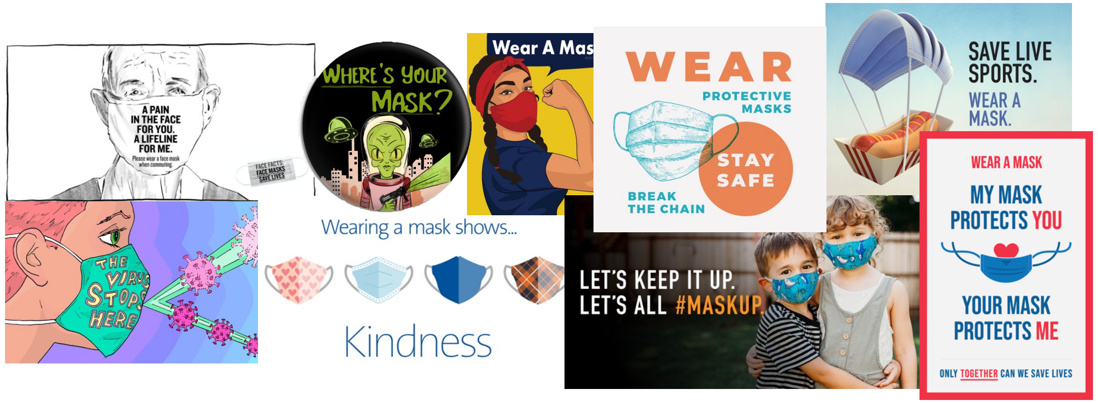
Graph Agnostic Randomized Experimental Design under Heterogeneous Linear Network Interference

Christina Lee Yu
Cornell University

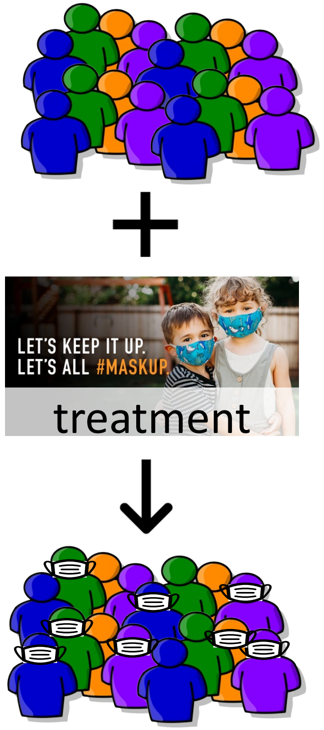
Joint work with Edoardo Airoldi, Christian Borgs, Jennifer Chayes,
Mayleen Cortez, and Matthew Eichhorn



Which ad is
most effective?

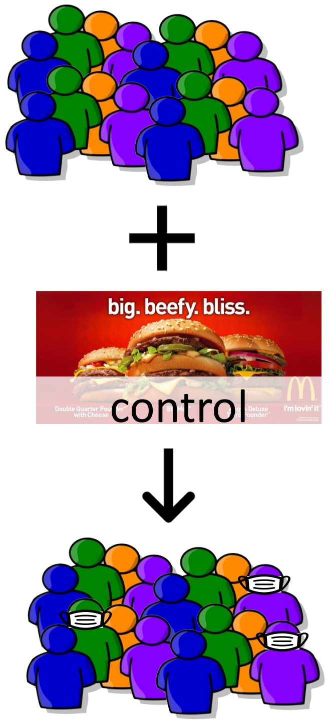
NEW YORK STATE | Department of Health

Hypothetical Scenario 1



treatment

Hypothetical Scenario 2

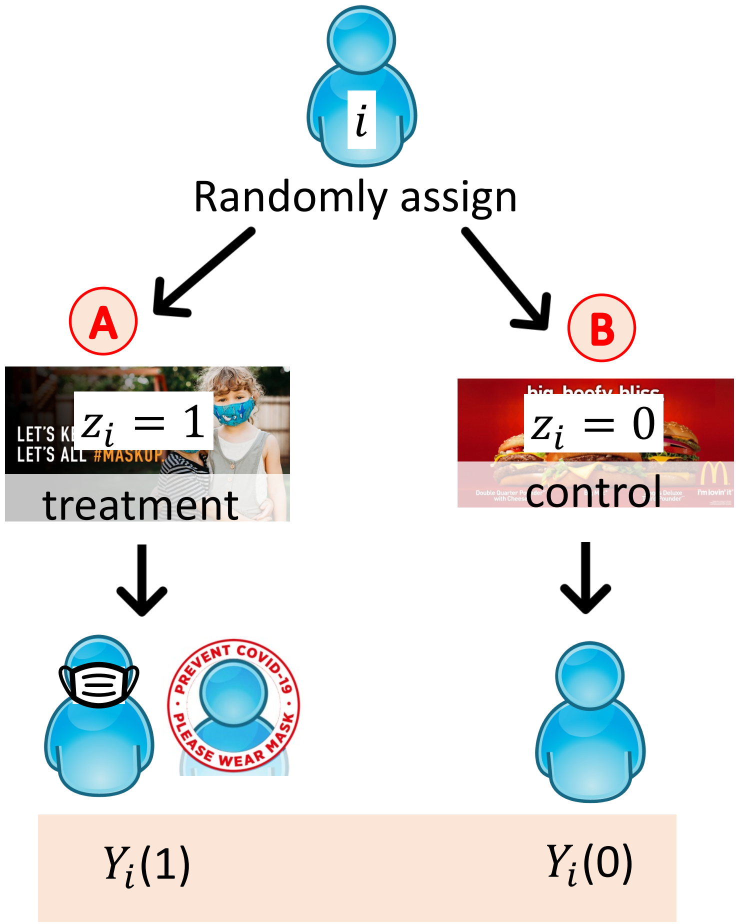


control

“Total Treatment Effect” – Average difference in total outcome

Randomized Experiment (A/B Testing)

NEW YORK STATE | Department of Health



i

Randomly assign

A

B

$z_i = 1$

treatment

$z_i = 0$

control

$Y_i(1)$

$Y_i(0)$

Assumes i's outcome only depends on z_i
"Stable Unit Treatment Value Assumption" (SUTVA)

