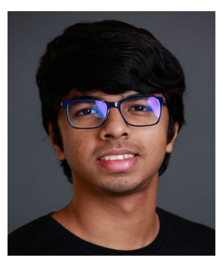
On Counterfactual Inference with Unobserved Confounding


Abhin Shah
abhin@mit.edu
MIT

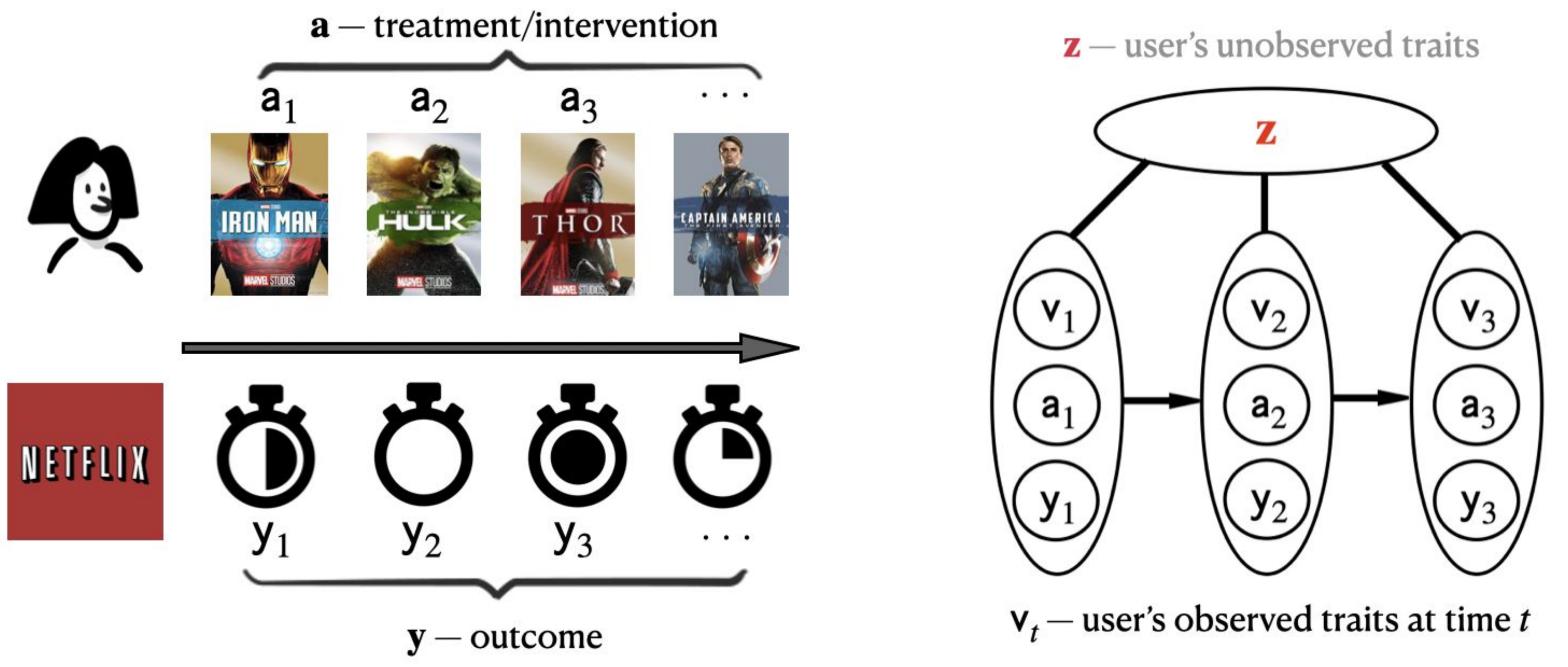

