Bipartite Causal Inference with Interference: Estimating Health Impacts of Power Plant Regulations

Corwin M. Zigler

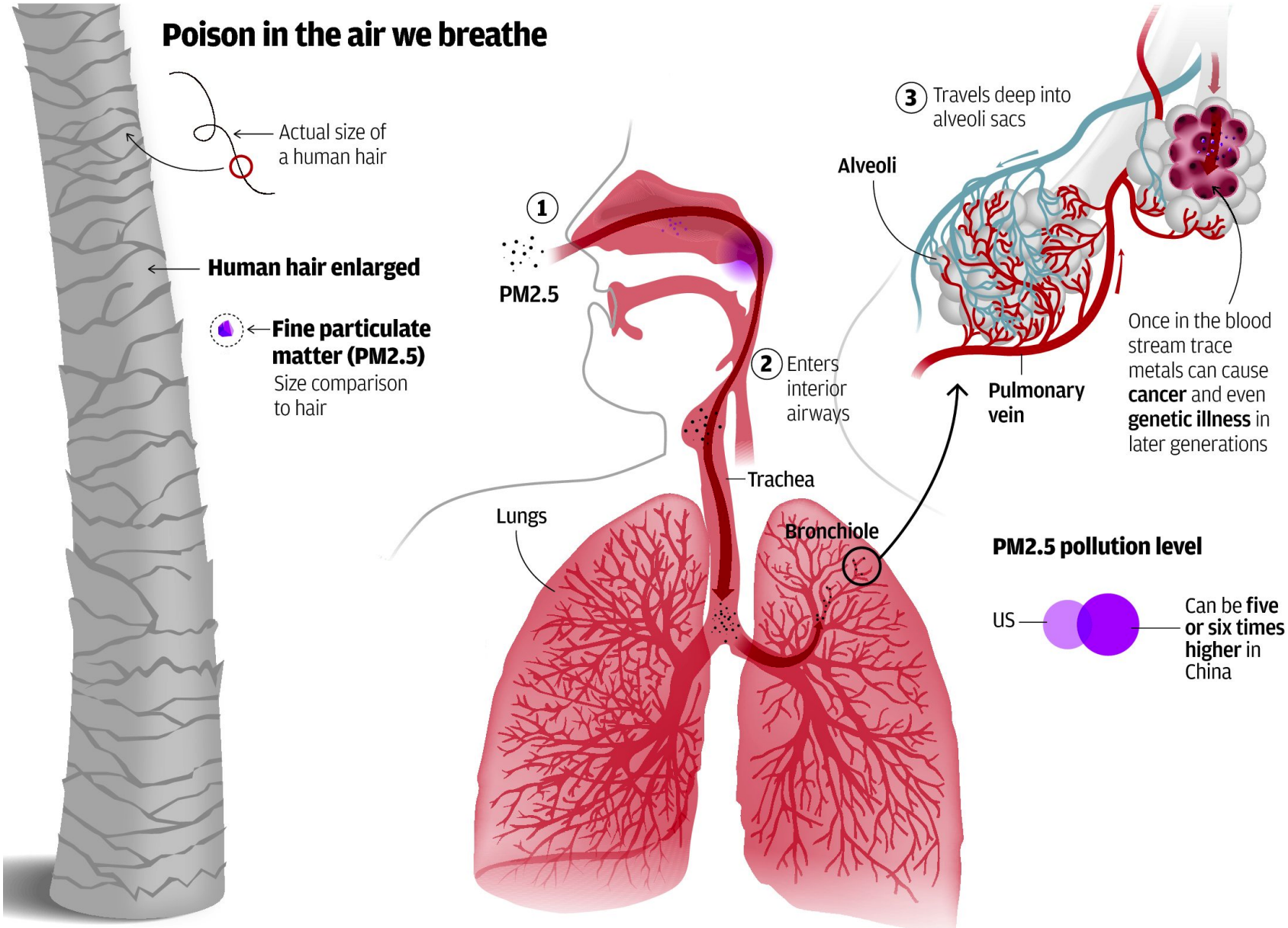
with Georgia Papdogeorgou, Lucas Henneman, Christine Choirat, Laura Forastiere, Fabrizia Mealli

June 12, 2018

The Motivating Problem

Evaluating Pollution Interventions on Power Plants

Air Pollution is Bad For You!



Sources:EPA, Environmental Protection Department, Greenpeace.

SCMP Graphic: Adolfo Arranz

"Effects of Pollution" $\rightarrow$ "Effectiveness of Interventions"

- $PM_{2.5}$ $\rightarrow$ increased health burden
- Regulations to limit exposure to air pollution
  - 58% to 80% of monetized benefits of all federal regulations
  - 44% to 54% of monetized costs
  - $\approx$ \$65 billion annually
  - High stakes, controversy, legal challenges, policy debates, etc.
  - Key pollutant: Fine particulate matter ($PM_{2.5}$)

"Effects of Pollution" $\rightarrow$ "Effectiveness of Interventions"

- $PM_{2.5}$ $\rightarrow$ increased health burden
- Regulations to limit exposure to air pollution
  - 58% to 80% of monetized benefits of all federal regulations
  - 44% to 54% of monetized costs
  - $\approx$ \$65 billion annually
  - High stakes, controversy, legal challenges, policy debates, etc.
  - Key pollutant: Fine particulate matter ($PM_{2.5}$)
- Questions surrounding the **effectiveness of regulatory actions** for
  - Improving air quality
  - Improving public health

Major Pollution Source: Power Plants

$\Rightarrow$ Many regulations to reduce emissions

- Interventions to reduce emissions
  1. Selective Catalytic Reduction (SCR) to reduce NO_x
  2. Scrubbers to reduce SO_2
- Both should reduce ambient pollution (e.g., $PM_{2.5}$

Major Pollution Source: Power Plants

$\Rightarrow$ Many regulations to reduce emissions

- Interventions to reduce emissions
  1. Selective Catalytic Reduction (SCR) to reduce NO_x
  2. Scrubbers to reduce SO_2
- Both should reduce ambient pollution (e.g., $PM_{2.5}$

Do these interventions improve hospitalization outcomes among Medicare beneficiaries?