- "Total Treatment Effect"

$$TTE = \frac{1}{n}\sum_{i\in[n]} Y_i(1) - \frac{1}{n}\sum_{i\in[n]} Y_i(0)$$

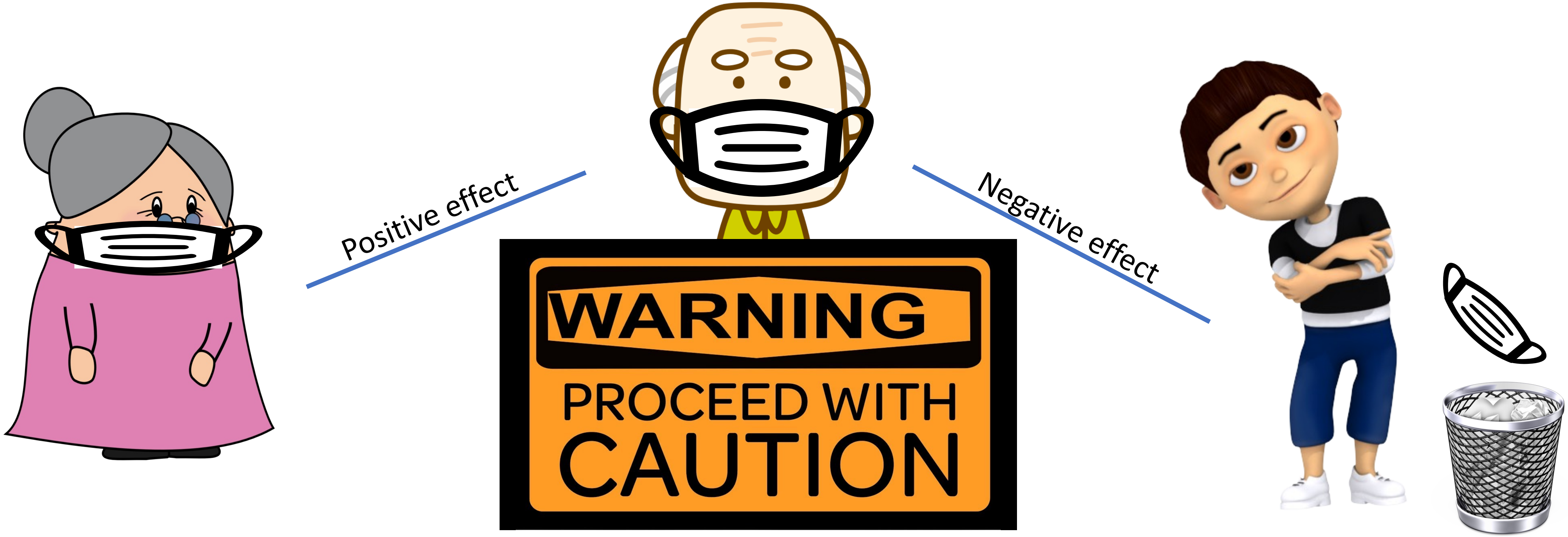
- Difference in Means Estimator

$$\widehat{TTE} = \overset{A}{\frac{\sum_{i\in[n]} z_i Y_i(z_i)}{\sum_{i\in[n]} z_i}} - \overset{B}{\frac{\sum_{i\in[n]} (1-z_i) Y_i(z_i)}{\sum_{i\in[n]} (1-z_i)}}$$

- Use randomization to get unbiasedness
- Relies on SUTVA

Network Interference

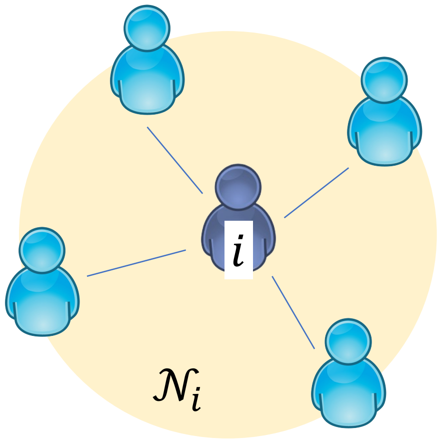
What if an individual’s outcome is also a function of the treatment of his/her neighbors?



SUTVA violated!

Challenge of Network Interference

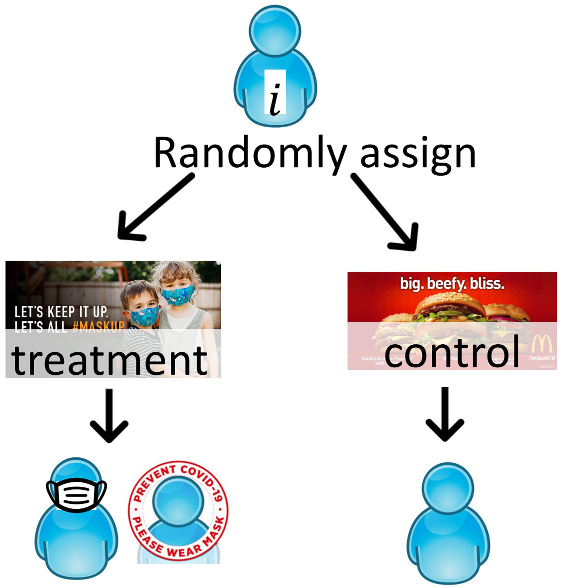
What if an individual's outcome is also a function of the treatment of his/her neighbors?