Raaz Dwivedi
raaz@mit.edu
MIT, Harvard University


Devavrat Shah
devavrat@mit.edu
MIT

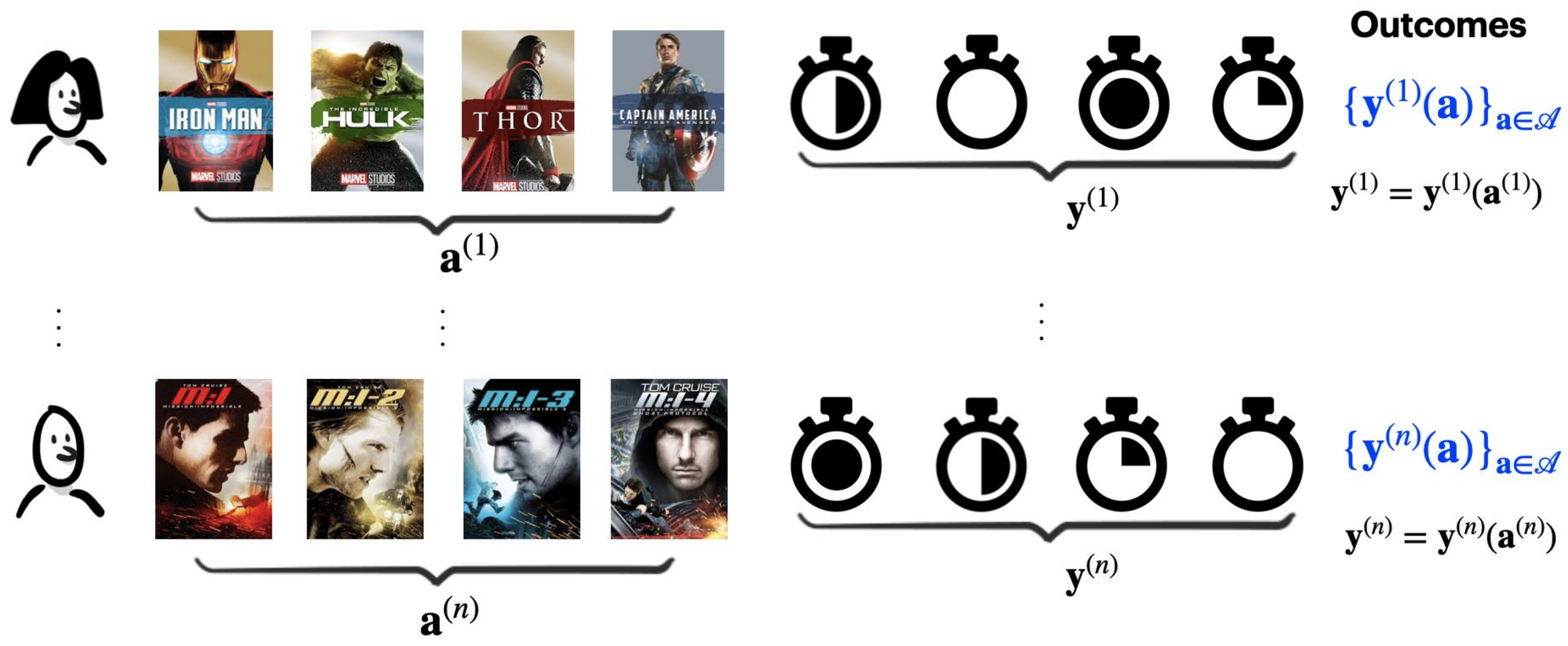

Greg Wornell
gww@mit.edu