1. Cardiovascular disease (CVD)
2. Ischemic Heart Disease (IHD)

Major Pollution Source: Power Plants

$\Rightarrow$ Many regulations to reduce emissions

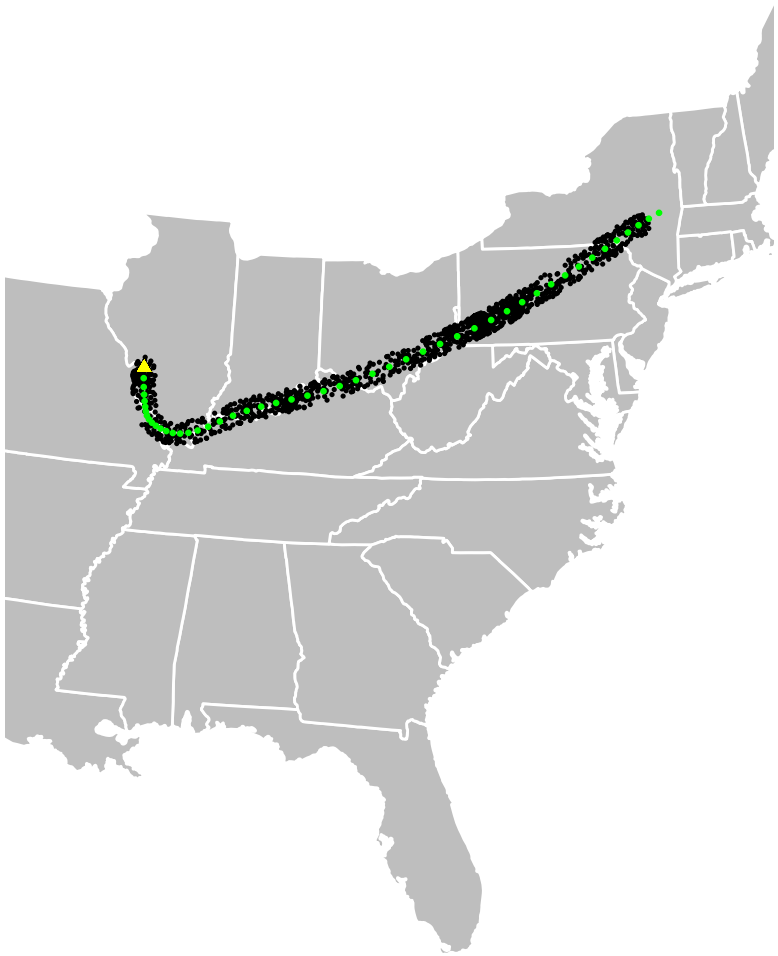
- Interventions to reduce emissions
  1. Selective Catalytic Reduction (SCR) to reduce NO_x
  2. Scrubbers to reduce SO_2
- Both should reduce ambient pollution (e.g., $PM_{2.5}$

Do these interventions improve hospitalization outcomes among Medicare beneficiaries?

1. Cardiovascular disease (CVD)
2. Ischemic Heart Disease (IHD)

Long-range Pollution Transport

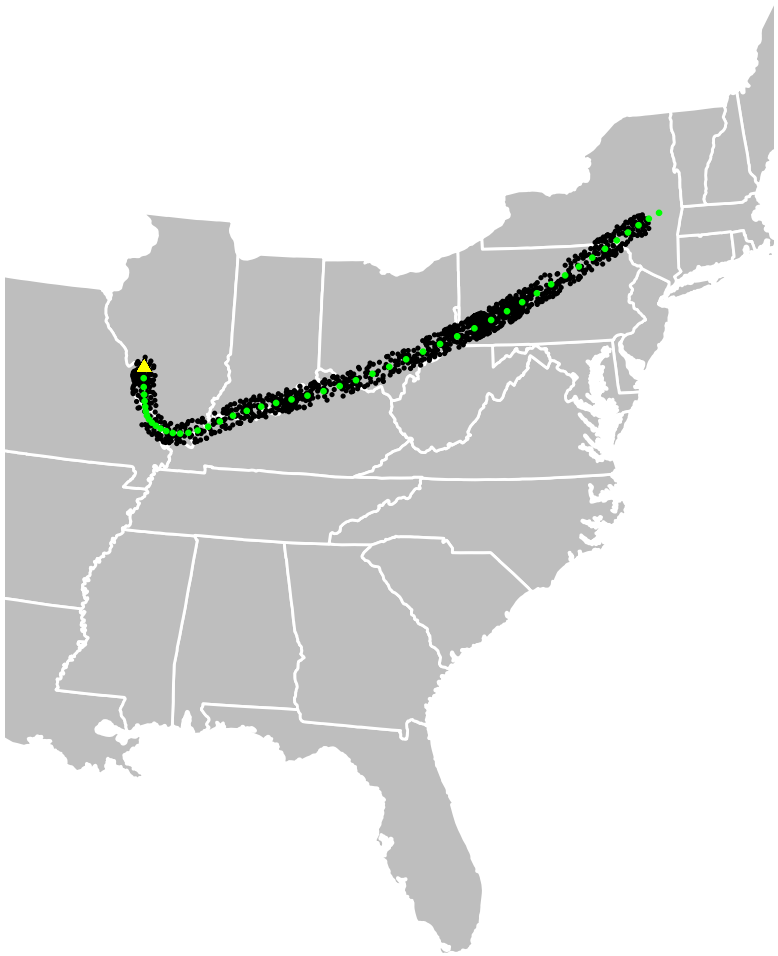
Movement of pollution from sources $\rightarrow$ populations



- Emissions originating at a power plant (**yellow triangle**) *move*
  - Long distances towards conversion to harmful pollution

