Fairness Interventions as (Dis)Incentives for Strategic Manipulation

Xueru Zhang\textsuperscript{1}, Mohammad Mahdi Khalili\textsuperscript{2}, Kun Jin\textsuperscript{3}, Parinaz Naghizadeh\textsuperscript{1}, Mingyan Liu\textsuperscript{3}

**Model: strategic interaction**

Two demographic groups $G_a, G_b$

- Sensitive attribute $S \in \{a, b\}$
- Features $X \in \mathbb{R}^d$
  - Feature generation $P_{X|S}(x|y, s)$
- Qualification state $Y \in \{0, 1\}$
  - Qualification rate $\alpha_s = P_{Y|S}(1|s)$
- Decision $D \in \{0, 1\}$
  - Policy $\pi_s(x) = P_{D|X,S}(1|x, s)$
- Manipulation action $M \in \{0, 1\}$
  - Manipulation doesn’t affect $Y$ but results in better feature distribution
  - Manipulation cost $C_s \geq 0$

**Individual best response**

- Individual chooses to manipulate at cost if manipulation brings the higher utility
- Manipulation probability:
  $$\Pr(C_s \leq P_{D|Y_M,S}(1|y, 1, s) - P_{D|Y_M,S}(1|y, 0, s))$$

**Decision-maker’s optimal (fair) policies**

- Maximize $\mathbb{E}[R(D, Y)]$
  - Subject to fairness constraint

**Theoretical results**

- Characterize the equilibrium strategies of individuals & decision-maker (four types of policies)
- Impact of decision-maker’s anticipation of strategic manipulation
  - Strategic policy over(under) accepts majority-qualified(majority-unqualified) group
  - Anticipation of manipulation can worsen the fairness of a strategic policy when one group is majority-qualified while the other is majority-unqualified
  - When both groups are majority-unqualified, a strategic policy may mitigate unfairness and even flip the disadvantaged group
- Impact of fairness interventions on policies and individuals’ manipulation
  - Conditions under which non-strategic decision maker may benefit from fairness constraints
  - Conditions under which fairness constraints serve as (dis)incentives for strategic manipulation

**Experiments: FICO credit score**

- Hispanic & Black: strategic policy mitigates unfairness
  - Black & $G_a$: strategic policy worsens unfairness

- More results: