Exploiting Additional Measurements in Staggered Rollout Designs for Graph Agnostic Estimators under Network Interference

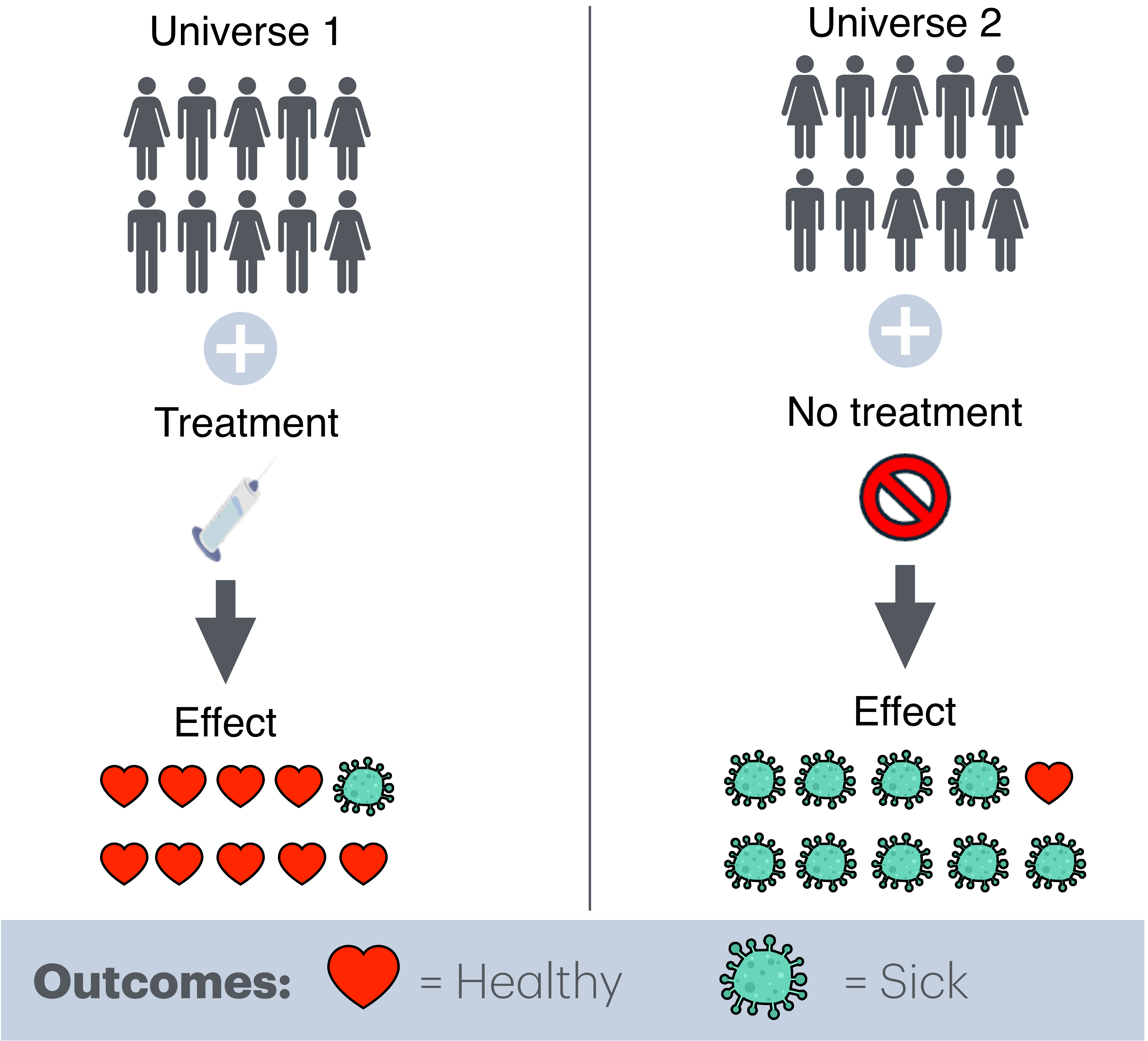
Emily Lopez
University of Minnesota - Twin Cities

Joint work with Mayleen Cortez-Rodriguez, Matt Eichhorn, Christina Lee Yu, Jennifer Kim, and Leo Phillips
Cornell University

AMS Special Session on Research Results by Mathematicians from the Enhancing Diversity in Graduate Education (EDGE) Program

Identifying Causation using the Total Treatment Effect

Motivation: Determine the efficacy of vaccine



Total Treatment Effect (TTE)

Average difference in outcomes

$$\mathrm{TTE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{1})}_{\text{Avg. effect when everyone is treated}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{0})}_{\text{Avg. effect when no one is treated}}$$

Identifying Causation using the Total Treatment Effect

Motivation: Determine the effectiveness of an advertisement

Universe 1

Advertisement

Effect

$ $ $ $ ✗
$ $ $ $ $

Universe 2

No Advertisement

Effect

✗ ✗ ✗ ✗ $
✗ ✗ ✗ ✗ ✗

Outcomes: $ = Buyer ✗ = Not a Buyer

Total Treatment Effect (TTE)

Average difference in outcomes

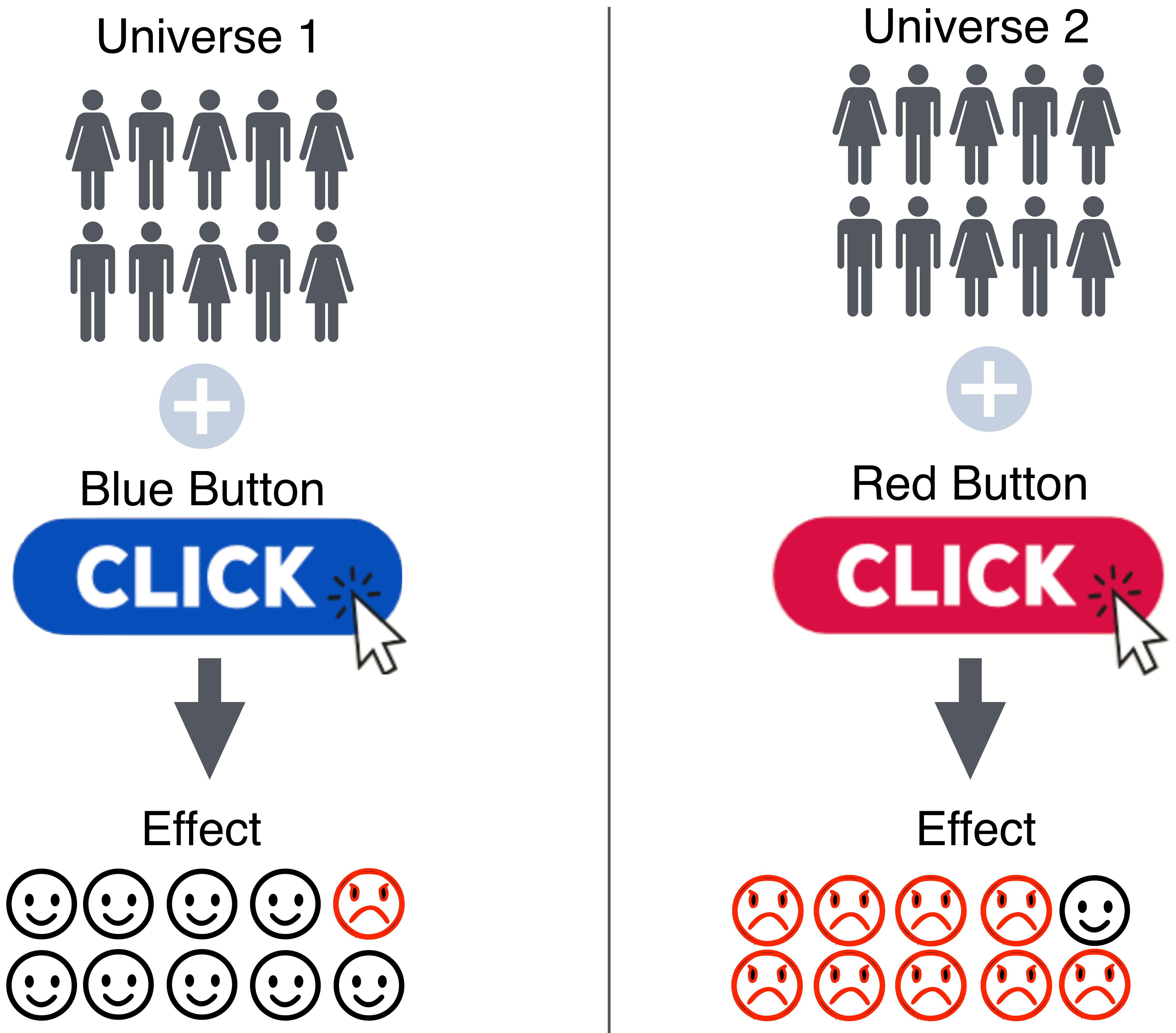
$$\text{TTE} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{0})$$