Long-range Pollution Transport

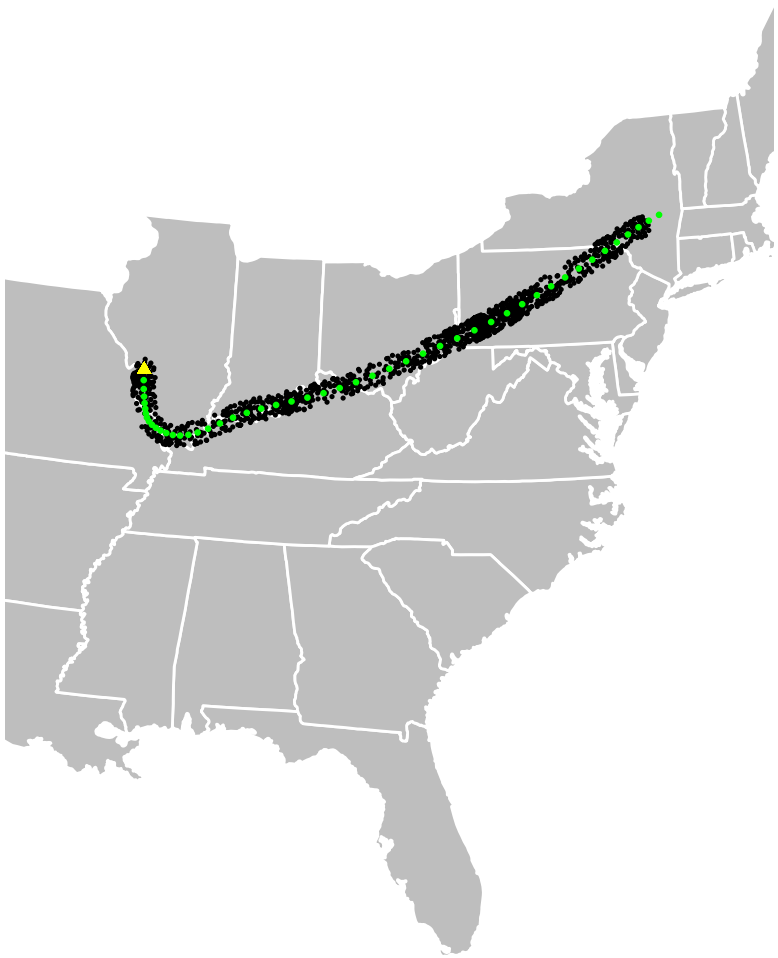
Movement of pollution from sources → populations



- Emissions originating at a power plant (**yellow triangle**) *move*
  - Long distances towards conversion to harmful pollution
- Intervening at a power plant has implications for pollution and health at *many* zip codes (**black dots**)

Long-range Pollution Transport

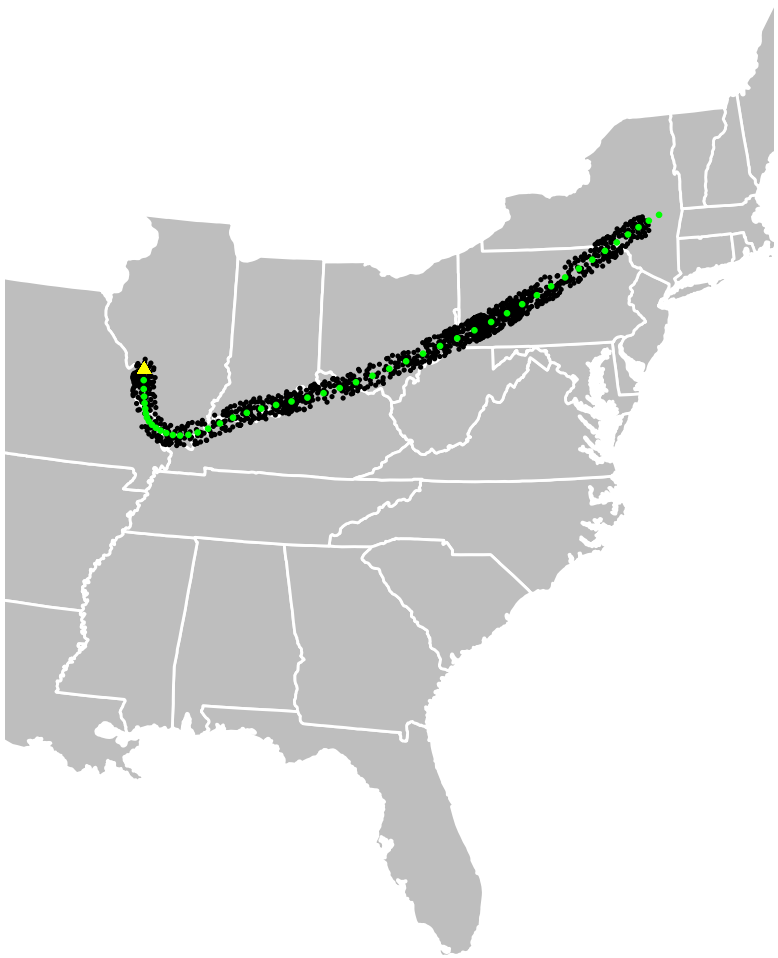
Movement of pollution from sources → populations



- Emissions originating at a power plant (**yellow triangle**) *move*
  - Long distances towards conversion to harmful pollution
- Intervening at a power plant has implications for pollution and health at *many* zip codes (**black dots**)
- Health at any given location is determined in part by interventions at *many* power plants

Long-range Pollution Transport

Movement of pollution from sources $\rightarrow$ populations

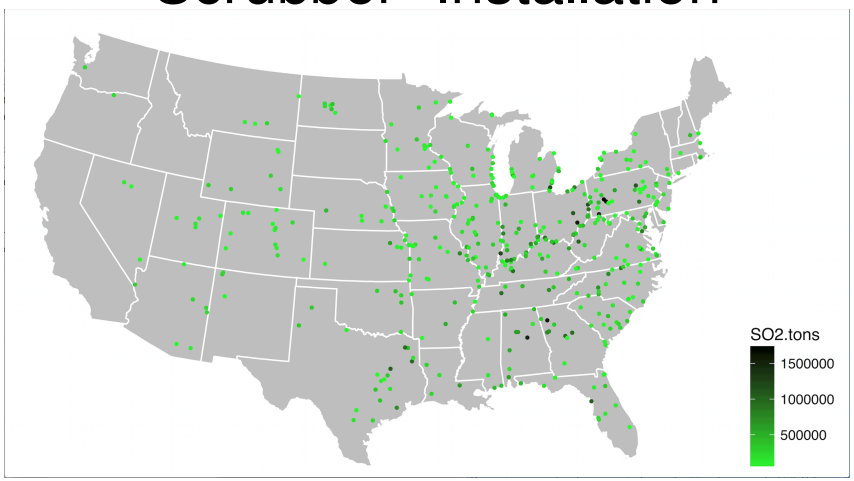


- Emissions originating at a power plant (**yellow triangle**) *move*
  - Long distances towards conversion to harmful pollution
- Intervening at a power plant has implications for pollution and health at *many* zip codes (**black dots**)
- Health at any given location is determined in part by interventions at *many* power plants
- $\Rightarrow$ inference on a "network" of interconnected power plants/population locations (**interference**)