$Y_i(z_i, z_{\mathcal{N}_i})$ denotes i's outcome

"Total Treatment Effect"

$$TTE = \frac{1}{n}\sum_{i\in[n]} Y_i(1,\mathbf{1}) - \frac{1}{n}\sum_{i\in[n]} Y_i(0,\mathbf{0})$$



Randomly assign

treatment

control

$Y_i(1, z_{\mathcal{N}_i})$

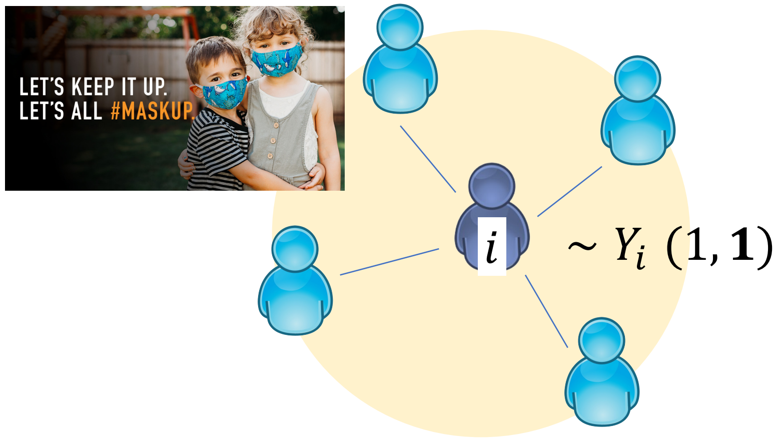
$Y_i(0, z_{\mathcal{N}_i})$

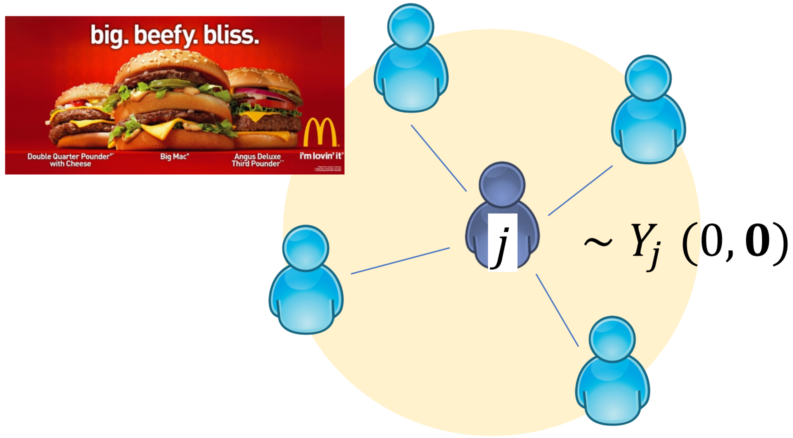
Problem: We may never observe $Y_i(1,\mathbf{1})$ or $Y_i(0,\mathbf{0})$!

Classical guarantees no longer hold with no further model assumptions.

Challenge of Network Interference

- Goal is to estimate $TTE \ = \frac{1}{n}\sum_{i\in[n]} Y_i(1,\mathbf{1}) \ - \ \frac{1}{n}\sum_{i\in[n]} Y_i(0,\mathbf{0})$
- Problem: in fullest generality, observing $Y_i\left(1, z_{\mathcal{N}_i}\right)$ or $Y_i\left(0, z_{\mathcal{N}_i}\right)$ tells nothing about $Y_i(1,\mathbf{1})$ or $Y_i(0,\mathbf{0})$
- Without further assumptions, estimators limited to sets $\{i \ \text{ s.t. } \ z_i = 1$ and $\ z_{\mathcal{N}_i} = \mathbf{1}\}$ and $\{j \ \text{ s.t. } \ z_j = 0 \ \text{ and } \ z_{\mathcal{N}_j} = \mathbf{0}\}$




big. beefy. bliss.
Double Quarter Pounder with Cheese
Big Mac
Angus Deluxe Third Pounder
i'm lovin' it
j
$\sim Y_j\ (0,\mathbf{0})$

Challenge of Network Interference

- Goal is to estimate $TTE = \frac{1}{n}\sum_{i\in[n]} Y_i(1, \mathbf{1}) - \frac{1}{n}\sum_{i\in[n]} Y_i(0, \mathbf{0})$
- Problem: in fullest generality, observing $Y_i\left(1, z_{\mathcal{N}_i}\right)$ or $Y_i\left(0, z_{\mathcal{N}_i}\right)$ tells nothing about $Y_i(1, \mathbf{1})$ or $Y_i(0, \mathbf{0})$
- Without further assumptions, estimators limited to sets $\{i \text{ s.t. } z_i = 1$ and $z_{\mathcal{N}_i} = \mathbf{1}\}$ and $\{j \text{ s.t. } z_j = 0 \text{ and } z_{\mathcal{N}_j} = \mathbf{0}\}$
- Lose statistical power as only a few measurements used
- Designing randomization that admits unbiased and low variance estimators is computationally challenging for complex networks

☹

Potential Outcomes Model

- Under SUTVA, degrees of freedom in model is $2n$

$$Y_i : \{0,1\} \rightarrow \mathbb{R}$$

- Under neighborhood interference, degrees of freedom is $2^d n$

$$Y_i : \{0,1\}^{\mathcal{N}_i+1} \rightarrow \mathbb{R}$$

- Total observations from experiment is n
- Goal is to estimate $TTE \ = \frac{1}{n}\sum_{i\in[n]}\left(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\right)$
- Minimal assumptions: constant treatment response, effective treatments, exposure mapping, neighborhood treatment response [Aronow12] [Manski13] [AronowSamii17] [SussmanAiroldi17]