Avg. effect when everyone is treated

Avg. effect when no one is treated

Identifying Causation using the Total Treatment Effect

Motivation: Determine the effect of a platform change on user experience



Outcomes: = Like = Dislike

Total Treatment Effect (TTE)

Average difference in outcomes

$$\mathrm{TTE} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i\,(\mathbf{0})$$

Avg. effect when everyone is treated

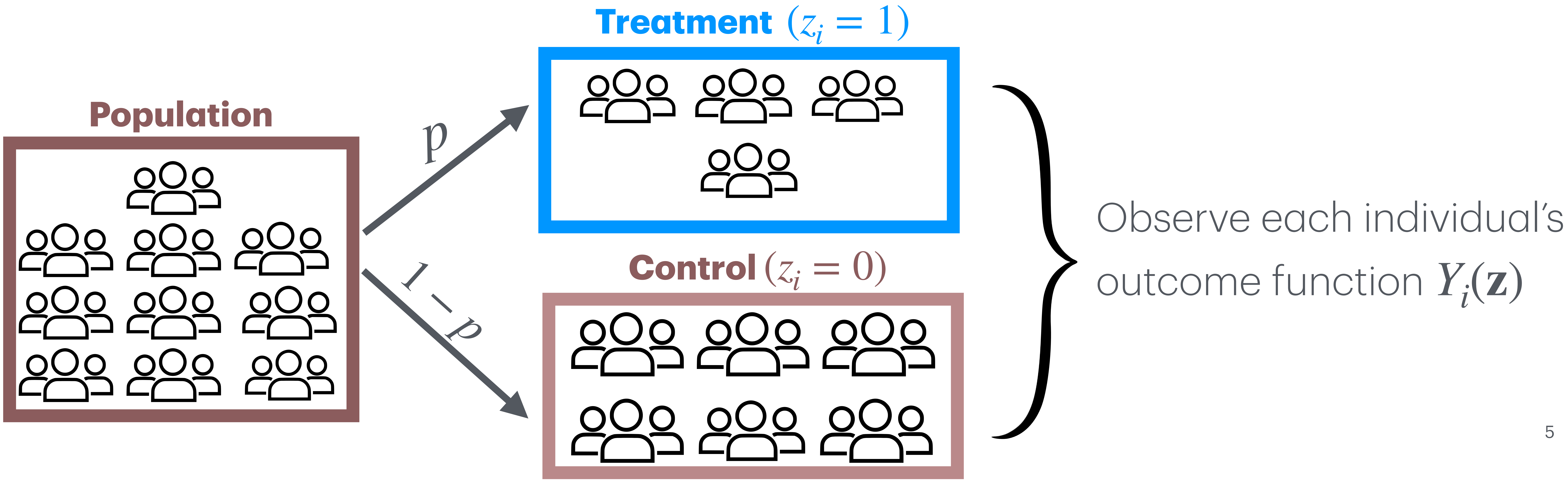
Avg. effect when no one is treated

Small Problem: We Cannot Observe the TTE Directly

- We are often limited to a **single experiment** and a **small portion of the population we can treat** $p \in [0,1]$.

- We estimate the TTE from a randomized experiment

- We assign treatment to each individual independently using $z_i \sim \text{Bernoulli}(p)$



How Should we Estimate the TTE?

Total Treatment Effect (TTE)

Average difference in outcomes

$$\mathrm{TTE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{0})$$

Avg. effect when everyone is treated

Avg. effect when no one is treated

- measure the "goodness" of an estimator using the **mean squared error** (MSE)

$$\mathrm{MSE}(\widehat{\mathrm{TTE}}) = \mathrm{bias}^2(\widehat{\mathrm{TTE}}) + \mathrm{Variance}(\widehat{\mathrm{TTE}})$$

- Want bias and variance to be small
- (bias-variance trade-off)

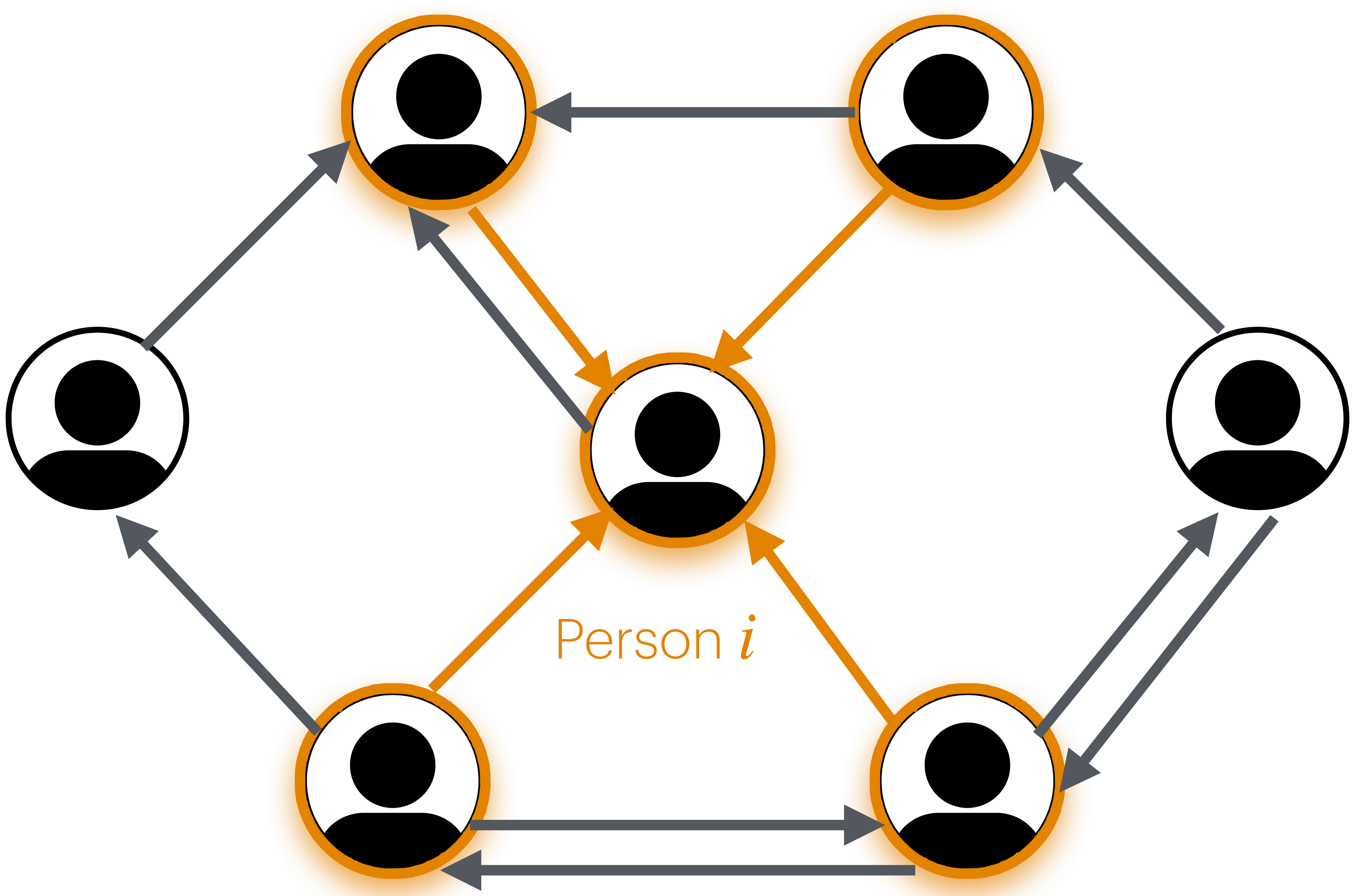
Classical estimator: difference in means (DM)

$$\widehat{\mathrm{TTE}}_{\mathrm{DM}} = \frac{\sum_{i=1}^{n} z_i Y_i(\mathbf{z})}{\sum_{i=1}^{n} z_i} - \frac{\sum_{i=1}^{n} (1-z_i) Y_i(\mathbf{z})}{\sum_{i=1}^{n} (1-z_i)}$$

Avg. effect among those in treatment

Avg. effect among those in control

DM Estimator Fails to Account for Interference



$\mathcal{N}_i$ = **neighborhood of** i

- **Interference** is when the outcome of an individual can be affected by the treatment of another

  ✳ Ex: Herd immunity (immunology)

  ✳ Ex: Hearsay influences buyer's outcomes and others' opinions

- **Neighborhood Interference** assumes a person is only affected by the treatment of its **direct** neighbors.