Intervention Effects Propagate Across a Network

Network of 100s of power plants and 10Ks of population locations

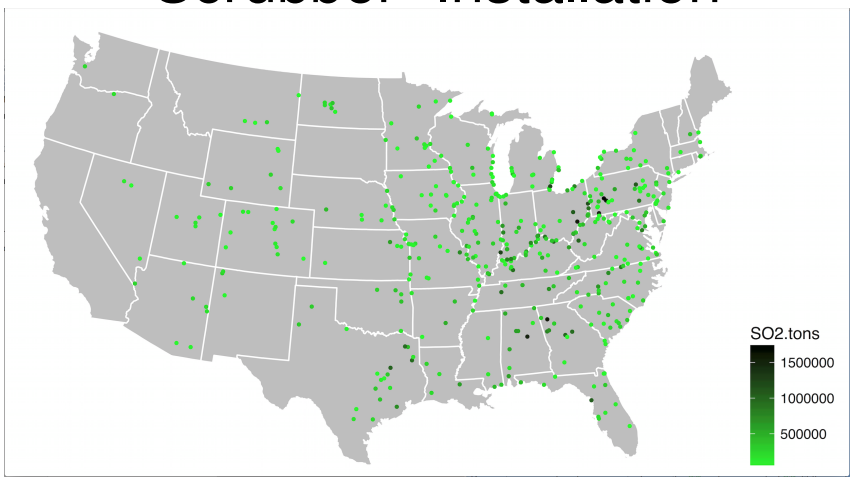
Interventions at Power Plants:
"Scrubber" Installation



Intervention Effects Propagate Across a Network

Network of 100s of power plants and 10Ks of population locations

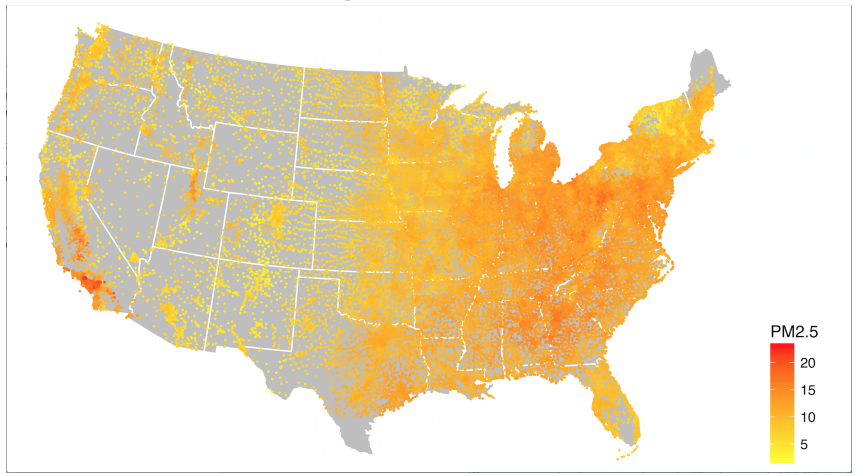
Interventions at Power Plants: "Scrubber" Installation



Affect

$\longrightarrow$

Health at Population Locations (Zip Codes)



Bipartite Causal Inference with Interference

Potential Outcomes Formalization

Defining Characteristics of Bipartite Causal Inference with Interference

Evaluating the causal effects of a (binary) intervention where:

1. "Bipartite" Structure

- The intervention is defined on one type of observational unit
  - E.g., whether a coal-fired power plant installs a "scrubber"
- Outcomes of interest are defined and measured on a second, distinct, type of unit
  - E.g., Hospitalizations measured at residential zip codes

Defining Characteristics of Bipartite Causal Inference with Interference

Evaluating the causal effects of a (binary) intervention where:

1. "Bipartite" Structure

- The intervention is defined on one type of observational unit
  - E.g., whether a coal-fired power plant installs a "scrubber"
- Outcomes of interest are defined and measured on a second, distinct, type of unit
  - E.g., Hospitalizations measured at residential zip codes

2. Interference

Interconnectedness between units $\Rightarrow$ outcomes for a particular unit depend upon treatments assigned to (multiple) other units

Formalization of Bipartite Structure

Two types of observational units:

1. *Interventional Units*
   - $\mathcal{P} = \{p_1, p_2, \ldots, p_i, \ldots, p_P\}$: a set of P power plants
   - Where interventions occur (or not)
   - $A_i = 1/0$ denotes presence/absence of intervention
   - Treatment allocation: $\mathbf{A} = (A_1, A_2, \ldots, A_P)$
   - Covariates $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_P)$

Formalization of Bipartite Structure

Two types of observational units:

1. *Interventional Units*
   - $\mathcal{P} = \{p_1, p_2, \ldots, p_i, \ldots, p_P\}$: a set of P power plants
   - Where interventions occur (or not)
   - $A_i = 1/0$ denotes presence/absence of intervention
   - Treatment allocation: $\mathbf{A} = (A_1, A_2, \ldots, A_P)$
   - Covariates $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_P)$

2. *Outcome Units*
   - $\mathcal{M} = \{m_1, m_2, \ldots, m_j, \ldots, m_M\}$: a set of M locations
   - Where outcomes are measured/of interest
   - Y_j denotes outcome if interest
   - Covariates: $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$

