Causal Multi-level Fairness

Vishwali Mhasawade[1], Rumi Chunara[1,2]

[1]Department of Computer Science and Engineering, Tandon School of Engineering, New York University

[2]Department of Biostatistics, College of Global Public Health, New York University

Motivation

- Algorithmic fairness approaches are limited to sensitive attributes at *individual level*.
- However, *Critical Theory* motivates the consideration of structural or macro-properties to understand social disparities.
- We propose a novel definition of fairness – *'causal multi-level fairness'* that accounts for both macro and individual properties to mitigate unfairness.

Introduction

- Macro-attribute shapes the resources and opportunities an individual may have, and algorithms need to mitigate any historical unfairness (Furze and Savy, 2014).
- Algorithmic fairness approaches have only considered unfairness due to individual-level sensitive attributes.

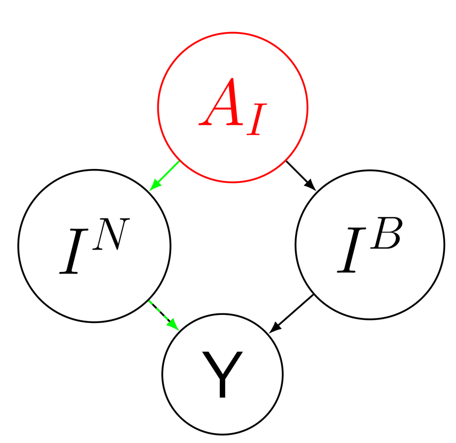

Figure 1: Proxy (I^N) for sensitive attributes (A_I)

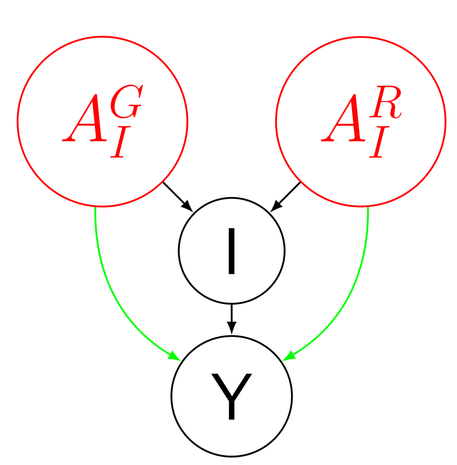

Figure 2: Multiple sensitive attributes A_I^G, A_I^R

Example causal graphs with sensitive variables represented by red nodes.

Problem Setup

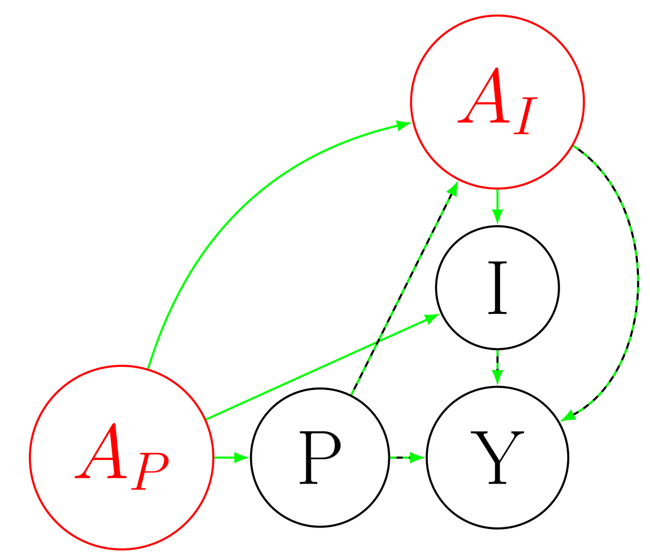

Figure 3: Macro-level variables, A_P (e.g. neighborhood SES), P (e.g. zipcode), and individual-level ones, A_I (e.g. perception of race), I (e.g. biological factors), affect the outcome Y (e.g. health outcomes).