- **Directed edge** (j, i): Person j is a direct neighbor to person i.

Cost-Benefit of Network Information in Causal Models

Can get more expressive potential outcome models to account for richer interactions

Complete network knowledge may not be known by the practitioner

The complexity of the network increases the “hardness” of inference.

Research Question

What is the best way to estimate the TTE when we do not know the interference network structure?

Structure of Potential Outcomes Function

- Under network interference, the potential outcomes model takes the form

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i, |\mathcal{S}| \leq \beta} c_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j .$$

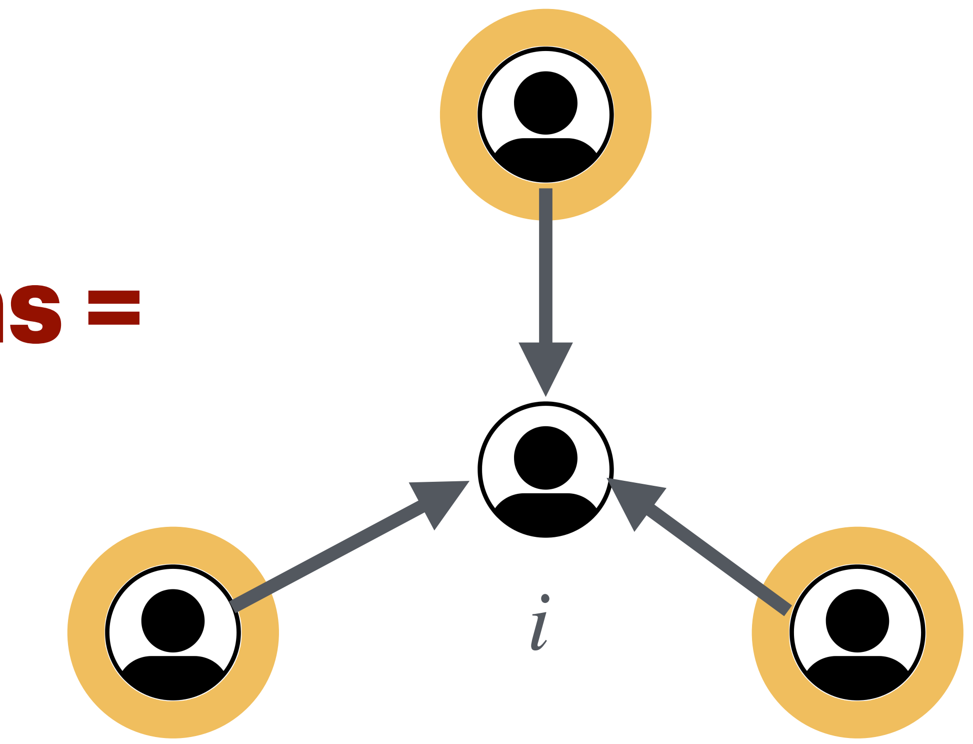
$c_{i,\mathcal{S}}$ = the influence of treating the subset of individuals $\mathcal{S}$ on person i's outcome.

$\prod_{j \in \mathcal{S}} z_j$ = indicator that checks all individuals in subset $\mathcal{S}$ were assigned treatment.

$Y_i(\mathbf{z})$ = sum of influence of all possible treated subsets of the neighborhood of person i.

β = order of interactions—serves as proxy for complexity of the model

$|\mathcal{S}| \leq 1$ interactions =



Structure of Potential Outcomes Function

- Under network interference, the potential outcomes model takes the form

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i, |\mathcal{S}| \leq \beta} c_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j .$$

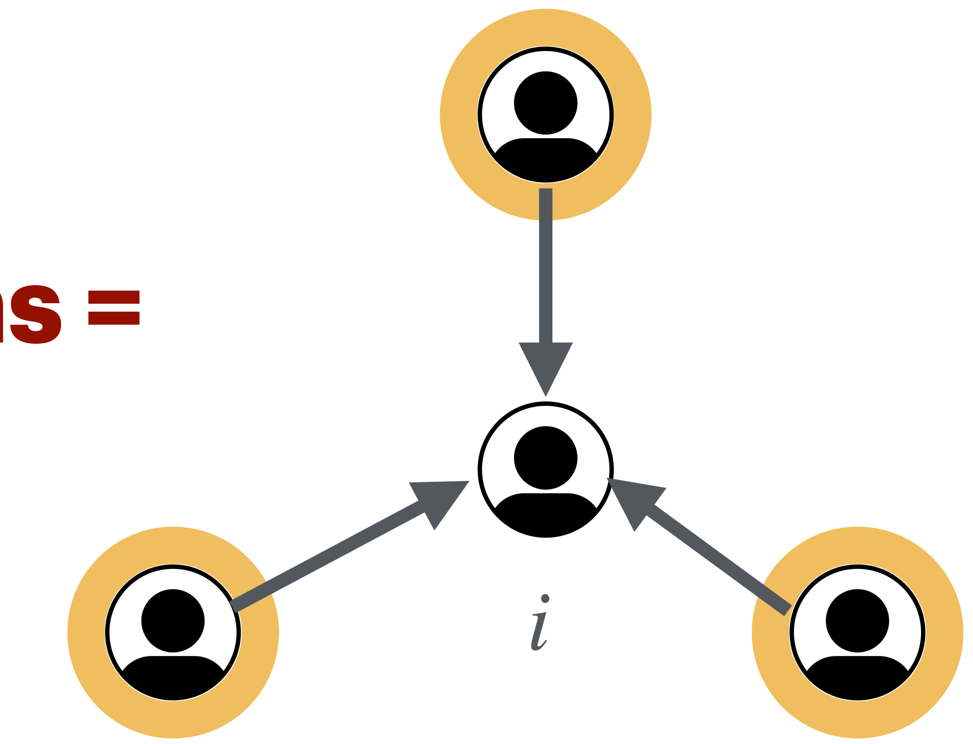
$c_{i,\mathcal{S}}$ = the influence of treating the subset of individuals $\mathcal{S}$ on person i's outcome.