Potential Outcomes with Interference

"Standard" Setting: No Interference

Stable Unit Treatment Value Assumption (SUTVA):

$$Y_j(\mathbf{A}) = Y_j(A_j)$$

$$E[Y_j(A_j = 1) - Y_j(A_j = 0)]$$

Potential Outcomes with Interference

"Standard" Setting: No Interference

Stable Unit Treatment Value Assumption (SUTVA):

$$Y_j(\mathbf{A}) = Y_j(A_j)$$

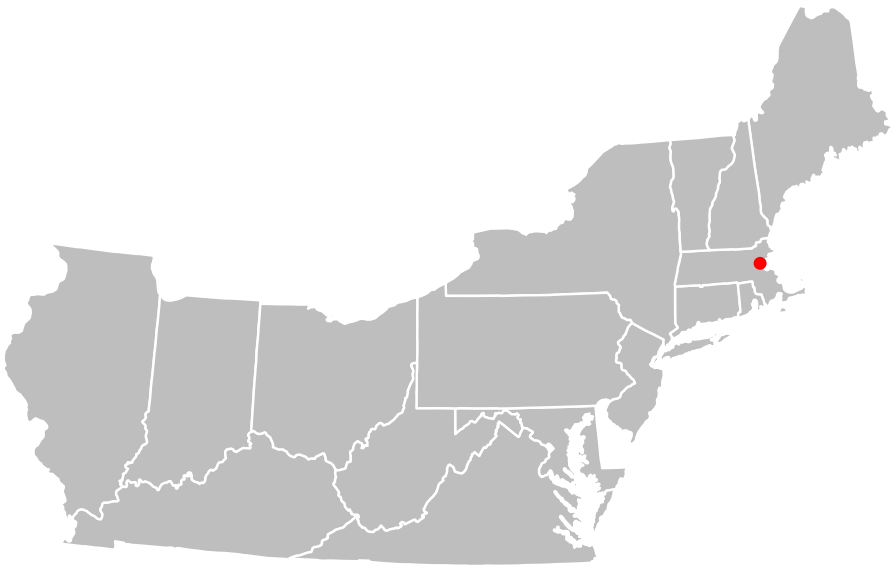
$$E[Y_j(A_j = 1) - Y_j(A_j = 0)]$$

Interference $\Rightarrow$ ~~SUTVA~~

- Potential outcome for j depends on $\mathbf{A}$ at *all other* locations
  - $Y_j(\mathbf{A}) = Y_j(A_1 = a_1, A_2 = a_2, \ldots, A_P = a_P)$
  - $\mathcal{A}(P) = \{0, 1\}^P$: space of possible allocations
  - 2^P unique potential outcomes for each j
  - Vastly complicates causal inference

Bipartite Causal Inference with Interference

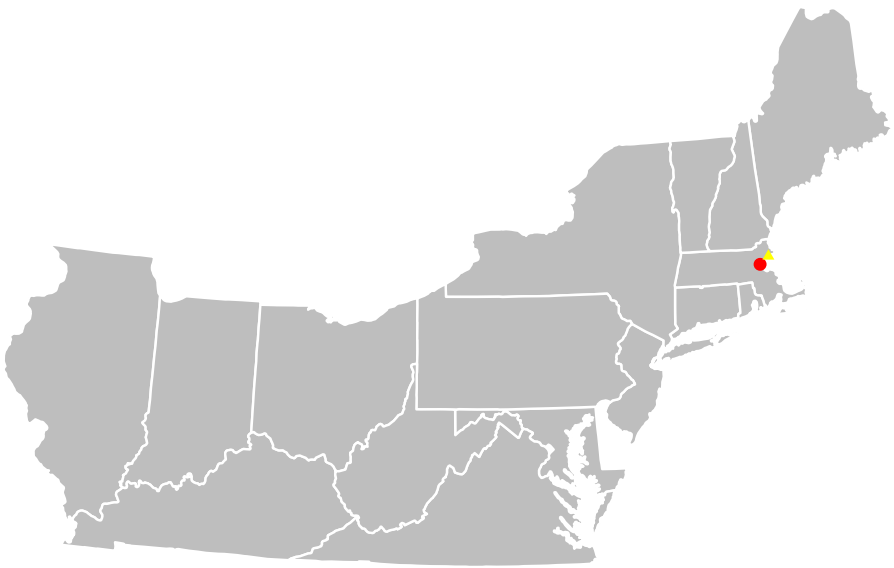
Amid the realities of pollution transport



Zip Code $j = 02138$ (**red dot**)
$Y_j(?) \equiv$ Potential IHD that would be observed...

Bipartite Causal Inference with Interference

Amid the realities of pollution transport

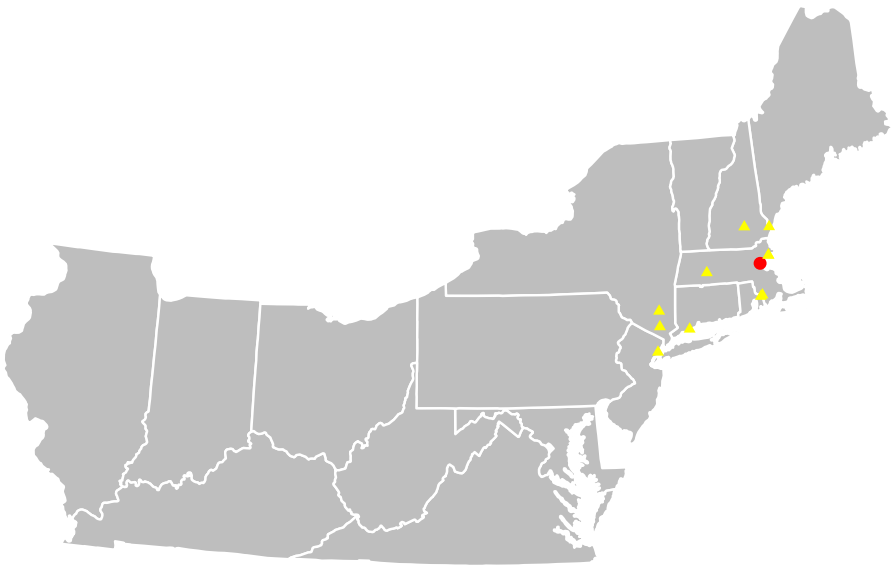


Zip Code $j = 02138$ (**red dot**)
$Y_j(A_i) \equiv$ Potential IHD that would be observed...