MIT

arXiv Link

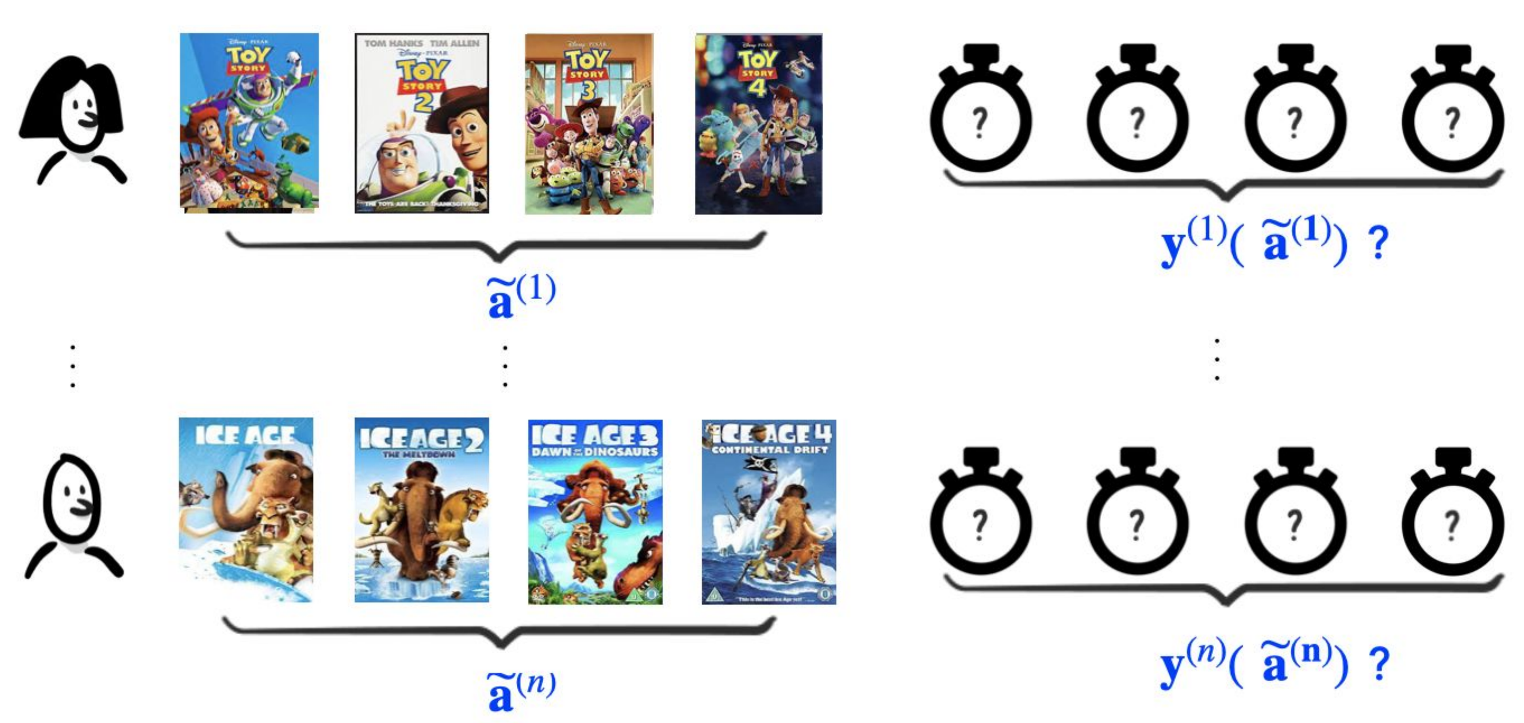
Observational Setting



Panel Data

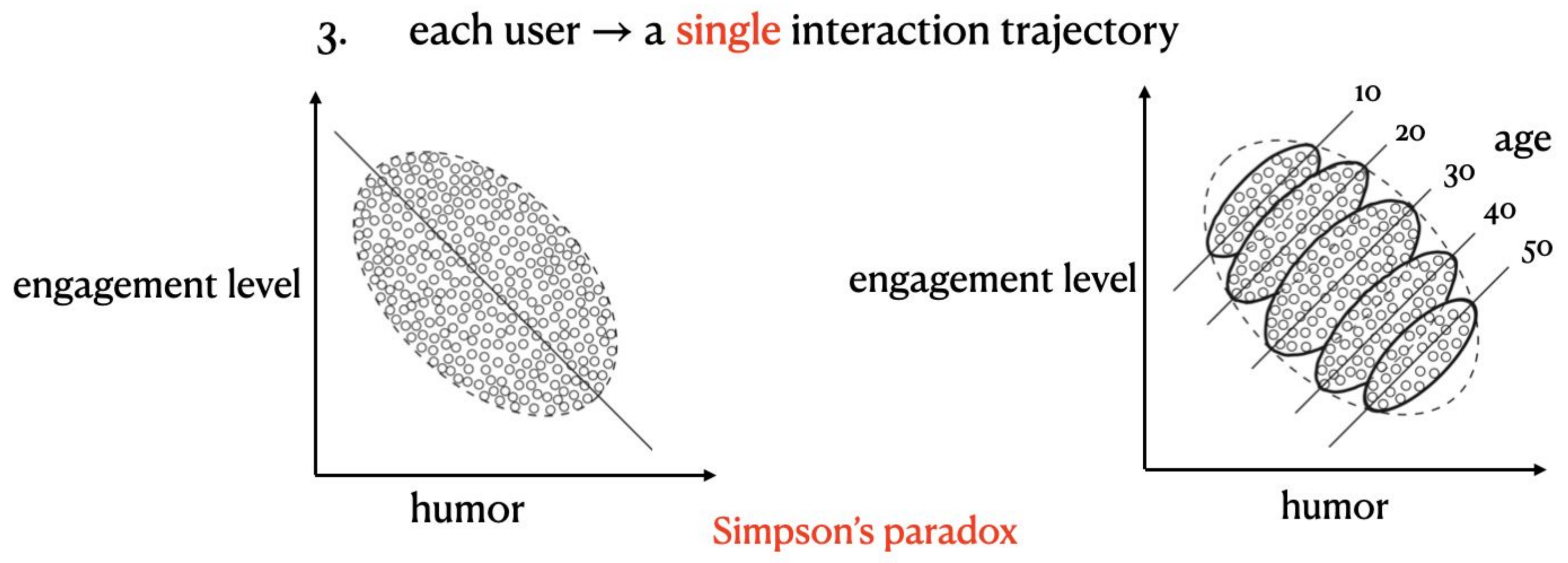


Goal : What-if?