Approach

- We assume access to a causal graph representing the data-generating process and knowledge of unfair pathways.
- The aim is to remove effects of sensitive variables along the unfair paths.
- We adopt counterfactual fairness to multi-level sensitive attributes (Kusner et al., 2017; Chiappa, 2019).
- We identify multi-level path-specific effects (PSE) along unfair pathways (Shpitser, 2013).
- Fair classifier: $\hat{Y}_{\text{fair}} = \hat{Y} - \text{PSE}$.

Our contribution

A multi-level fairness approach to mitigate unfairness while accounting for macro and individual-level sensitive attributes.

Results

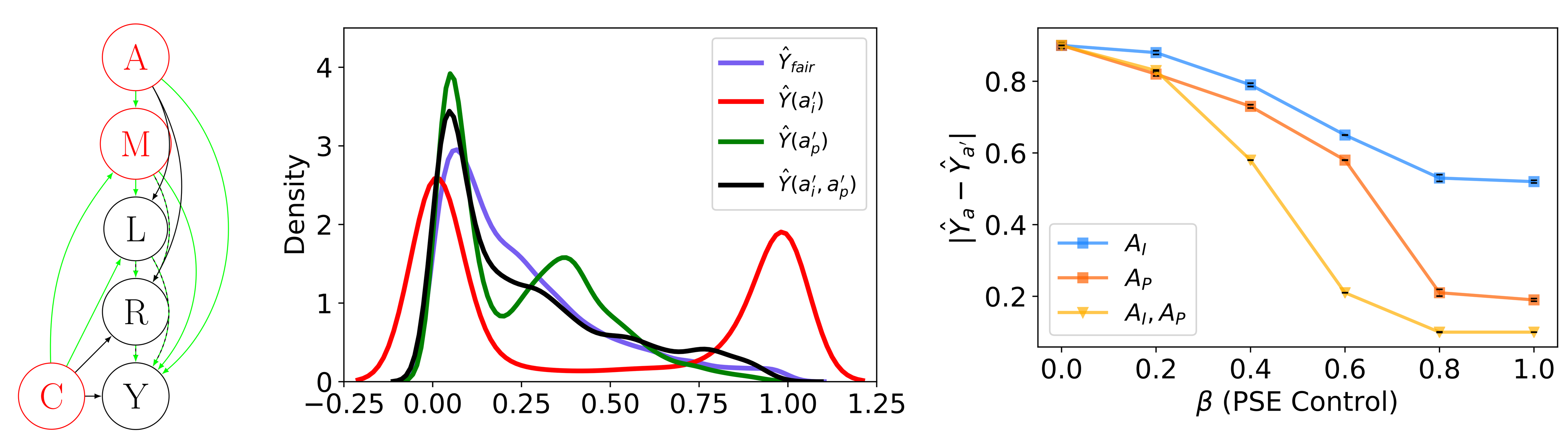

Figure 4: *Left*: Causal graph for the UCI Adult dataset (Chiappa, 2019), A and M represent the individual-level protected attributes, sex and marital status, respectively, C is nationality, L is the level of education, R corresponds to working class, occupation, and hours per week, Y is the income class, unfair paths are represented in green, *Center:* Density of $\hat{Y}$, *Right:* path-specific unfairness, $|\hat{Y}_a - \hat{Y}_{a'}|$ controlling for the effects of just A_I (blue), A_P (orange) and both A_I, A_P (yellow).

Conclusion

- Our work extends algorithmic fairness to account for the multi-level and socially-constructed nature of forces that shape unfairness.
- A framework like this can be used to assess unfairness at each level, and identify the places for intervention that would reduce unfairness best (e.g. via macro-level policies versus individual attributes).
- We illustrate the importance of accounting for macro-level sensitive attributes by exhibiting residual unfairness if they are not accounted for.

References

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Brian Furze and Pauline Savy. *Sociology in today's world.* Cengage AU, 2014.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

William Little, Ron McGivern, and Nathan Kerins. *Introduction to Sociology-2nd Canadian Edition.* BC Campus, 2016.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

Acknowledgements

We acknowledge funding from the National Science Foundation, award number 1845487.

NEW YORK UNIVERSITY