Simple first attempt

- Horvitz-Thompson estimator

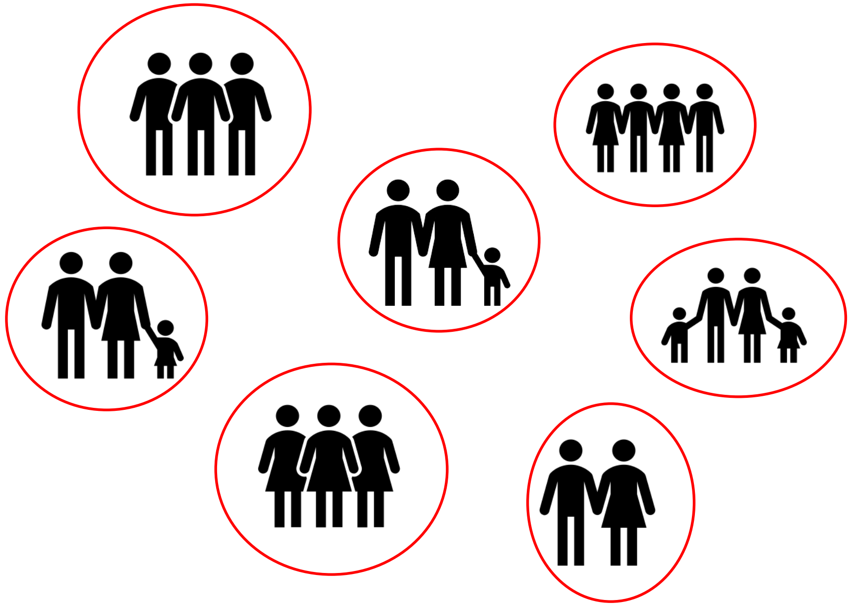
$$\widehat{TTE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}) \left(\frac{\mathbb{I}(\mathbf{z}_{\mathcal{N}_i \cup \{i\}} = \mathbf{1})}{\mathbb{P}(\mathbf{z}_{\mathcal{N}_i \cup \{i\}} = \mathbf{1})} - \frac{\mathbb{I}(\mathbf{z}_{\mathcal{N}_i \cup \{i\}} = \mathbf{0})}{\mathbb{P}(\mathbf{z}_{\mathcal{N}_i \cup \{i\}} = \mathbf{0})} \right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}) \left(\prod_{j \in \mathcal{N}_i \cup \{i\}} \frac{z_j}{p_j} - \prod_{j \in \mathcal{N}_i \cup \{i\}} \frac{1 - z_j}{1 - p_j} \right)$$

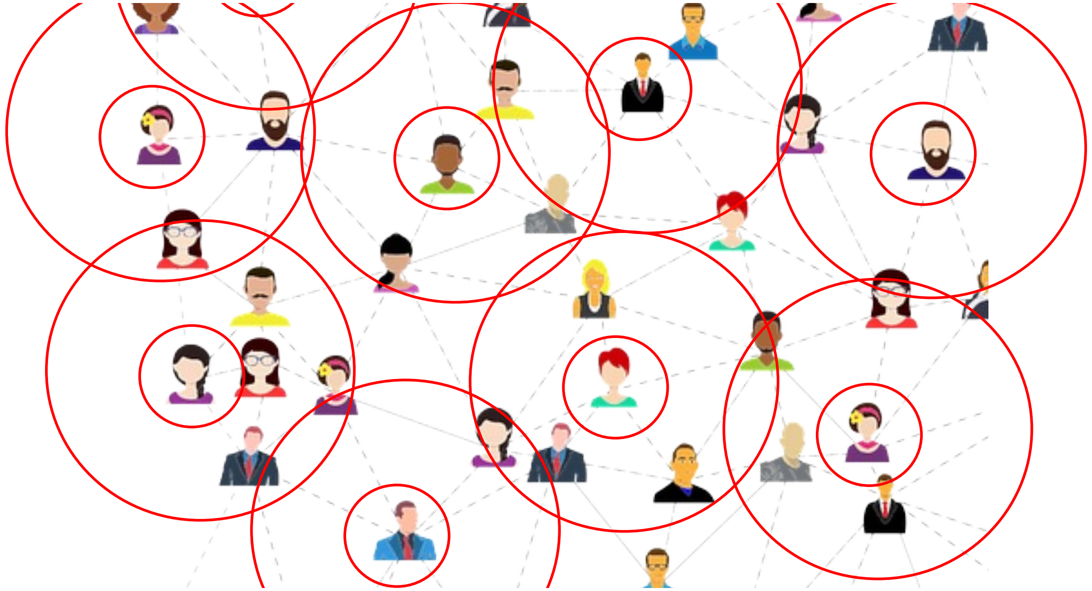
- Variance under Bernoulli design is $O\left(\frac{Y_{\max}^2 d^2}{np^d}\right)$
- Can we do better?

Brief Literature Review

- “nonparametric” approaches – focus on designing clever designs
  - Depends heavily on graph structure (clusterable)
  - Computationally complex randomized designs or high bias/variance



(a) fully disconnected [Sobel06][Rosenbaum07] [HudgensHalloran08][TchetgenVanderWeele12] and more



(a) 3-net clustering for restricted-growth graphs [GuiXuBhasinHan15] [EcklesKarrerUgander17] [UganderKarrerBackstromKleinberg13]

Brief Literature Review

- “nonparametric” approaches – focus on designing clever designs
    - Depends heavily on graph structure (clusterable)
    - Computationally complex randomized designs or high bias/variance
- “parametric” approaches
    - Requires more data than parameters to fully identify model
    - Regression style methods, includes many ML approaches [ToulisKao13] [GuiXuBhasinHan15] [BasseAiroldi15] [Cai2015] [Parker2016] [Chin2019]
    - Fragile to model misspecification, but fewer requirements on randomization

Brief Literature Review

- “nonparametric” approaches – focus on designing clever designs
  - Depends heavily on graph structure (clusterable)
  - Computationally complex randomized designs or high bias/variance
- “parametric” approaches
  - Requires more data than parameters to fully identify model
  - Regression style methods, includes many ML approaches [ToulisKao13] [GuiXuBhasinHan15] [BasseAiroldi15] [cai2015social] [parker2016optimal] [chin2019regression]
  - Fragile to model misspecification, but fewer requirements on randomization
- All previous solutions require (approx) knowledge of network!!
- In nonparametric setting, how can we exploit model structure?

Key Question

In the presence of network interference, does there exist any simple and efficient solution for estimating Total Treatment Effect without critically relying on the knowledge of network structure or restrictive network properties?

