The Intersectionality Problem for Algorithmic Fairness

Johannes Himmelreich [1] Arbie Hsu [2] Kristian Lum [3] Ellen Veomett [2]

[1]Syracuse University [2]University of San Francisco [3]University of Chicago

Problem: Statistical Uncertainty

Intersectionality makes typical fairness definitions meaningless because of statistical uncertainty due to increasingly small subgroups

- **Intersectionality:** Fairness for *subgroups*
  - e.g., for Maghrebi older women in France simultaneously
  - instead of each ethnic origin, age, gender, location separately
- But: Number of intersectional subgroups grows exponentially:
  - $\prod k^n$ (for n k-valued attributes)
- Thus: High **statistical uncertainty** in fairness "metrics"
- Problem: Widely-used definitions of **fairness** become meaningless

$$|m(G) - m(\cdot)| < \epsilon \quad \forall G$$

where $m(\cdot)$ $m(G)$ model performance (however understood) for group G

Solutions: Desiderata

Based on consensus in literature, uncontroversial assumptions

1. **Minimal Justice:** Don't lower fairness standard for certain groups; i.e., "don't disadvantage the disadvantaged"
2. **Incentive Compatibility:** Don't discourage further data collection, and don't encourage deliberate mistakes

Existing Solutions Violate Desiderata

Example: Kearns et al. (2018)

$$\alpha(G)\,|m(G) - m(\cdot)| < \epsilon \quad \forall G$$

where $\alpha(G) = \Pr(G)$, proportion of group G in population

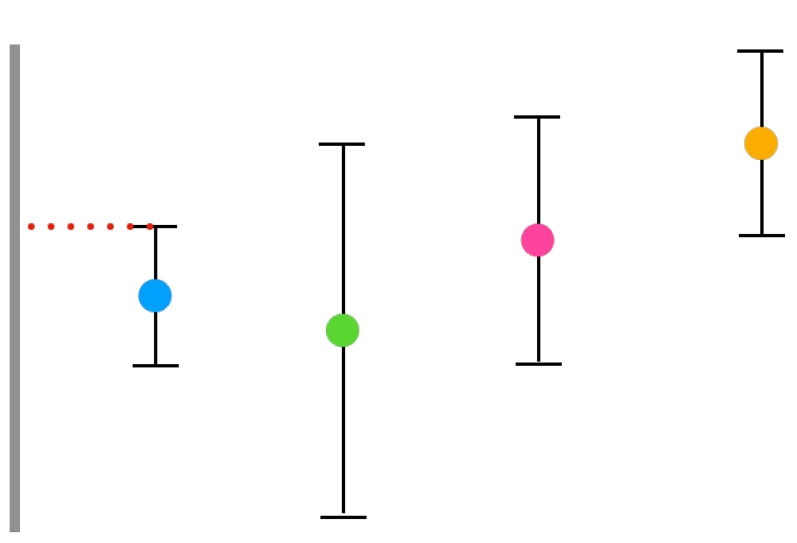
- Violates Minimal Justice: fairness proportional to group size
  - small groups are often disadvantaged, i.e., *less* fairness for them
- Violates Incentive Compatibility
  - discourages minority group data collection (since model subgroup performance is typically lower than current estimate)
  - generally, one *can* improve fairness by making deliberately inaccurate predictions (on group with high model performance)

Alternative: Metrics Based on Hypothesis Testing

Optimist's Metric

Null hypothesis: Model is **fair**

$$H_0 : m(G) > c \quad \forall G$$
$$H_1 : m(G) \le c \quad \exists G$$



Maximal c such that $\forall G$:

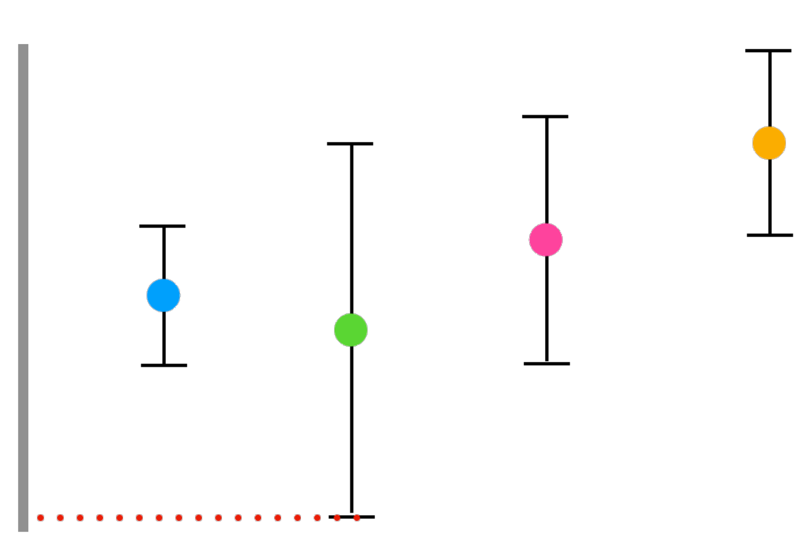
$$c \le m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}$$

Interpretation: Model is 'fair up to c'—likely performs up to c-well for all groups.

