50%	70%
	25%		35%		55%		65%	

Overall admission rate:
Female: 37%
Male: 47%

TABLE II: Summary statistics of Example 2.

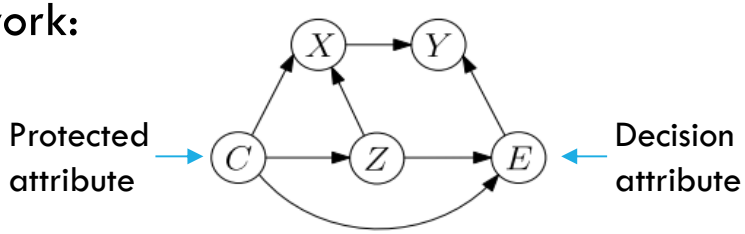
major	CS				EE			
gender	female		male		female		male	
test score	L	H	L	H	L	H	L	H
No. applicants	450	300	150	100	600	300	200	100
admission rate	30%	50%	36%	40%	40%	60%	45%	50%
	38%		38%		47%		47%	

Overall admission rate:
Female: 43%
Male: 43%

We use the causal network to find the meaningful partitions.

CAUSAL NETWORK (CAUSAL GRAPH, CAUSAL BAYESIAN NETWORK, ETC.)-BASED APPROACH

The causal network:



Causal effects are carried by the paths that trace arrows pointing from the cause to the effect.

In the causal network theory, a path can be blocked by conditioning on a set of nodes not containing the two end-nodes.

- Upon blocked, the effect originally transmitted through the path is suppressed in each subpopulation under the partition defined by the set of nodes.
- Use path blocking to identify the causal effects that we want to measure for discrimination discovery.

The criterion of blocking a path is given in the *d*-separation criterion.

- E.g., the path C → Z → E is blocked when conditioning on Z.

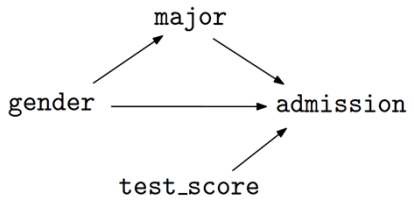
DISCRIMINATION CATEGORIZATION

Different criteria are needed for different types of discrimination.

We categorize discrimination into:

	system-level (e.g., against the university)	group-level (e.g., against a department)	individual-level
direct discrimination	two criteria		
indirect discrimination			

SYSTEM-LEVEL DIRECT DISCRIMINATION



Direct: the direct causal effect carried by C → E.

- The direct causal effect can be identified under the partition determined by a certain set of attributes, which we refer to as the block set.
  1. All paths other than C → E are blocked.
  2. Does not contain E's decedents.
- For each block set B (defines a partition) and its value assignment b (defines a subpopulation), $\Delta P \mid b = P(e+ \mid c+, b) - P(e+ \mid c-, b)$ can be used to measure the direct causal effect.

System-level: the discriminatory effect should be measured as a persistent negative effect under all meaningful partitions.

- Discriminatory effect needs to measured for each value assignment b of each block set B.

SYSTEM-LEVEL DIRECT DISCRIMINATION

Rigorous discrimination criterion: Discrimination is considered to present if inequality $\Delta P \mid b > \tau$ holds for each value assignment b of each block set B.

- In real situations, $\Delta P \mid b$ may vary due to randomness and small sample size.

Relax (τ,α)-discrimination criterion: Discrimination is considered to present if $Pr(\Delta P \mid B > \tau) > \alpha$, where α is a user-defined parameter.

- Can be estimated using the Cantelli's inequality: $Pr(\Delta P \mid B > \tau) > \alpha$ if

$$\mu_B - \sqrt{\frac{\alpha}{1-\alpha}} \cdot \sigma_B > \tau$$

SYSTEM-LEVEL DIRECT DISCRIMINATION

The previous two criteria require to examine all block sets. A brute force algorithm of finding all block sets has an exponential complexity.

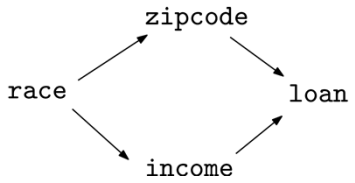
It is proved that, it is sufficient to examine Q instead of examining each block set, where Q is the set of E's all parents except C.

Rigorous discrimination criterion: Discrimination is considered to present if inequality $\Delta P \mid q > \tau$ holds for each value assignment q of set Q.

Relax (τ,α)-discrimination criterion: Discrimination is considered to present if

$$\mu_Q - \sqrt{\frac{\alpha}{1-\alpha}} \cdot \sigma_Q > \tau$$

SYSTEM-LEVEL INDIRECT DISCRIMINATION



Concerns about the indirect causal effects that are transmitted through intermediate attributes along the causal paths from C to E other than the direct arc C $\rightarrow$ E.

- Not all of them.

We refer to the attributes on the causal paths whose usage cannot be legally justified as the redlining attributes.

We propose to measure the indirect causal effect through the paths that each contains at least one redlining attribute.

Similarly, system-level indirect discrimination can be claimed if persistent negative effects are measured.

GROUP-LEVEL DIRECT DISCRIMINATION

Occurs in a particular subsystem other than across the whole system.

- The group G can be specified by analysts to denote a subsystem, determined by a subset of profiling attributes.
- E.g., when we determine whether there exists group-level discrimination in a particular major (e.g., CS) in university admission, G contains all applicants in CS.

The determination of block set B needs to be adjusted based on the given group G to form a partition within the given group.

- E.g., when focusing on group-level discrimination in CS major, B may contain test scores.
- For each test score b, group-level discrimination can be claimed after we examine $\Delta P|b$ across all test scores among CS applicants.