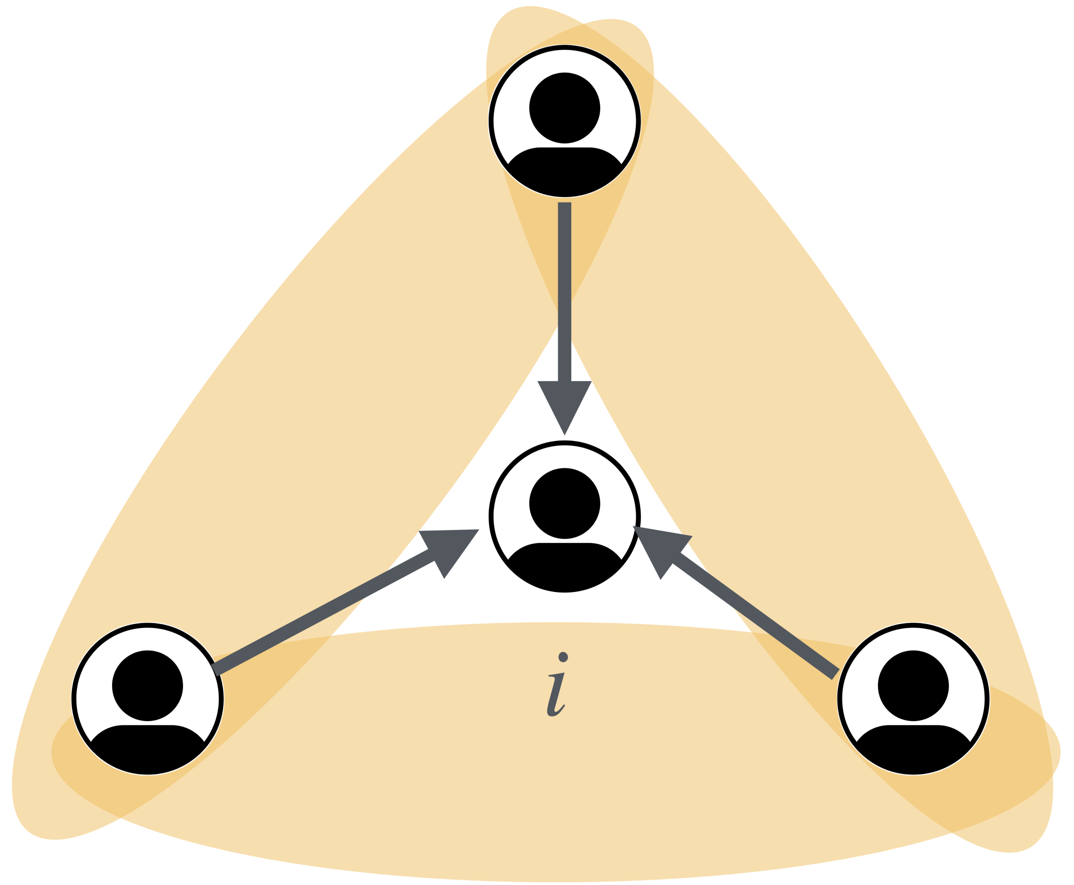
$\prod_{j \in \mathcal{S}} z_j$ = indicator that checks all individuals in subset $\mathcal{S}$ were assigned treatment.

$Y_i(\mathbf{z})$ = sum of influence of all possible treated subsets of the neighborhood of person i.

β = order of interactions—serves as proxy for complexity of the model

$|\mathcal{S}| \leq 2$ **interactions =**

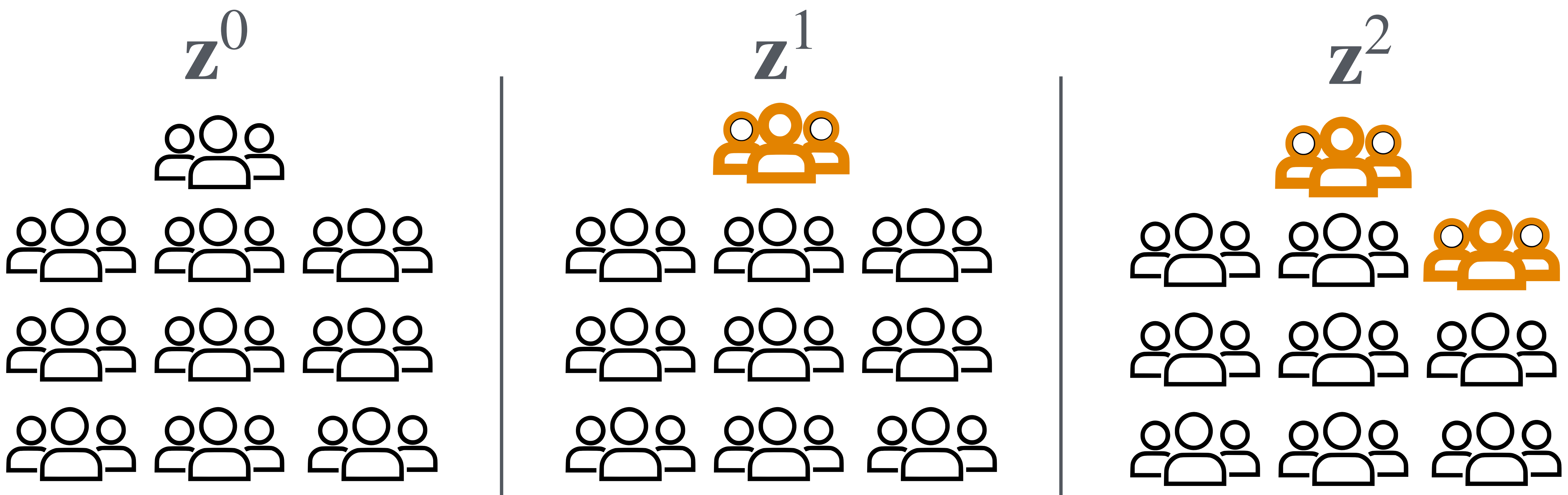




Polynomial Interpolation Estimator

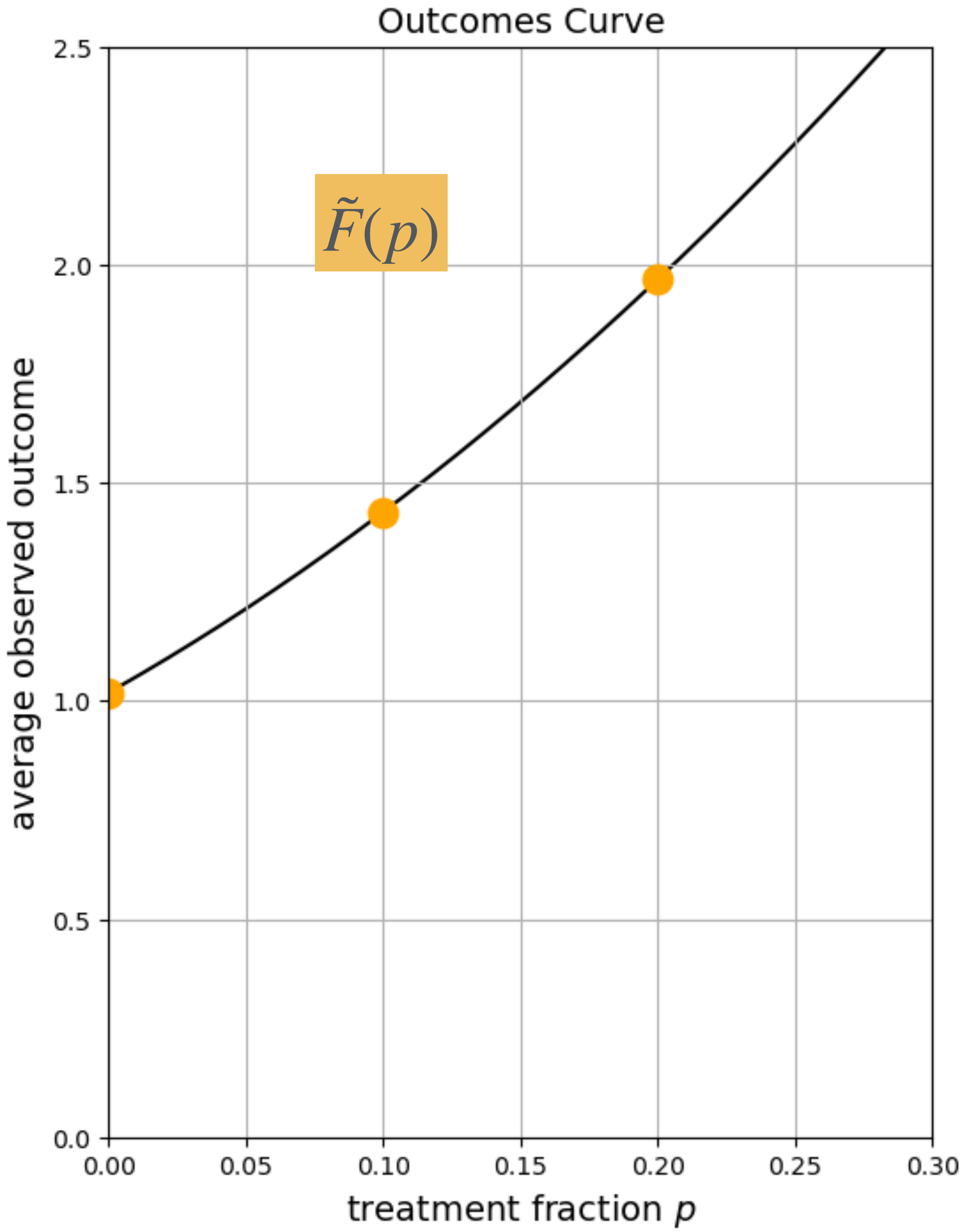
- If we can observe the outcomes of treating each fraction of the population $p \in [0,1]$, we can **fit a polynomial** to guess what would happen at higher fractions.
  - The degree of the polynomial = order of interactions β
- Staggered Roll out enables polynomial interpolation