Pessimist's Metric

Null hypothesis: Model is **unfair**

$$H_0 : m(G) < c \quad \exists G$$
$$H_1 : m(G) \ge c \quad \forall G$$



Maximal c such that $\forall G$:

$$c \le m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}$$

Interpretation: Model is 'unfair above c'—model likely does not perform at least c'-well for some group at any $c' > c$.

Theoretical Analysis: Meets Desiderata?

Minimal Justice

- Same fairness standard c for all groups
- Fairness as *sufficiency* instead of equality

Incentive Compatibility

- Not susceptible to gaming (no "levelling down") because fairness defined in terms of *absolute* model performance
- Pessimistic metric incentivizes data collection (to reject hypothesis)
- But optimistic metric may *dis*courage data collection on small groups

Empirical Analysis

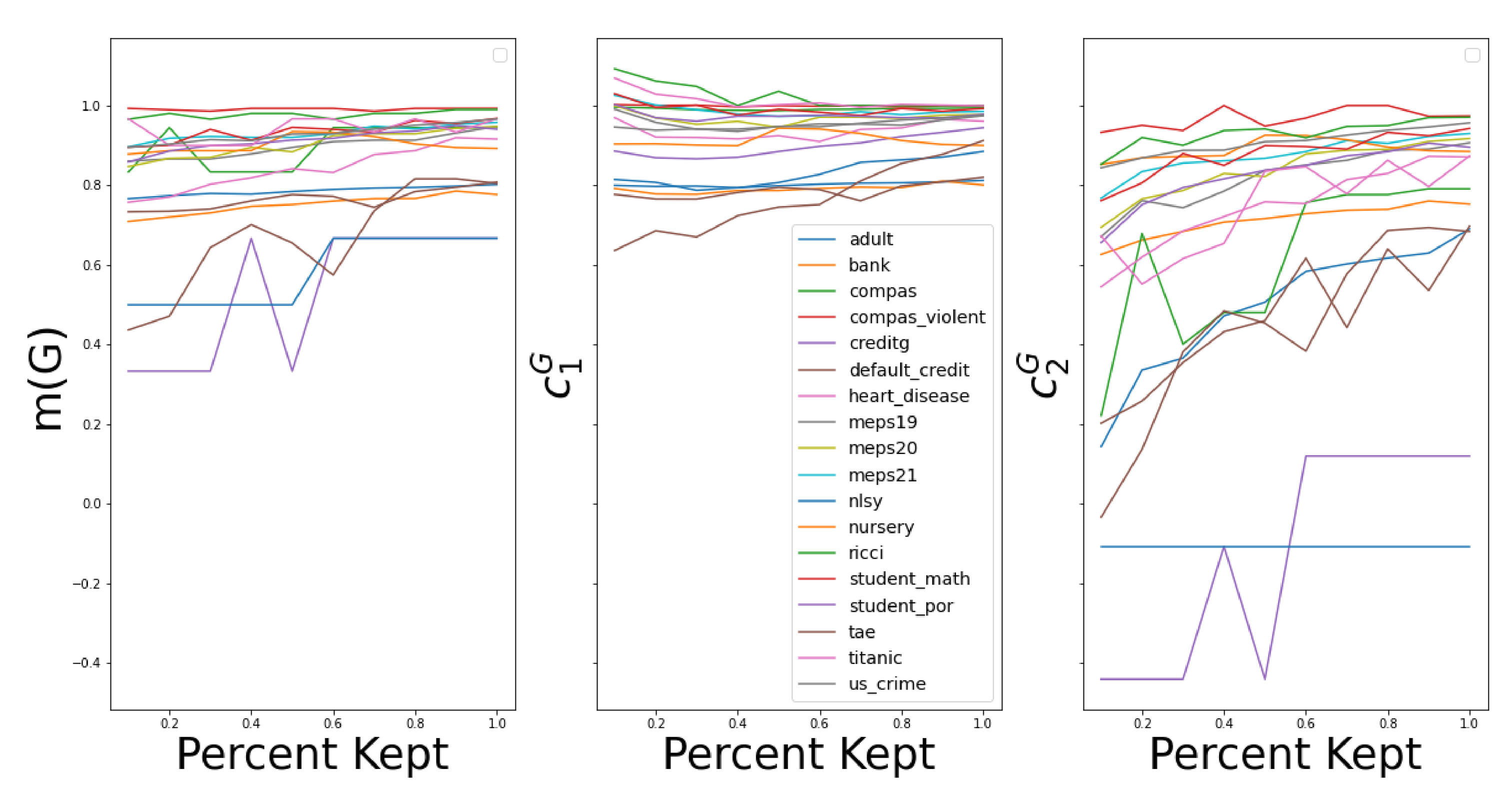
- Test whether proposed metrics meet Incentive Compatibility
- Method
  - **18 fairness datasets** from IBM's lale library
  - XGBoost models with 3-fold cross-validation using lale
  - **Identify critical subgroups**: Minimum accuracy, minimum c_1, minimum c_2
  - **Subsampling** experiments on critical subgroups and full datasets