if the *most influential (or closest)* power plant were intervened upon

Bipartite Causal Inference with Interference

Amid the realities of pollution transport

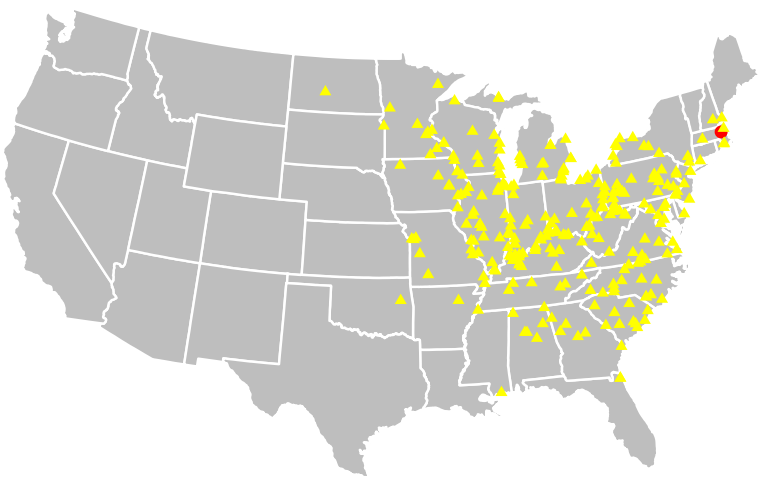


Zip Code $j = 02138$ (**red dot**)
$Y_j(A_1, \ldots, A_{10}) \equiv$ Potential IHD that would be observed...

if the10 most influential power plants were intervened upon

Bipartite Causal Inference with Interference

Amid the realities of pollution transport



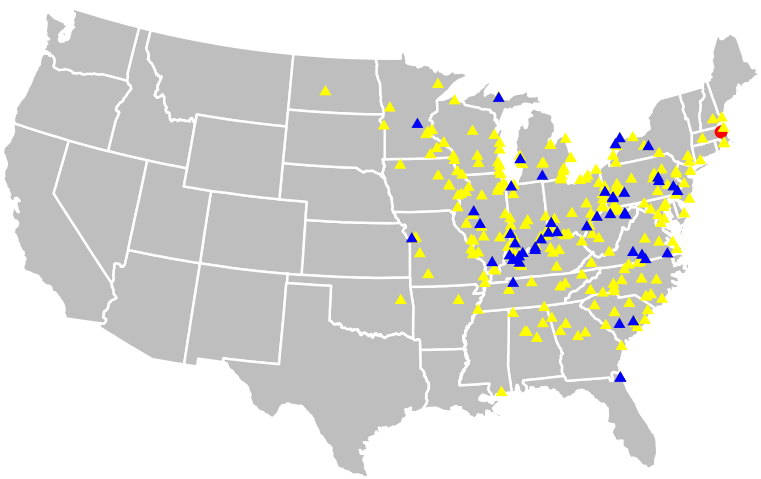
Zip Code $j = 02138$ (**red dot**)
$Y_j(A_1, \ldots, A_{278}) \equiv$ Potential IHD that would be observed...

under a particular intervention allocation to all 278 plants that influence this location

Intervening at any of these 278 plants affects hospitalization at 02138

Bipartite Causal Inference with Interference

Amid the realities of pollution transport



Zip Code $j = 02138$ (**red dot**)
$Y_j(A_1, \ldots, A_{278}) \equiv$ Potential IHD that would be observed...

under a particular scrubber allocation to all 278 plants that influence this location

Intervening at any of these 278 plants affects hospitalization at 02138
48 observed to have intervention

Interference $\rightarrow$ New Causal Quantites

Familiar question: “What is the (average) effect of the intervention on a the outcome?”

Interference $\rightarrow$ New Causal Quantites

Familiar question: "What is the (average) effect of the intervention on a the outcome?"

With **interference**,

(a) impossible to answer such questions without knowing what treatments are applied to *others* that are connected in some way

(b) new causal quantities may be of scientific interest

Interference $\rightarrow$ New Causal Quantites

Familiar question: "What is the (average) effect of the intervention on a the outcome?"

With **interference**,

(a) impossible to answer such questions without knowing what treatments are applied to *others* that are connected in some way

(b) new causal quantities may be of scientific interest

- More than just accounting for or addressing structure (e.g., spatial correlation)
- Budding research on interference in specialized settings
  - Infectious diseases
  - Social networks

Interconnectedness of Causal Questions

Question: What is the average effect (e.g., on hospitalizations) of intervening on the closest power plant?
Answer: It depends on how many interventions are on the "upwind" power plants.

Interconnectedness of Causal Questions

Question: What is the average effect (e.g., on hospitalizations) of intervening on the closest power plant?
Answer: It depends on how many interventions are on the "upwind" power plants.

Question: What is the average effect of intervening on "upwind" power plants?
Answer: It depends whether the closest power plant has the intervention.

Direct and Indirect (or Spillover) Effects

A_i: Treatment assignment of the the i^{th} power plant

$\mathbf{A}_{(-i)}$ Treatment assignments of all power plants *except i*

Direct and Indirect (or Spillover) Effects

A_i: Treatment assignment of the the i^{th} power plant

$\mathbf{A}_{(-i)}$ Treatment assignments of all power plants *except i*

Direct Effect (at individual $\mathcal{M}$-level)

$$DE_{(i,j)} = Y_j(A_i = 1, \mathbf{A}_{(-i)} = \boldsymbol{a}) - Y_j(A_i = 0, \mathbf{A}_{(-i)} = \boldsymbol{a})$$

- Effect on location *j* of intervening (vs. not) at power plant *i*, under a given allocation to all other plants
- Could be different for all possible values of $\mathbf{A}_{(-i)}$

Direct and Indirect (or Spillover) Effects

A_i: Treatment assignment of the the i^{th} power plant

$\mathbf{A}_{(-i)}$ Treatment assignments of all power plants *except i*

Direct Effect (at individual $\mathcal{M}$-level)

$$DE_{(i,j)} = Y_j(A_i = 1, \mathbf{A}_{(-i)} = \boldsymbol{a}) - Y_j(A_i = 0, \mathbf{A}_{(-i)} = \boldsymbol{a})$$