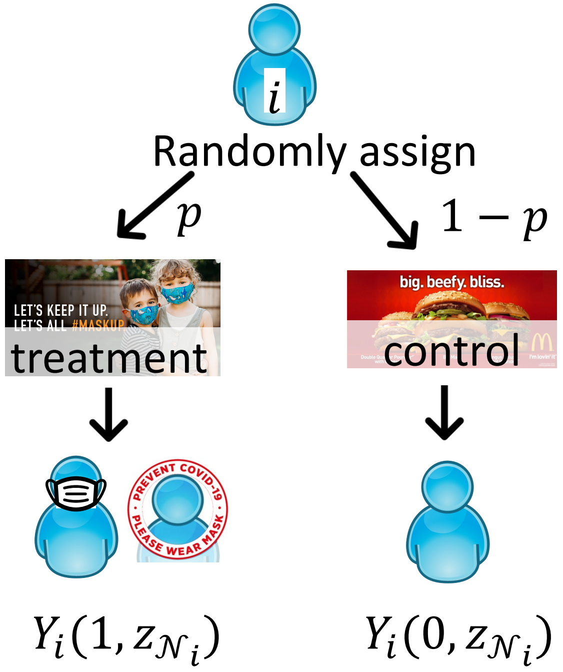
*solution must require imposing appropriate model assumptions …

Preview of Result

- Assume we have knowledge of average baseline $\bar{\alpha} := \frac{1}{n}\sum_{i\in[n]} Y_i(\mathbf{0})$
- $\widehat{TTE} = \frac{1}{p}\left(\frac{1}{n}\sum_{i\in[n]} Y_i(\mathbf{z}) - \bar{\alpha}\right)$

+ suitable model

- Under Bernoulli randomized design, $\widehat{TTE}$ is unbiased with variance $O\left(\frac{d^2}{pn}\right)$
- No knowledge/assumptions on graph!!



i

Randomly assign

p

$1 - p$

treatment

control

$Y_i(1, z_{\mathcal{N}_i})$

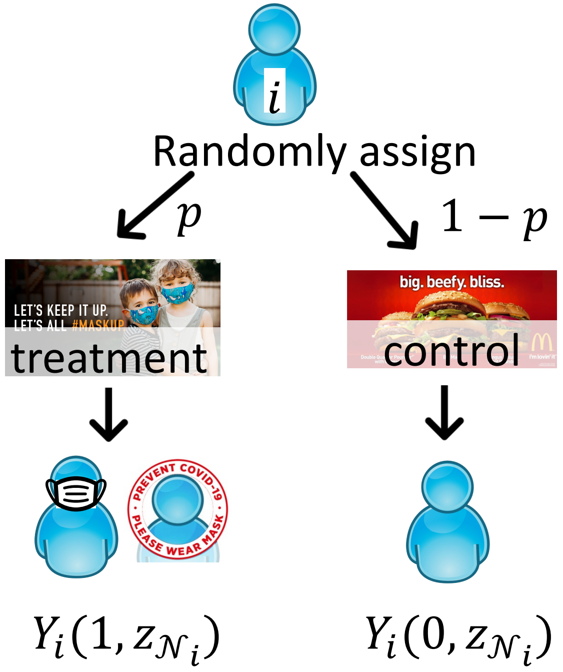
$Y_i(0, z_{\mathcal{N}_i})$

Preview of Result

- Assume we have knowledge of average baseline $\bar{\alpha} := \frac{1}{n}\sum_{i\in[n]} Y_i(\mathbf{0})$
  - Easy to estimate from historical data or pilot surveys
  - Easy to collect this data before experiment begins
- $\widehat{TTE} = \frac{1}{p}\left(\frac{1}{n}\sum_{i\in[n]} Y_i(\boldsymbol{z}) - \bar{\alpha}\right)$

+ suitable model

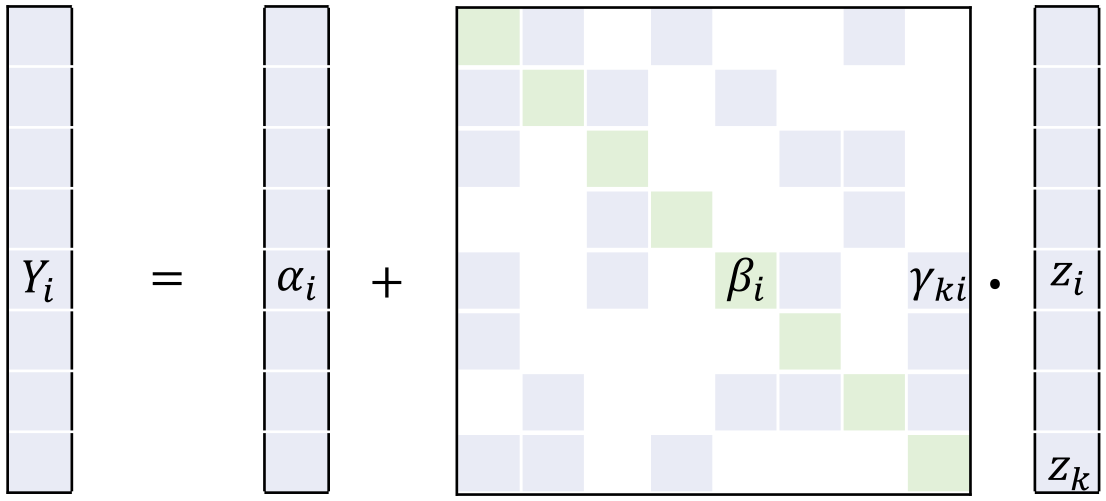
- Under Bernoulli randomized design, $\widehat{TTE}$ is unbiased with variance $O\left(\frac{d_{\max}^2}{pn}\right)$
- No knowledge/assumptions on graph!!