Ex: Order of interactions $\beta = 2$, $p = 0.2$



Stage 0: 0% treated | **Stage 1:** 10% treated | **Stage 2:** 20% treated

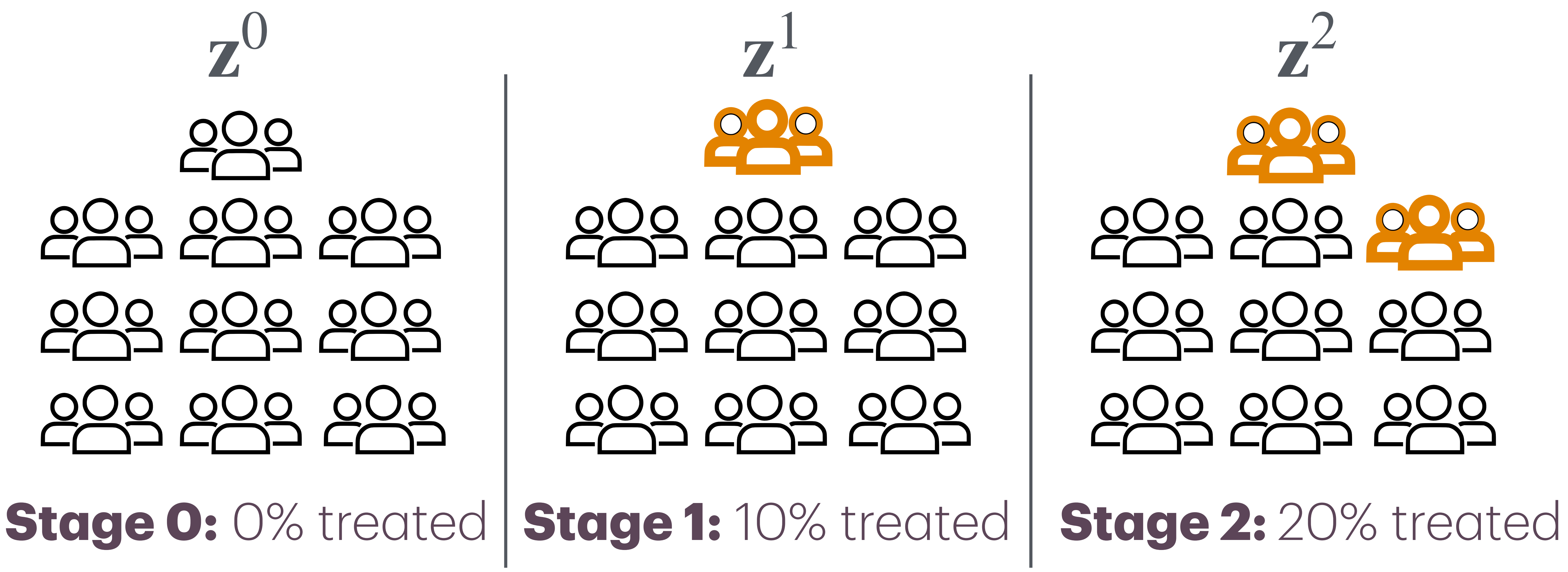
$$\widehat{\mathrm{TTE}}_{\mathrm{PI}} = \tilde{F}(1) - \tilde{F}(0)$$



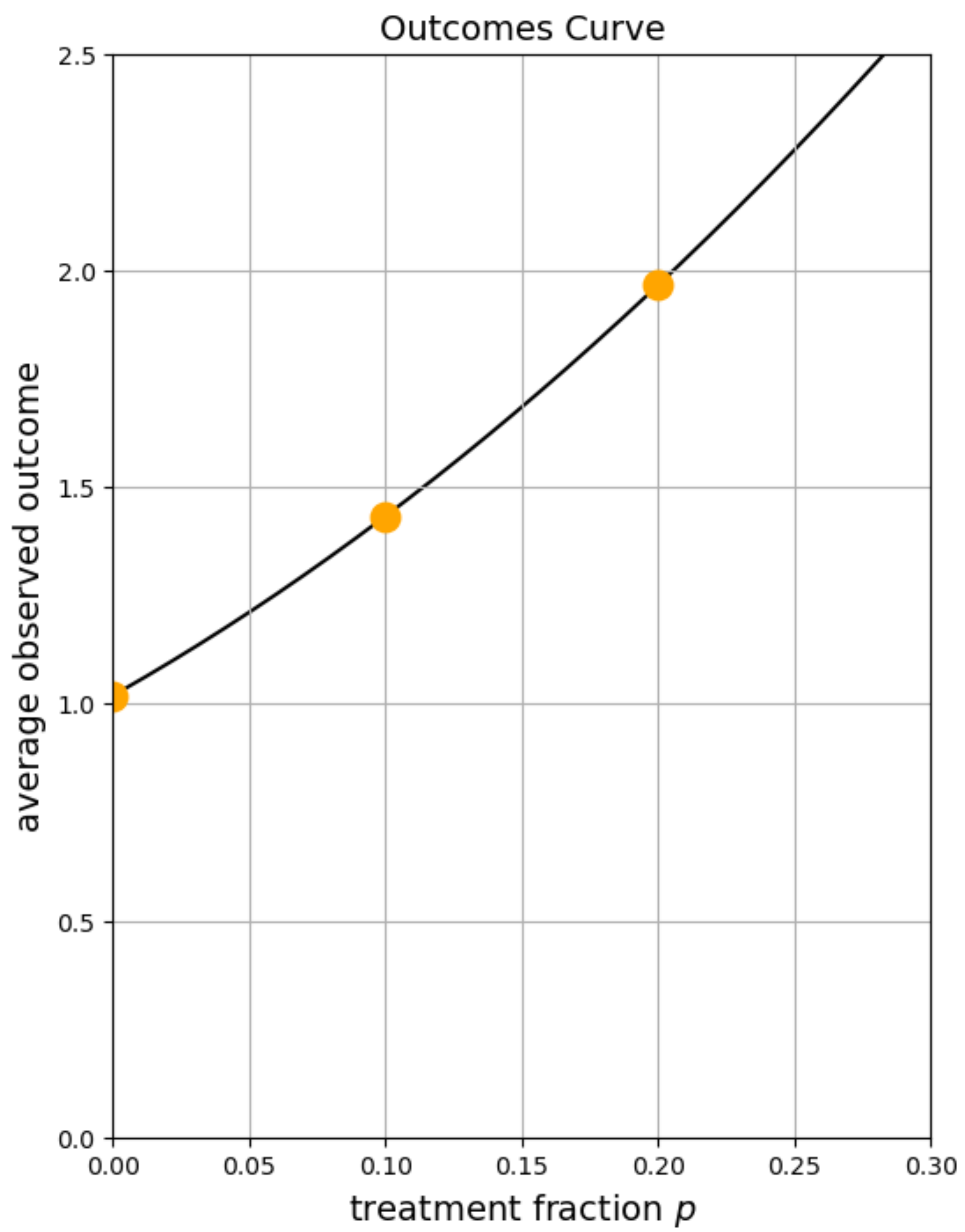
[Cortez et.al 2022]