- Effect on location *j* of intervening (vs. not) at power plant *i*, under a given allocation to all other plants
- Could be different for all possible values of $\mathbf{A}_{(-i)}$

Indirect Effect (at individual $\mathcal{M}$-level)

$$IE_{(i,j)} = Y_j(A_i = 0, \mathbf{A}_{(-i)} = \boldsymbol{a}) - Y_j(A_i = 0, \mathbf{A}_{(-i)} = \boldsymbol{a'})$$

- Effect of changing intervention allocation to *other* upwind plants

Two Avenues for Methods Development

Drawing from several works in progress

Strategies for defining and estimating causal effects:

1. Clustering + "Partial Interference"
   - Simplify pollution transport as **clusters** of units
   - Inverse probability of treatment weighted (IPTW) estimator
2. Effects of "individual" and "neighborhood" treatments
   - Repurpose atmospheric engineering tools to characterize pollution transport
   - Generalized propensity score estimator

Bipartite Causal Inference with Interference

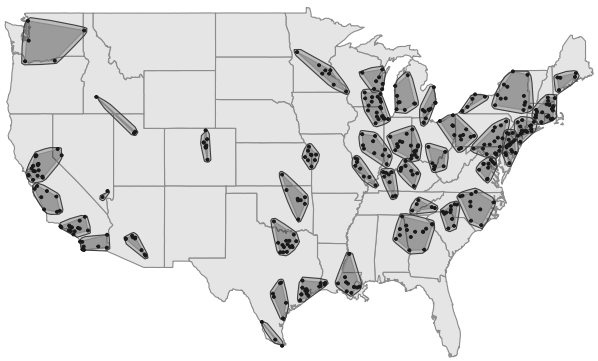
IPTW Estimators under Partial Interference

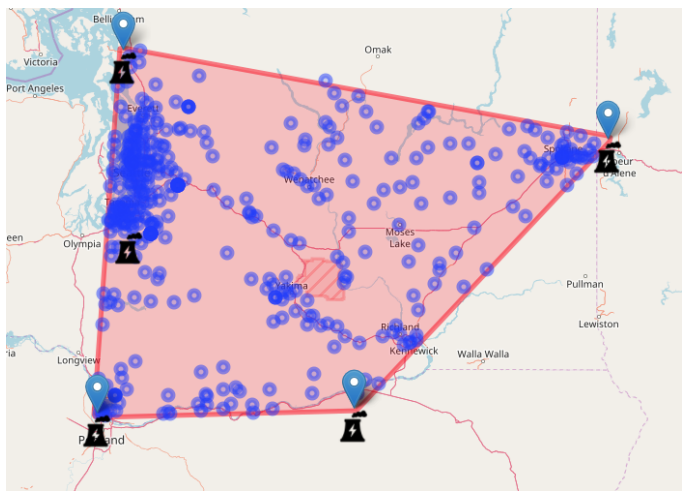
Partitioning Units into Clusters

A *very* simplified representation of pollution transport

Partition $\mathcal{M}, \mathcal{P}$ observations into K "aligned" groups

- $\{P^1, P^2, \ldots, P^K\}$, each of size $|P^k|$
  - Geographic clusters of power lants
- $\{M^1, M^2, \ldots, M^K\}$, each of size $|M^k|$
  - All zip codes within complex hull of P^k
- Cluster-level treatment assignment vector:
  - $\mathbf{A}^k = (A_1^k, A_2^k, \ldots, A_{|P^k|}^k) \in \mathcal{A}(|P^k|)$





Partial Interference Assumption

Commonly used in work on interference

- $Y_j^k(\mathbf{A})$ denote the outcome of unit j in the k^{th} cluster
- $\mathbf{A}^k$ is the treatment assignment vector of the k^{th} cluster

Partial Interference:

$$Y_j^k(\mathbf{A}) = Y_j^k(\mathbf{A}^k)$$

Potential outcomes only depend on treatments of other power plants in the corresponding cluster

Individual Average Potential Outcomes

- Let α denote a cluster-level propensity of intervention
- Define **individual average potential outcome** over all possible allocations described by α

$$\bar{Y}_j(A_i = a; \alpha) =$$

Individual Average Potential Outcomes

- Let α denote a cluster-level propensity of intervention
- Define **individual average potential outcome** over all possible allocations described by α

$$\bar{Y}_j(A_i = a; \alpha) = \sum_{\mathbf{s} \in \mathcal{A}(|P^k|-1)} Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{s})$$

Individual Average Potential Outcomes

- Let α denote a cluster-level propensity of intervention
- Define **individual average potential outcome** over all possible allocations described by α

$$\bar{Y}_j(A_i = a; \alpha) = \sum_{\mathbf{s} \in \mathcal{A}(|P^k|-1)} Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{s}) \pi(s | A_i = a; \alpha)$$

($\pi(s|A_i = a; \alpha)$ is probability of a given allocation, depending on covariates)

Individual Average Potential Outcomes

- Let α denote a cluster-level propensity of intervention
- Define **individual average potential outcome** over all possible allocations described by α

$$\bar{Y}_j(A_i = a; \alpha) = \sum_{\mathbf{s} \in \mathcal{A}(|P^k|-1)} Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{s}) \pi(s | A_i = a; \alpha)$$

($\pi(s|A_i = a; \alpha)$ is probability of a given allocation, depending on covariates)

For every location, average potential outcomes over all possible treatment allocations where:

- Power plant *i* gets treatment *a*
- Cluster-level treatment propensity is α

"Key-Associated" Causal Effects

In the bipartite setting

- $\bar{Y}_j(A_i = a; \alpha)$ is defined for every (i, j) pair within the cluster

"Key-Associated" Causal Effects

In the bipartite setting

- $\bar{Y}_j(A_i = a; \alpha)$ is defined for every (i, j) pair within the cluster
- Consider for each location, a particular power plant of interest
  - E.g., the closest plant to location j
  - Denote with $p^*_{i(j)}$
- Define estimands based on $(p^*_{i(j)}, m_j)$
- Only consider one $\bar{Y}_j(A_{i^*} = a; \alpha)$ for every j in the cluster