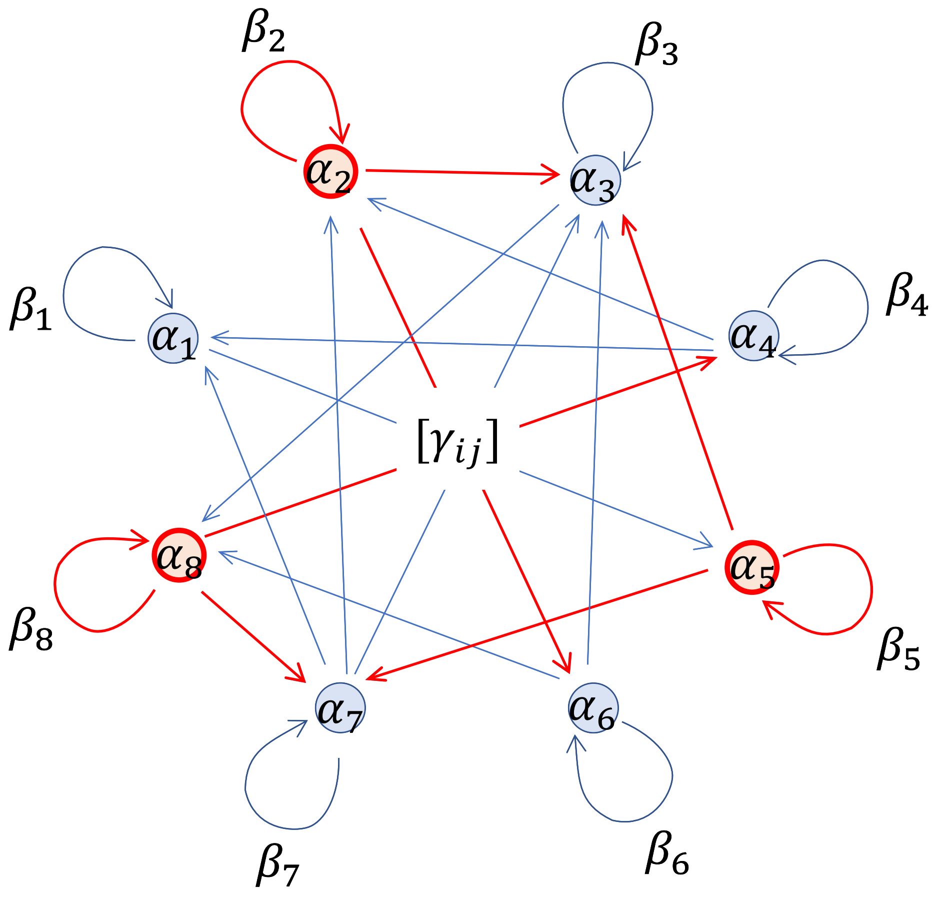
i
Randomly assign
p
$1-p$
LET'S KEEP IT UP. LET'S ALL #MASKUP
treatment
big. beefy. bliss.
control
PREVENT COVID-19 · PLEASE WEAR MASK ·
$Y_i(1, z_{\mathcal{N}_i})$
$Y_i(0, z_{\mathcal{N}_i})$

Heterogeneous Linear Outcomes Model

$$Y_i(\boldsymbol{z}) = \alpha_i + \beta_i z_i + \overbrace{\textstyle\sum_{k\in\mathcal{N}_i}\gamma_{ki} z_k}^{\text{Additive network effects}}$$



$$[\boldsymbol{Y}] \quad = \quad [\boldsymbol{\alpha}] \quad + \quad (\mathrm{diag}(\boldsymbol{\beta}) + [\boldsymbol{\gamma}]^T) \cdot [\boldsymbol{z}]$$



Heterogeneous Linear Outcomes Model

$$\boxed{Y_i(\boldsymbol{z}) = \alpha_i + \beta_i z_i + \sum_{k \in \mathcal{N}_i} \gamma_{ki} z_k}$$

- Allows for full heterogeneity in $\alpha_i, \beta_i, \gamma_{ki}$, can be positive or negative
- More parameters (2n + #edges) than possible measurements (n)
- Can capture endogenous peer effects such as contagion

$$Y_i(\boldsymbol{z}) = a_i + b_i z_i + \sum_{k \in \mathcal{N}_i} c_{ki} Y_k(\boldsymbol{z})$$

$$(I \ - C)\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{a} + \mathrm{diag}(\boldsymbol{b}) \cdot \boldsymbol{z}$$

$$\boldsymbol{Y}(\boldsymbol{z}) = \underbrace{(I \ - C)^{-1}\boldsymbol{a}}_{\boldsymbol{\alpha}} + \underbrace{\sum_t^{\infty} C^t \,\mathrm{diag}(\boldsymbol{b})}_{\mathrm{diag}(\boldsymbol{\beta}) + [\boldsymbol{\gamma}]^T} \cdot \boldsymbol{z}$$

Heterogeneous Linear Outcomes Model

$$\boxed{Y_i(\boldsymbol{z}) = \alpha_i + \beta_i z_i + \sum_{k \in \mathcal{N}_i} \gamma_{ki} z_k}$$

- Allows for full heterogeneity in $\alpha_i, \beta_i, \gamma_{ki}$, can be positive or negative
- More parameters (2n + #edges) than possible measurements (n)
- Can capture endogenous peer effects such as contagion
- Can easily add mean zero independent measurement noise

Total Treatment Effect with baseline estimates

Given knowledge of average baselines $\bar{\alpha}$, for any randomized design such that $\mathbb{E}[z_i] = p$ for all $i \in [n]$, the following simple estimator

$$\widehat{TTE} = \frac{1}{p}\left(\frac{1}{n}\sum_{i\in[n]} Y_i(\boldsymbol{z}) - \overbrace{\frac{1}{n}\sum_{i\in[n]} \alpha_i}^{\bar{\alpha}}\right)$$

is an unbiased and efficient estimator for any network under the heterogeneous linear outcomes model.

Does not require knowledge of the network!!