Result: Both metrics satisfy Incentive Compatibility

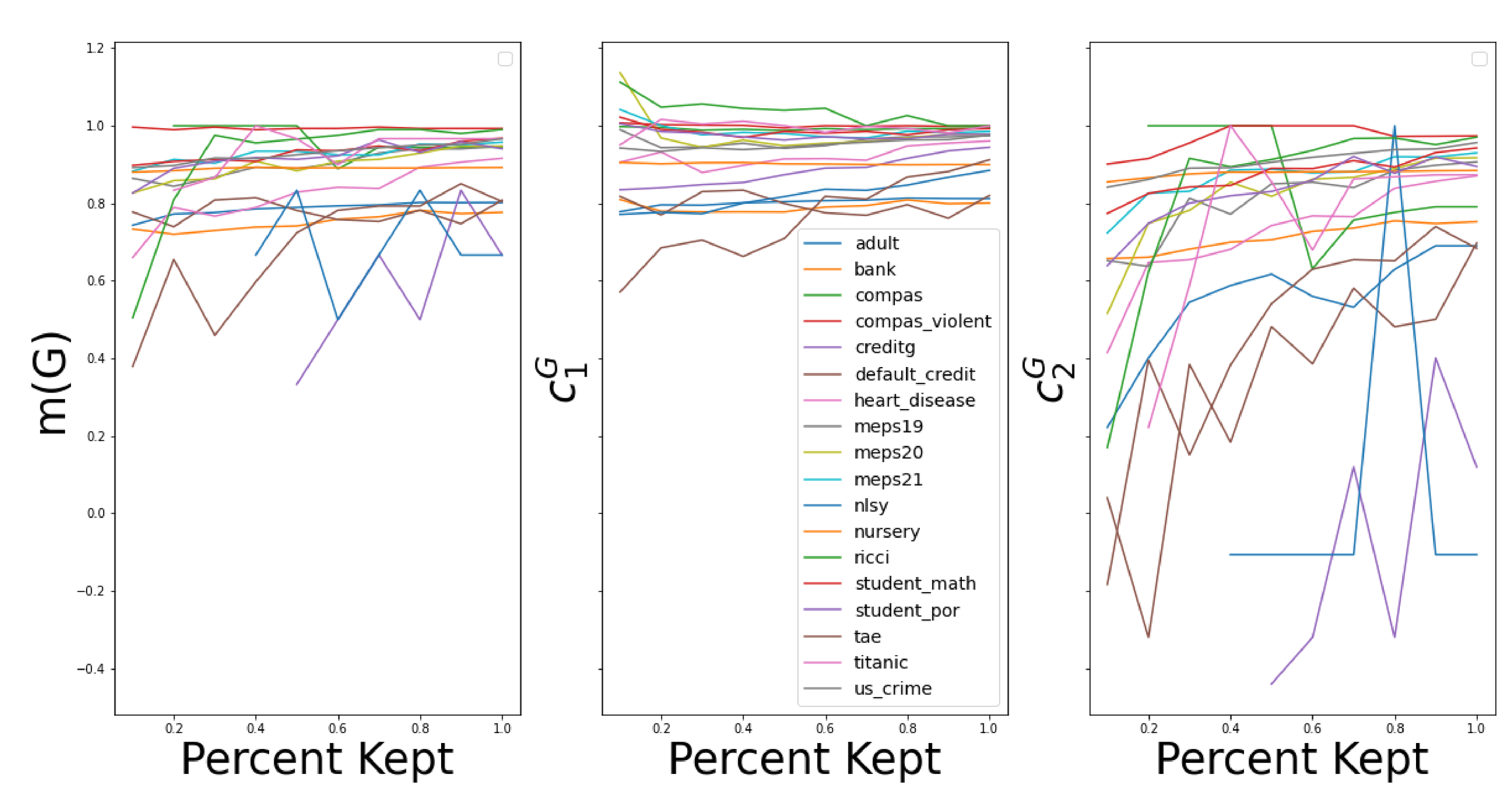
Both the Optimist's Metric and Pessimist's Metric increase as data increase, indicating they satisfy Incentive Compatibility.

$m(G)$ = accuracy, group G $\quad c_1^G$ = Optimist's $\quad c_2^G$ = Pessimist's metric

Subsampling only the Critical Subgroup



Subsampling the Entire Dataset



Summary

- Describe intersectionality problem for fairness estimation
- Develop desiderata to guide search for fairness metrics
- Illustrate desiderata with metrics based on hypothesis testing
- **Explore fundamentally different approach**: fairness as *sufficiency* (not equality), accounting for *uncertainty* (not point estimates)
- *Does existing literature sufficiently consider statistical uncertainty in estimating fairness?*

NEURAL INFORMATION PROCESSING SYSTEMS