Performance of the Polynomial Interpolation Estimator

Ex: Order of interactions $\beta = 2$, $p = 0.2$



$$\widehat{\mathrm{TTE}}_{\mathrm{PI}} = \tilde{F}(1) - \tilde{F}(0)$$



Pros

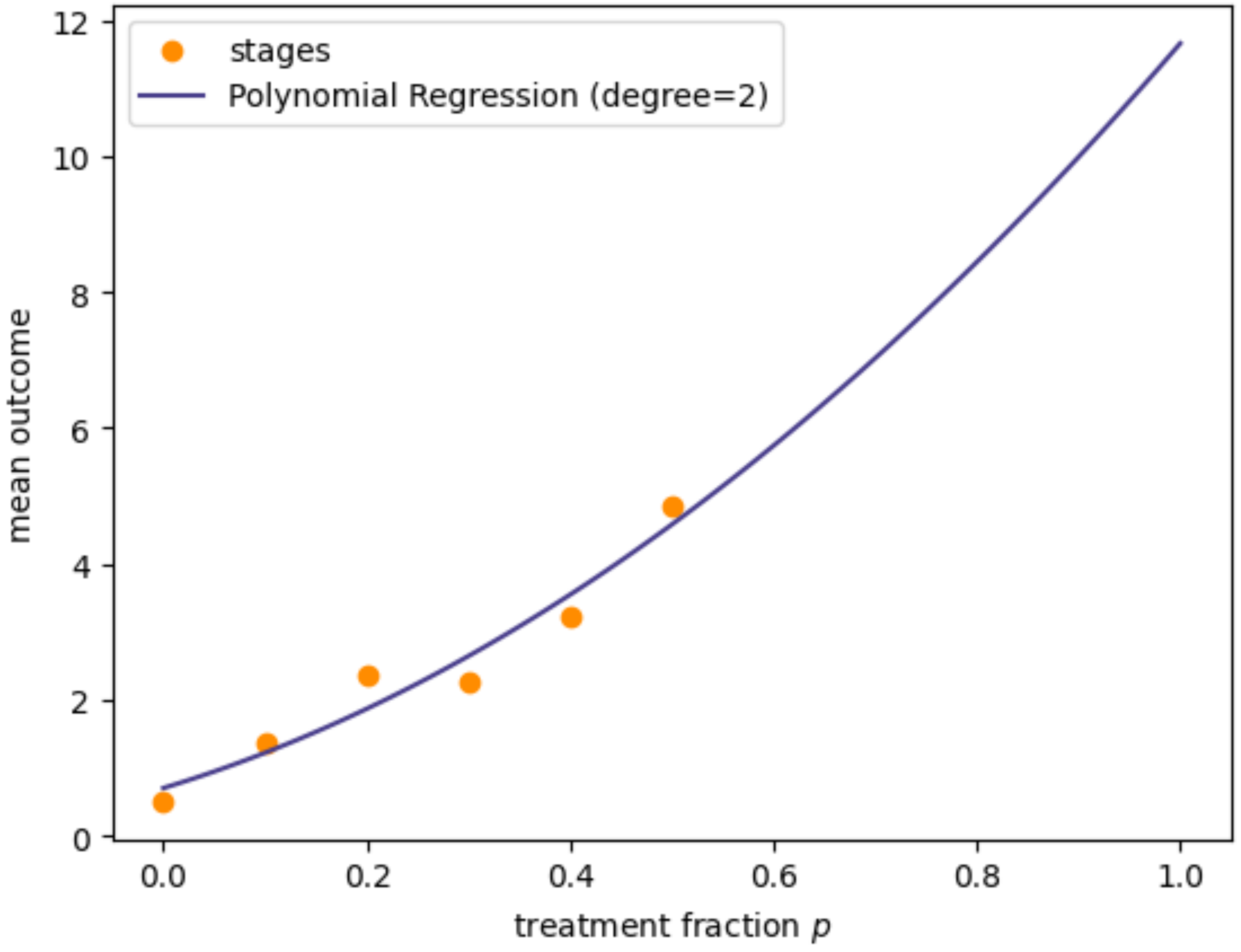
- Unbiased estimator
- Graph agnostic

Cons

- Variance scales $(\beta/p)^{2\beta}$

[Cortez et.al 2022]

Can We Lower the Variance by Extending our Rollout Length?



- By sampling at more than $\beta + 1$ points, we can instead use **polynomial regression (OLS)** rather than interpolation
  - ✳ Can regression decrease the variance?

Theorem 1 [Cortez et.al 2022] If the potential outcomes function is linear, and a staggered rollout is implemented with a set of distinct treatment probabilities, the **best-unbiased estimator** for the TTE when we can obtain **perfect observations** is

$$\widehat{\text{TTE}}_{\text{end-pt}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}^T) - Y_i(\mathbf{z}^0)}{p}.$$

Main Point: The only informative points in the staggered roll-out design with perfect observations are the first and last observations.

How Much Noise Needs to be Present for OLS to Improve Variance?

Suppose $Y_i^{\text{obs}}(\mathbf{z}^t) = Y_i(\mathbf{z}^t) + \epsilon_i$ where the noise $\epsilon_i \sim N(0, \sigma^2)$, and $\beta = 1$

Theorem 2 [LCEY 2024+] When the linear potential outcomes function is affected by Gaussian noise $\epsilon_i \sim N(0, \sigma^2)$ and we ramp up the rollout an equal amount at each stage, $\text{MSE}(\widehat{TTE}_{\text{OLS}}) \leq \text{MSE}(\widehat{\text{TTE}}_{\text{end-pt}})$ if and only if

$$\sigma^2 \geq \underbrace{\frac{p}{10n}\sum_{j=1}^{n} L_j^2}_{\text{noise threshold}} ,$$

where L_j denotes the total influence of individual j.

Main point: The "noise threshold" is independent of the number of stages T in the rollout.

Numerical Experiments ($\beta = 1$)

- Consider the potential outcomes model

$$Y_i(\mathbf{z}) = \sum_{j \in \mathcal{N}_i \setminus i} c_{ij} z_j \quad \text{where} \quad c_{ij} = \frac{1}{\text{out degree } j}$$

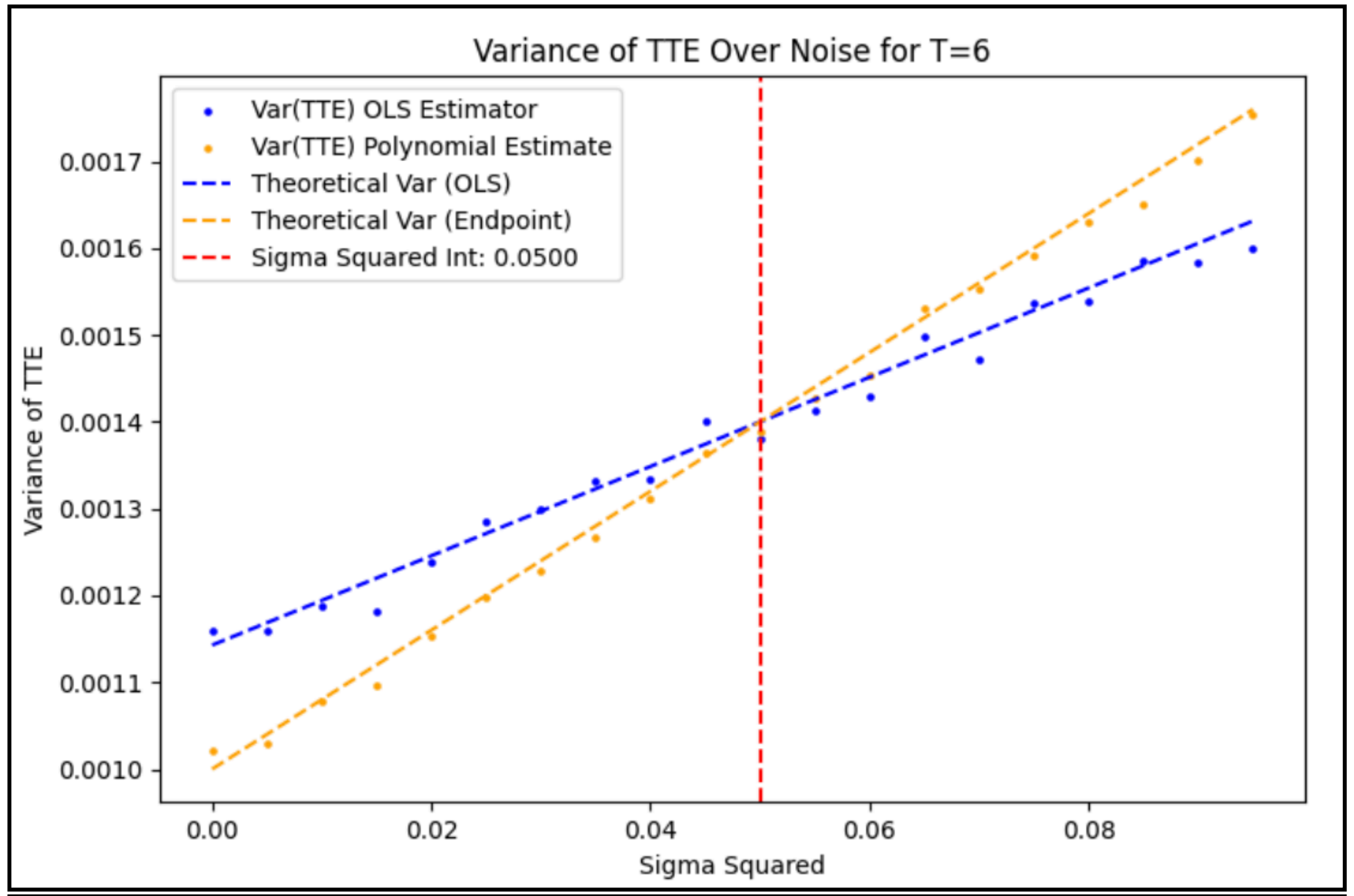
- When the treatment proportion $p = 0.5$, a graph theory arguments gives