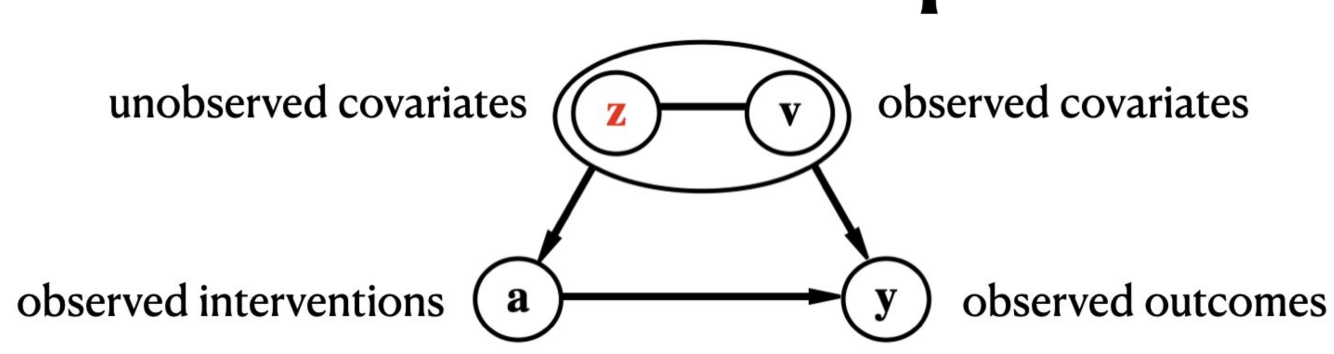
Challenges

1. unobserved factors → spurious associations
2. users → heterogeneous
3. each user → a single interaction trajectory



Simpson's paradox

Problem Setup



n heterogenous and independent users with one observation each - $\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$ (p–dimensional)

Goal : Counterfactual Questions

For user $i \in [n]$, what would have happened if alternative treatments were assigned?

≡

Estimate $y^{(i)}(\tilde{a}^{(i)})$ for $\tilde{a}^{(i)} \in \mathcal{A}$?

Suffices to learn $f(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, z^{(i)}, v^{(i)})$ for all $i \in [n]$, but each user may have *different* z

Can we learn n different distributions with *one* sample per distribution ?

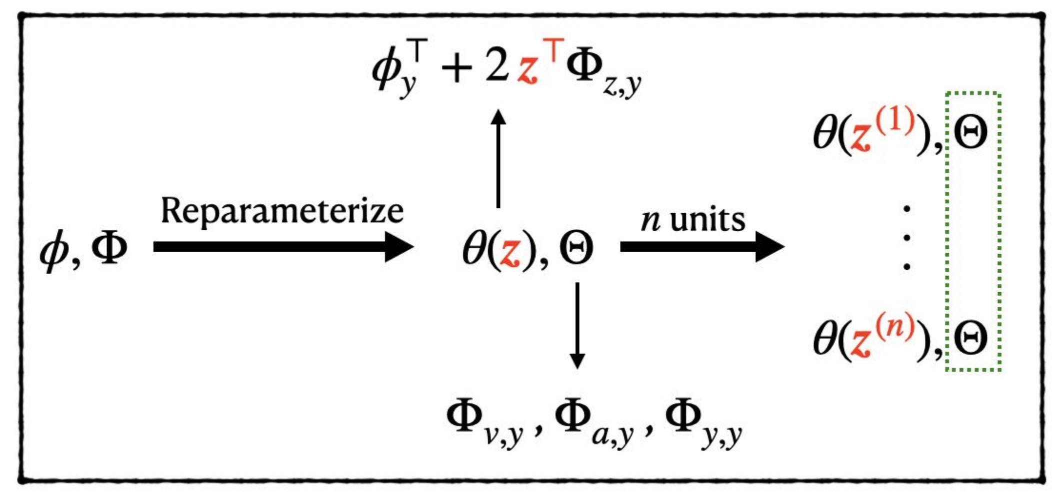
Our Approach

We posit a joint exponential family distribution for $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

$$f(\boldsymbol{w}) \propto \exp\left(\boldsymbol{\phi}^\top \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Phi} \boldsymbol{w}\right)$$

$$f(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\underbrace{\phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y}}_{\text{different for different users}} + 2\boldsymbol{a}^\top\Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y}\right)$$

n heterogeneous conditional distributions ⇒ same exp. family but with diff. parameters