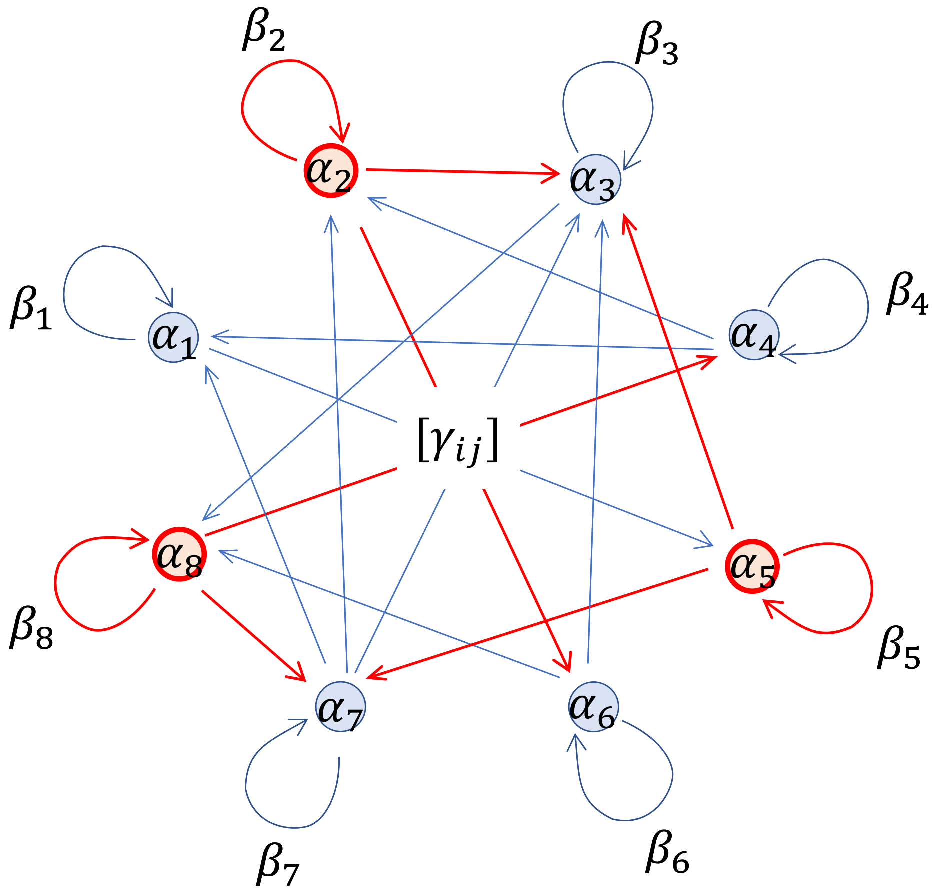
Total Treatment Effect with baseline estimates

- Total treatment effect equals sum of weighted edges

$$TTE = \frac{1}{n}\sum_i\left(\beta_i + \sum_k \gamma_{ki}\right)$$

- Treating an individual "activates" its outgoing edges
- Estimator is sum of activated edges

$$\widehat{TTE} = \frac{1}{pn}\sum_{i\in[n]} Y_i(\mathbf{z}) - \frac{1}{pn}\sum_{i\in[n]} \alpha_i$$
$$= \frac{1}{pn}\sum_{i\in[n]} \underbrace{\left(\beta_i + \sum_{k\in[n]} \gamma_{ik}\right)}_{\text{“influence” } L_i} z_i$$



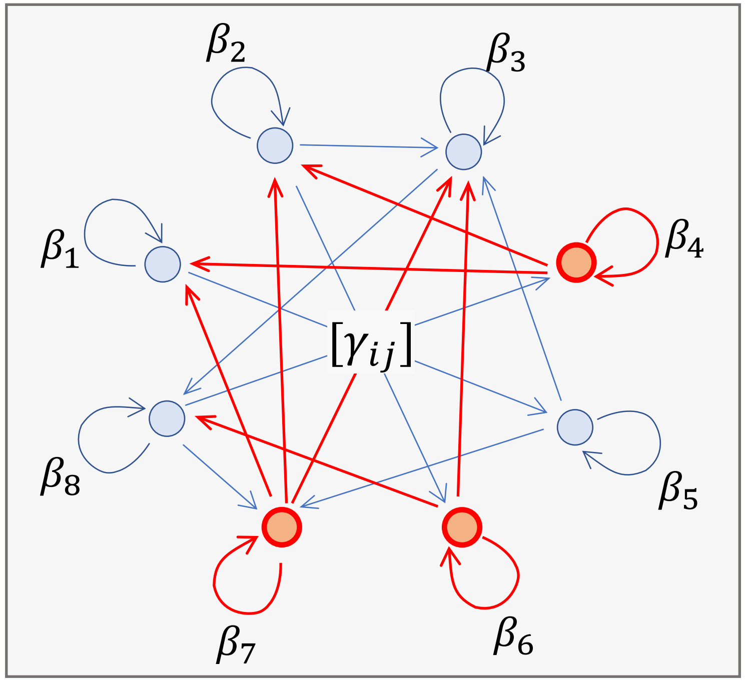
Reduction to estimating population mean

$$\widehat{TTE} = \frac{1}{pn}\sum_{i\in[n]}\overbrace{\left(\beta_i + \sum_{k\in[n]}\gamma_{ik}\right)}^{L_i} z_i \qquad \widehat{TTE} = \frac{1}{pn}\sum_{i\in[n]} L_i z_i$$

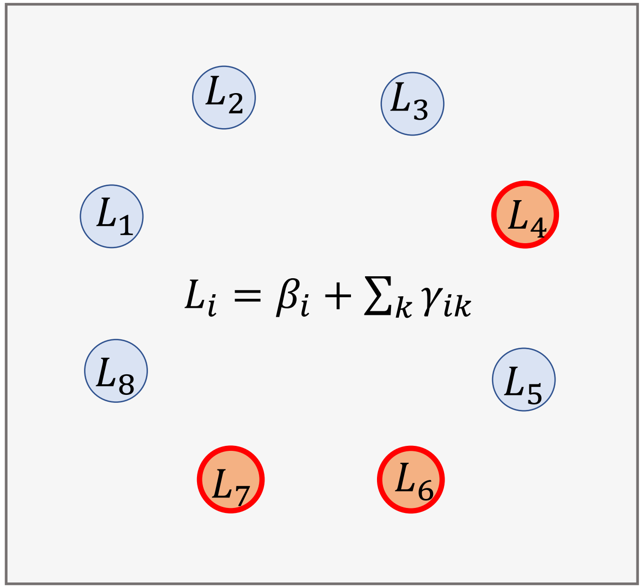
equivalent

$$TTE = \frac{1}{n}\sum_i\left(\beta_i + \sum_k \gamma_{ki}\right) \qquad TTE = \frac{1}{n}\sum_{i\in[n]} L_i$$