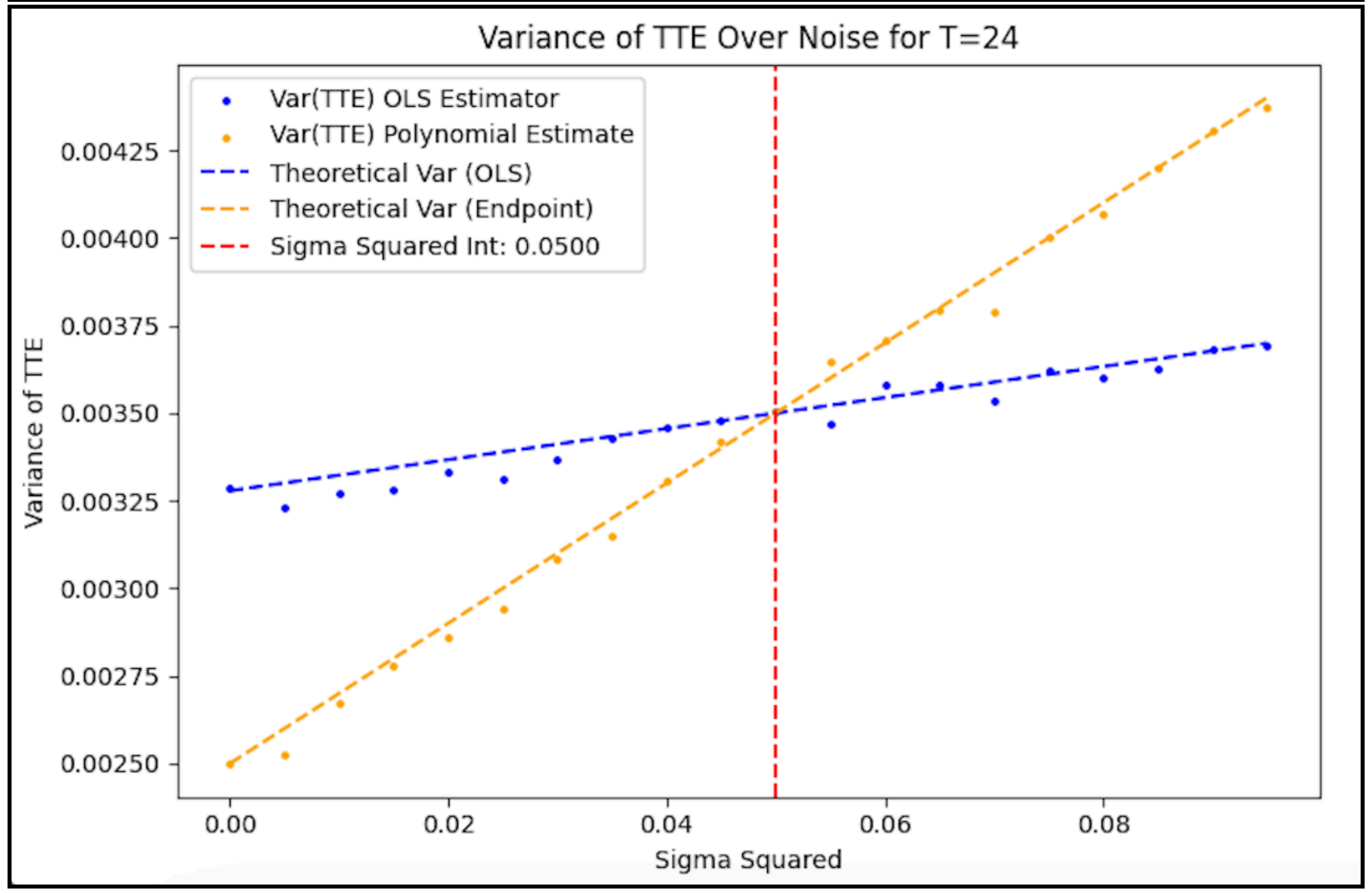
$$\frac{p}{10n} \sum_{j=1}^{n} L_j^2 = 0.05$$

- For a fixed number of stages, we

1. Generate 10 Erdos-Renyi Graphs (n = 400, p-edge = 0.1), and for each graph, simulate a staggered rollout design 1000 times for different noise levels.

2. Compute the empirical variance among the TTE estimates





How should we estimate the TTE when $\beta = 2$?

- In $\beta = 2$ case, the underlying curve $F(p)$ is a parabola
- It becomes less clear which observations of the roll-out procedure are the **most important** or if there is benefit to aggregate them

Candidate Estimators

Beginning-Middle-End Polynomial Interpolation Estimator $\widehat{\mathrm{TTE}}_{\mathrm{BME}} = \tilde{F}(1) - \tilde{F}(0)$