Inference Tasks

1. Parameters: User-level — $\theta^\star(z^{(i)})$ for all $i \in [n]$; Population-level — $\Theta^\star$ → counterfactual distribution

2. Potential Outcomes: $\mu^{(i)} \triangleq \mathbf{E}\left[y^{(i)}(\tilde{a}^{(i)}) \mid \mathbf{z} = z^{(i)}, \mathbf{v} = v^{(i)}\right]$ → counterfactual mean

Parameter Estimation

$\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$ pool all n samples →

$$\min_{\theta^{(1)},\cdots,\theta^{(n)},\Theta} \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\left(-\left[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\right] x_t^{(i)}\right)$$

→ $\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(n)}, \widehat{\Theta}$ estimates

Assum 1: $\Theta^\star$ has sparse rows

Assum 2: $\theta^\star(z^{(i)}) \in$ set $\mathcal{B}$

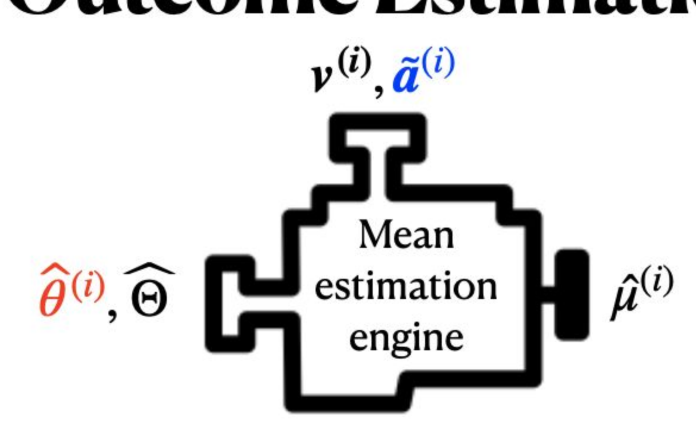
$$\|\Theta^\star - \widehat{\Theta}\|_{2,\infty} \le \epsilon \quad \text{when } n \ge O\left(\frac{p^2(p + M_n(\epsilon^2))}{\epsilon^4}\right)$$

$$\text{For all } i, \text{MSE}\left(\theta^\star(z^{(i)}), \hat{\theta}^{(i)}\right) \le \max\left\{\epsilon^2, \frac{M(c)}{p}\right\} \quad \text{when } n \ge O\left(\frac{p^2(pM(c) + M_n(\epsilon^2))}{\epsilon^4}\right)$$

metric entropy of $\mathcal{B}$ — $M_n(\epsilon) = nM(n\epsilon)$

★ When $\mathcal{B} = s-$sparse linear combinations of k known vectors,
$M(c) = O(s\log(k))$ and $M_n(\epsilon) = O\left(\frac{s\log k}{\epsilon}\right)$

Outcome Estimation



$$\hat{f}(\boldsymbol{y} \mid \boldsymbol{a} = \tilde{\boldsymbol{a}}^{(i)}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\hat{\theta}(z^{(i)}) + 2\boldsymbol{v}^{(i)\top}\widehat{\Phi}_{v,y} + 2\tilde{\boldsymbol{a}}^{(i)\top}\widehat{\Phi}_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top\widehat{\Phi}_{y,y}\boldsymbol{y}\right)$$

For all i and any $\tilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}$,

$$MSE\left(\mu^{(i)}, \hat{\mu}^{(i)}\right) \le \epsilon^2 + \frac{M(c)}{p} \quad \text{when } n \ge O\left(\frac{p^2(pM(c) + M_n(\epsilon^2))}{\epsilon^4}\right)$$

Application: Denoise User-wise Data

No systematically unobserved covariates

Noisy observed data = true data + measurement error
($\bar{\mathbf{x}}$, $\mathbf{x}$, $\Delta\mathbf{x}$)

Assum 1: Only half users have error: $\Delta\mathbf{x}^{(i)} = \mathbf{0}$ for $i \in \{n/2, \cdots, n\}$

Assum 2: Data has sparse error: $\|\Delta\mathbf{x}^{(i)}\|_0 \le s$ for $i \in \{1, \cdots, n/2\}$

Goal: Estimate the true data

$$\text{For all } i,\ \|\Delta\mathbf{x}^{(i)}, \widehat{\Delta\mathbf{x}^{(i)}}\|^2 \le \max\left\{\frac{\epsilon^2}{s}, \frac{s}{p}\right\} + \epsilon^2 \quad \text{when } n \ge O\left(\frac{s^2 p}{\epsilon^4}\right)$$