Given baseline estimates, network causal inference is as easy as estimating population mean!



equivalent



Reduction to estimating population mean

$$\widehat{TTE} = \frac{1}{pn}\sum\nolimits_{i\in[n]}\overbrace{\left(\beta_i + \sum\nolimits_{k\in[n]}\gamma_{ik}\right)}^{L_i} z_i$$

$$TTE = \frac{1}{n}\sum\nolimits_i\left(\beta_i + \sum\nolimits_k \gamma_{ki}\right)$$

$$\xLeftrightarrow{\text{equivalent}}$$

$$\widehat{TTE} = \frac{1}{pn}\sum\nolimits_{i\in[n]} L_i z_i$$

$$TTE = \frac{1}{n}\sum\nolimits_{i\in[n]} L_i$$

Given baseline estimates, network causal inference is as easy as estimating population mean!

- Easy to show unbiasedness, i.e. $\mathbb{E}\left[\widehat{TTE}\right] = TTE$
- Easy to show low variance under simple designs, e.g. for Bernoulli design

$$\mathrm{Var}\left[\widehat{TTE}\right] = \frac{1-p}{pn}\left(\frac{1}{n}\sum_{i\in[n]} L_i^2\right) \approx \frac{\overline{L^2}}{pn}$$

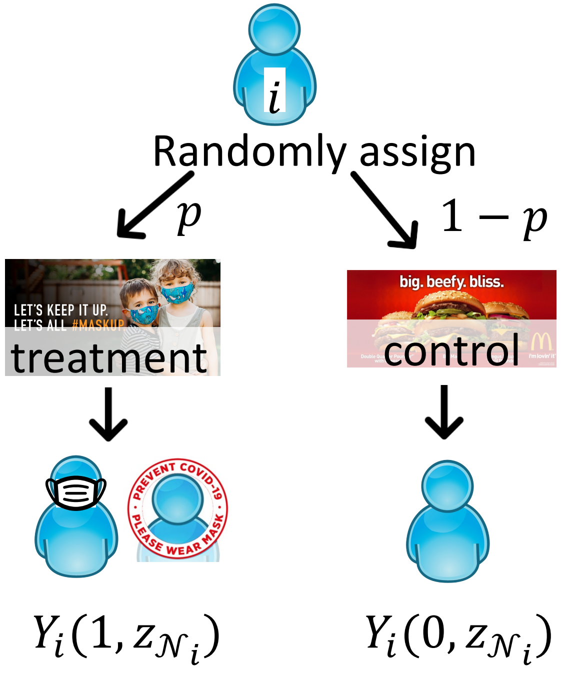
- Approach + guarantees allows for fully dense network
- Does NOT require any knowledge of underlying network

Summary of Results (part 1)

- Given baseline estimates, assuming a heterogeneous linear outcomes model, network causal inference is as easy as estimating population mean!
- Works for *any* arbitrary and unknown network

$$\widehat{TTE} = \frac{1}{pn}\sum_{i\in[n]} Y_i(z) - \frac{1}{pn}\sum_{i\in[n]} \alpha_i$$

- Unbiased and statistically consistent for $p \gg \frac{L_{\max}^2}{n}$



i

Randomly assign

p

$1 - p$

treatment

control

$Y_i(1, z_{\mathcal{N}_i})$

$Y_i(0, z_{\mathcal{N}_i})$

So easy!!

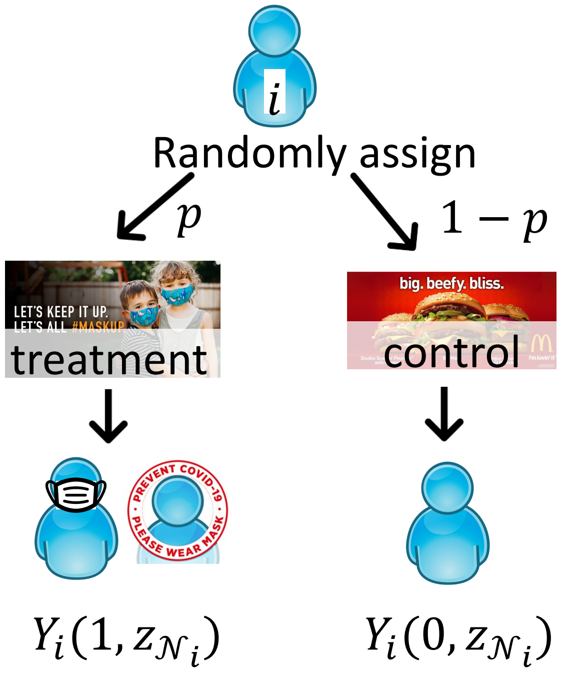
"Graph Agnostic Randomized Experimental Design under Heterogeneous Linear Network Interference". Christina Lee Yu, Edo Airoldi, Christian Borgs, and Jennifer Chayes.
Preprint at https://people.orie.cornell.edu/cleeyu/Network_Causal_Inference_full.pdf.

Summary of Results

- Given baseline estimates, assuming a heterogeneous linear outcomes model, network causal inference is as easy as estimating population mean!
- Works for *any* arbitrary and unknown network

$$\widehat{TTE} = \frac{1}{pn}\sum_{i\in[n]} Y_i(z) - \frac{1}{pn}\sum_{i\in[n]} \alpha_i$$

- Unbiased and statistically consistent for $p \gg \frac{L_{\max}^2}{n}$
- What if we don't know $\bar{\alpha}$? e.g. time fixed effects
- What about observational datasets?
- What if the model is not linear?



i

Randomly assign

p

$1 - p$

treatment

control

$Y_i(1, z_{\mathcal{N}_i})$

$Y_i(0, z_{\mathcal{N}_i})$

So easy!!

Ongoing work with Mayleen Cortez and Matthew Eichhorn