0	1	2	3	4	5	6	7	8

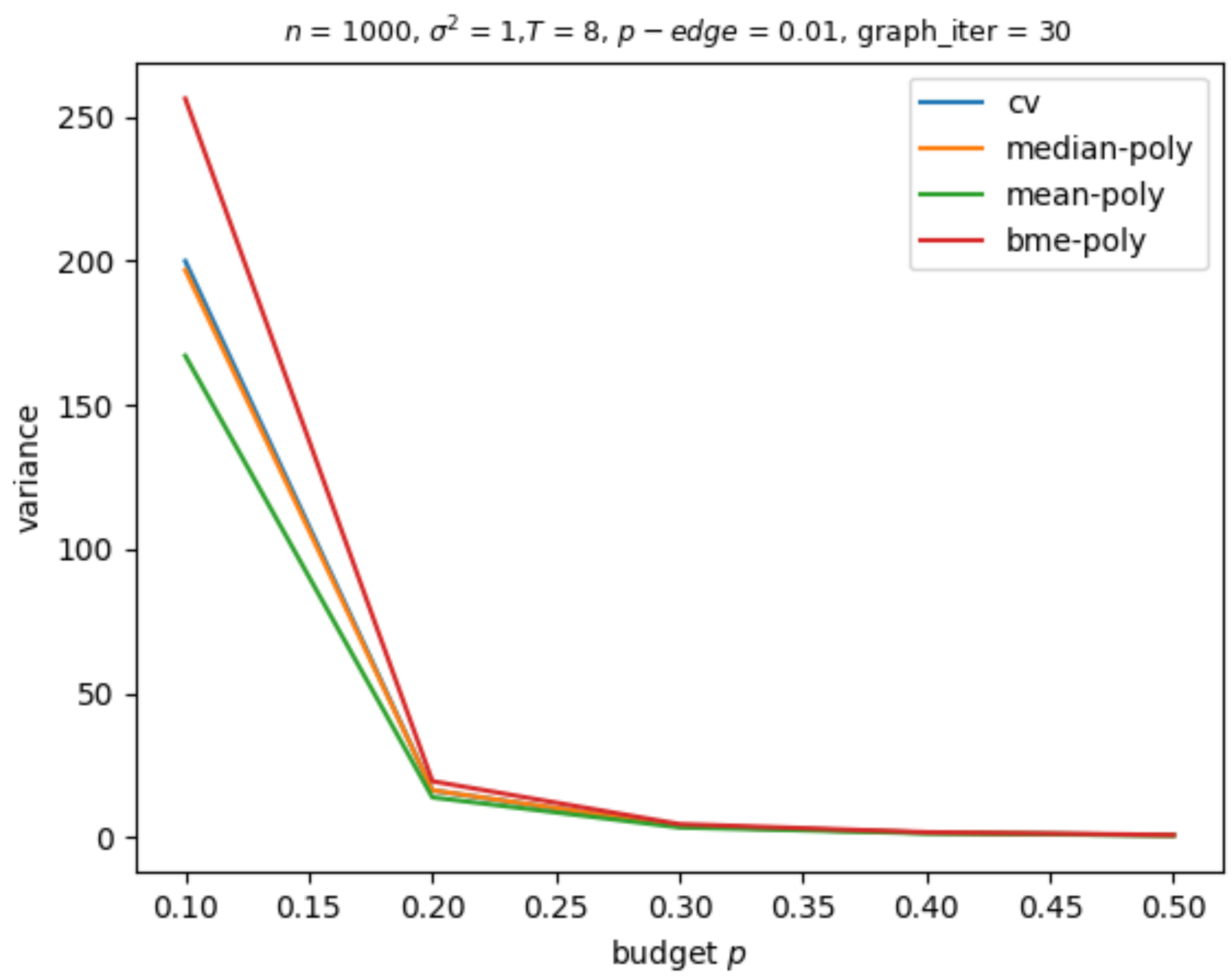
Cross-Validation Polynomial Interpolation Estimator: Choose 3 stages for interpolation that minimize square error among the other stages observations

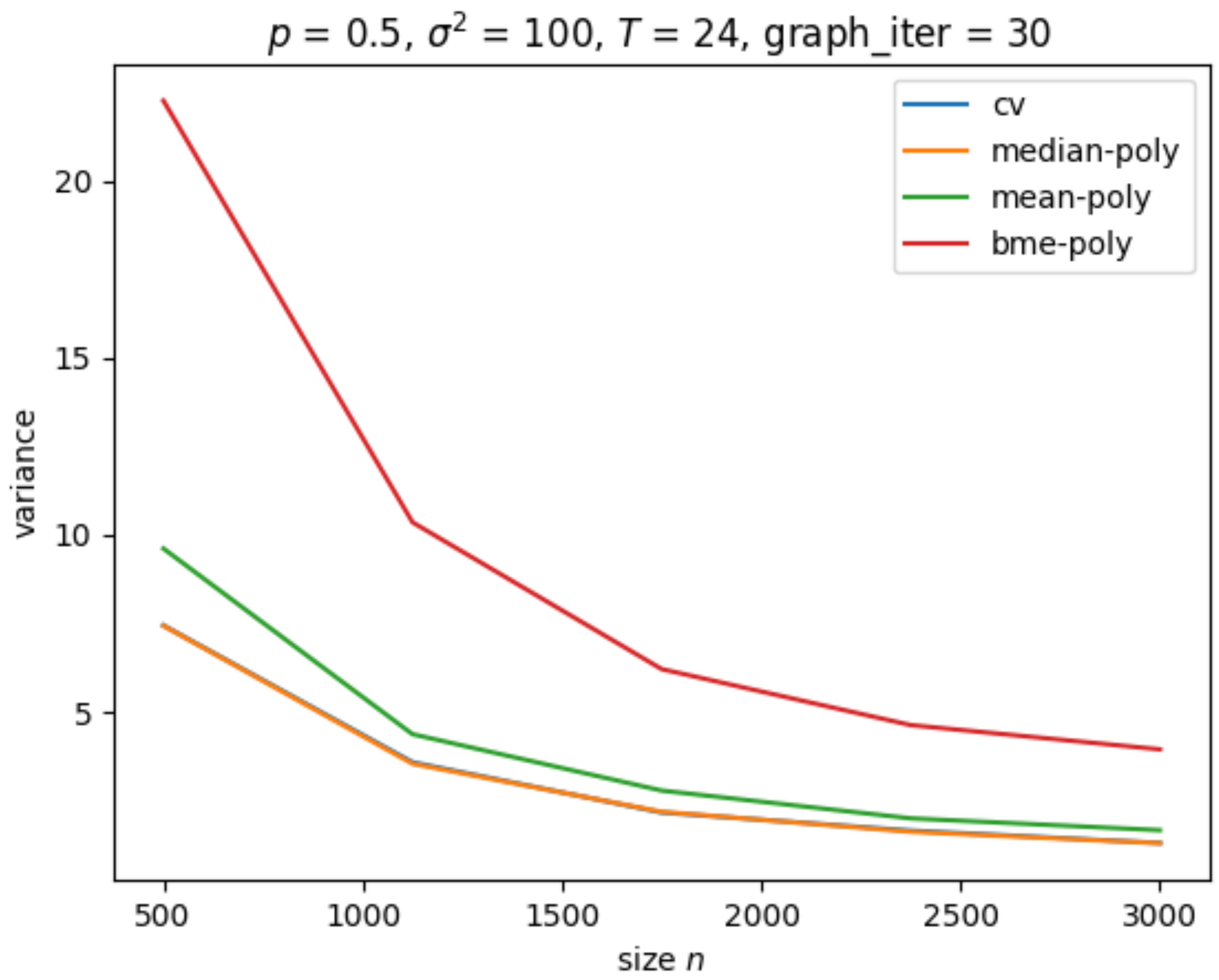
0	1	2	3	4	5	6	7	8

Aggregate Polynomial Interpolation estimates using Mean and Median

Preliminary Results

- Empirically, all estimators **unbiased**
- See major difference in their variances for **smaller treatment fractions**
- Median tends to perform best with **higher noise levels**





Key Takeaways and Future Directions

- **First-order interactions:**
  - OLS reduces variance only if there is a certain amount of noise in observations
- **Second-order interactions:**
  - Aggregation methods show the most difference in performance with lower treatment proportions
  - The optimal estimator may be highly dependent on the set of parameters
- **Next Steps:**
  - Apply methods to larger networks and study computational complexity

Additional Slides

Key Takeaways and Future Directions

- **First-order interactions:**
  - OLS reduces variance only if there is a certain amount of noise in observations
- **Next Steps:**
- Investigate second order interactions:
  - ✴ In $\beta = 2$ case, the underlying curve $F(p)$ is a parabola
  - ✴ It becomes less clear which observations of the roll-out procedure are the **most important** or if there is benefit to aggregate them

Models Restricted to Neighborhood Interference

Because treatments z_i are binary, any potential outcome function can be written as a **polynomial**

Effects of treating $\mathcal{S}$ on i

$$Y_i(\mathbf{z}) = \sum_{\boxed{\mathcal{S} \subseteq \mathcal{N}_i}} \boxed{a_{\mathcal{S}}} \prod_{j \in \mathcal{S}} z_j \prod_{j \in \mathcal{N}_{j'} \setminus \mathcal{S}} (1 - z_{j'}),$$

Subset of i's neighborhood

which can be written as a monomial basis

$$Y_i(z) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i} \tilde{a}_{\mathcal{S}} \prod_{i \in \mathcal{S}} z_j$$

The TTE then evaluates to be

$$\mathrm{TTE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i(\mathbf{1}) - Y_i(\mathbf{0}))$$

$$= \frac{1}{n} \sum_{i}^{n} \sum_{\mathcal{S} \subseteq \mathcal{N}_i} \tilde{a}_{\mathcal{S}}$$

Estimating the TTE reduces to get a good fit of the mean outcome (a polynomial)

Lower polynomial degree makes it **easier** to estimate the TTE.