INDIVIDUAL-LEVEL DIRECT DISCRIMINATION

To identify discrimination for a specific individual, i.e., an entire record in the dataset.

Find two neighborhood groups that contain similar individuals from the protected group and the non-protected group. The individual is considered as discriminated if significant difference is observed between the decisions from the two groups.

Use the causal network to find similar individuals.

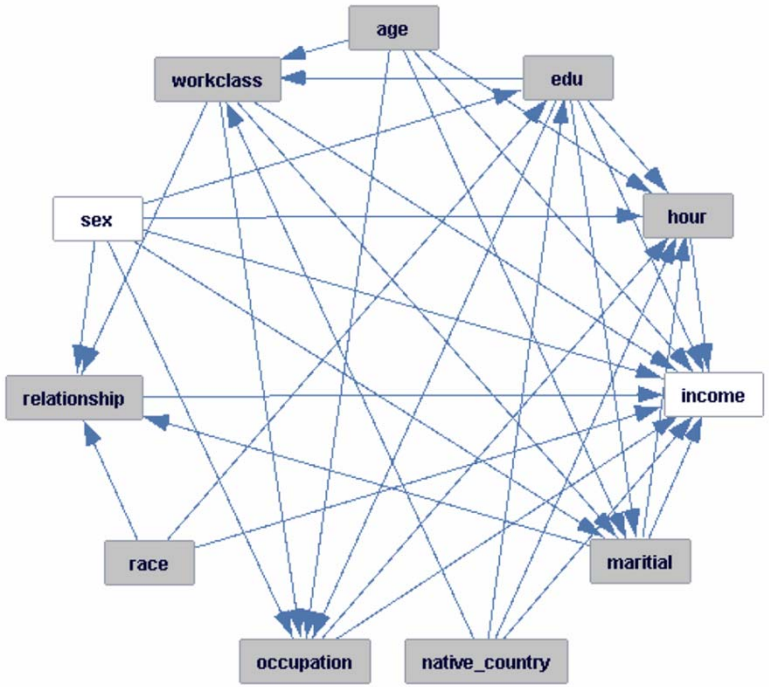
The causal structure and the causal effect of each attribute on the decision attribute are used for the calculation of inter-individual distances.

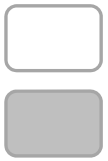
PRELIMINARY RESULTS

Apply the rigorous discrimination criterion and the relax (τ,α)-discrimination criterion to discovery system-level direct discrimination.

- Datasets: Adult dataset and Dutch Census 2001
- The causal network is built using the PC algorithm on TETRAD.
- τ is set as 0.05.
- α is set as 60%.

PRELIMINARY RESULTS (ADULT)





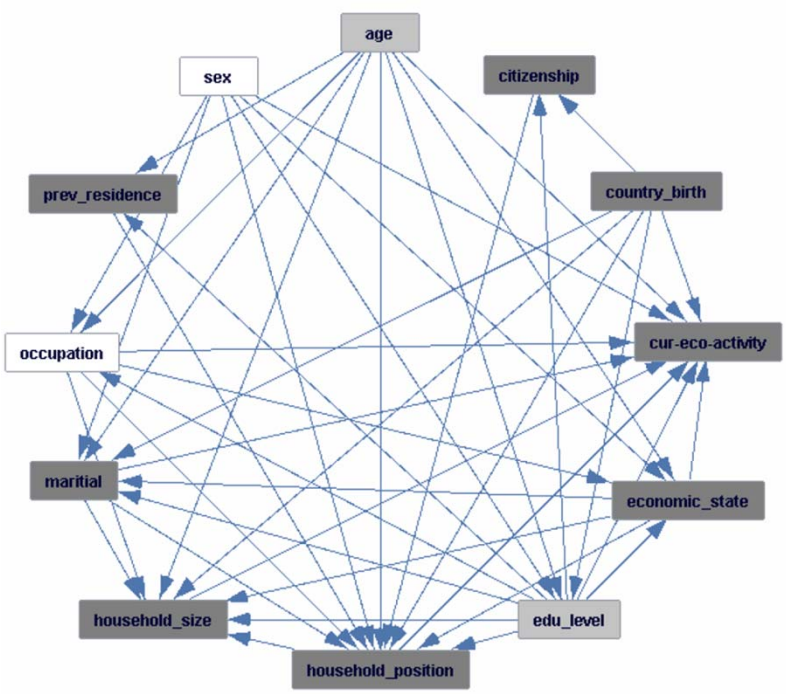
Protected and decision attributes

Parents of decision attribute

Rigorous: the values of $\Delta P|q$ range from -0.614 to 0.524

Relaxed: Pr($\Delta P|Q > 0.05$) > 0.09% according to the Cantelli's inequality

PRELIMINARY RESULTS (DUTCH)



Protected and decision attributes

Parents of decision attribute

Rigorous: the values of $\Delta P|q$ range from 0.062 to 0.435

Relaxed: Pr($\Delta P|Q > 0.05$) > 65.48% according to the Cantelli's inequality

CONCLUSIONS AND FUTURE WORK

Studied the problem of discrimination discovery.

Categorize different discrimination types based on discrimination scale and discrimination manner.

Based on the causal network, established two criteria for system-level direct discrimination: a rigorous criterion and relaxed criterion.

In the future work: extend the results to other types of discrimination, and study the problem of discrimination prevention, which aim to remove discrimination by modifying the data before conducting predictive analysis.

THANK YOU!

Acknowledgment

- This work was supported in part by U.S. National Institute of Health (1R01GM103309).

Contact

- Lu Zhang, Email: *lz006@uark.edu*