Average Key-Associated Direct/Indirect Effects

Cluster-Average Potential Outcome:

$$\bar{Y}^k(a_{i*}, \alpha) = \frac{1}{|M^k|} \sum_{j \in M^k} \bar{Y}_j(A_{i*} = a, \alpha)$$

Average Key-Associated Direct/Indirect Effects

Cluster-Average Potential Outcome:

$$\bar{Y}^k(a_{i*}, \alpha) = \frac{1}{|M^k|} \sum_{j \in M^k} \bar{Y}_j(A_{i*} = a, \alpha)$$

Cluster-Average Direct Effect

$$\bar{Y}^k(A_{i*} = 1, \alpha) - \bar{Y}^k(A_{i*} = 0, \alpha) \equiv DE^{k*}(\alpha)$$

Cluster-Average Indirect Effect

$$\bar{Y}^k(A_{i*} = a, \alpha) - \bar{Y}^k(A_{i*} = a, \alpha') \equiv IE^{k*a}(\alpha, \alpha')$$

Population-Average Direct Effect

$$DE^*(\alpha) = \frac{1}{K} \sum_k DE^{k*}(\alpha)$$

Population-Average Indirect Effect

$$IE^{a*}(\alpha, \alpha') = \frac{1}{K} \sum_k IE^{k*a}(\alpha, \alpha')$$

IPTW Estimator

Tchetgen Tchetgen and Vanderweele (2012)
Papadogeorgou et al. (2018)

"Familiar" IPTW estimator

$$\widehat{Y}^k(A_{i^*} = a; \alpha) = \frac{1}{|M^k|} \sum_{j \in M^k} \frac{\pi(\mathbf{A}^k_{(-i^*)} | A_{i^*} = a, \alpha)}{f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k | \mathbf{W}^k, \mathbf{X}^k)} I(A_{i^*} = a) Y_j$$

- $f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k | \mathbf{W}^k, \mathbf{X}^k)$ is the cluster-level propensity score
- $\pi(\mathbf{A}^k_{(-i^*)} | A_{i^*} = a, \alpha)$ specifies probabilities of realistic counterfactual treatment allocations

IPTW Estimator

Tchetgen Tchetgen and Vanderweele (2012)
Papadogeorgou et al. (2018)

"Familiar" IPTW estimator

$$\widehat{Y}^k(A_{i^*} = a; \alpha) = \frac{1}{|M^k|} \sum_{j \in M^k} \frac{\pi(\mathbf{A}^k_{(-i^*)} | A_{i^*} = a, \alpha)}{f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k | \mathbf{W}^k, \mathbf{X}^k)} I(A_{i^*} = a) Y_j$$

- $f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k | \mathbf{W}^k, \mathbf{X}^k)$ is the cluster-level propensity score
- $\pi(\mathbf{A}^k_{(-i^*)} | A_{i^*} = a, \alpha)$ specifies probabilities of realistic counterfactual treatment allocations

Corresponding estimator for the population-average potential outcome:

$$\widehat{Y}(A_{i^*} = a; \alpha) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Y}^k(A_{i^*} = a, \alpha)$$

Properties of Estimator

Following work by Tchetgen Tchetgen and VanderWeele (2012)
Liu et al. (2016), Papadogeorgou et al. (2018)

$\widehat{Y}(A_{i*} = a, \alpha)$ is consistent for the population-average potential outcome and asymptotically normal if:

Properties of Estimator

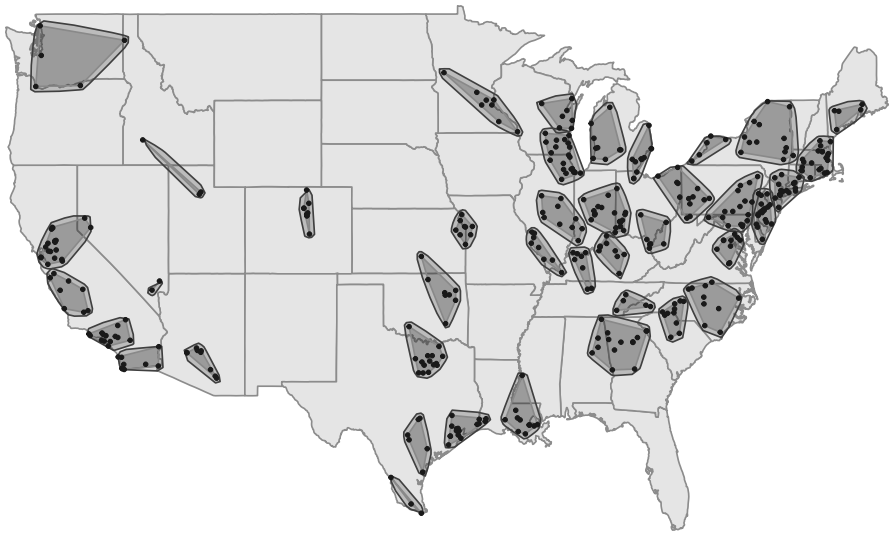
Following work by Tchetgen Tchetgen and VanderWeele (2012)
Liu et al. (2016), Papadogeorgou et al. (2018)

$\widehat{Y}(A_{i*} = a, \alpha)$ is consistent for the population-average potential outcome and asymptotically normal if:

1. *Positivity*: Probability of observing cluster treatment vector $\mathbf{A}^k = \boldsymbol{a}^k$, $f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k = \boldsymbol{a}^k|\mathbf{W}^k, \mathbf{X}^k)$, is positive for all $\boldsymbol{a}^k \in \mathcal{A}(|P^k|)$
2. *Ignorability*: Observed cluster treatment $\mathbf{A}^k$ is conditionally independent of cluster potential outcomes $\mathbf{Y}^k(\cdot)$ given the cluster covariates $\mathbf{W}^k, \mathbf{X}^k$: $\mathbf{A}^k \amalg \mathbf{Y}^k(\cdot)|\mathbf{W}^k, \mathbf{X}^k$

Bipartite Causal Inference with Interference

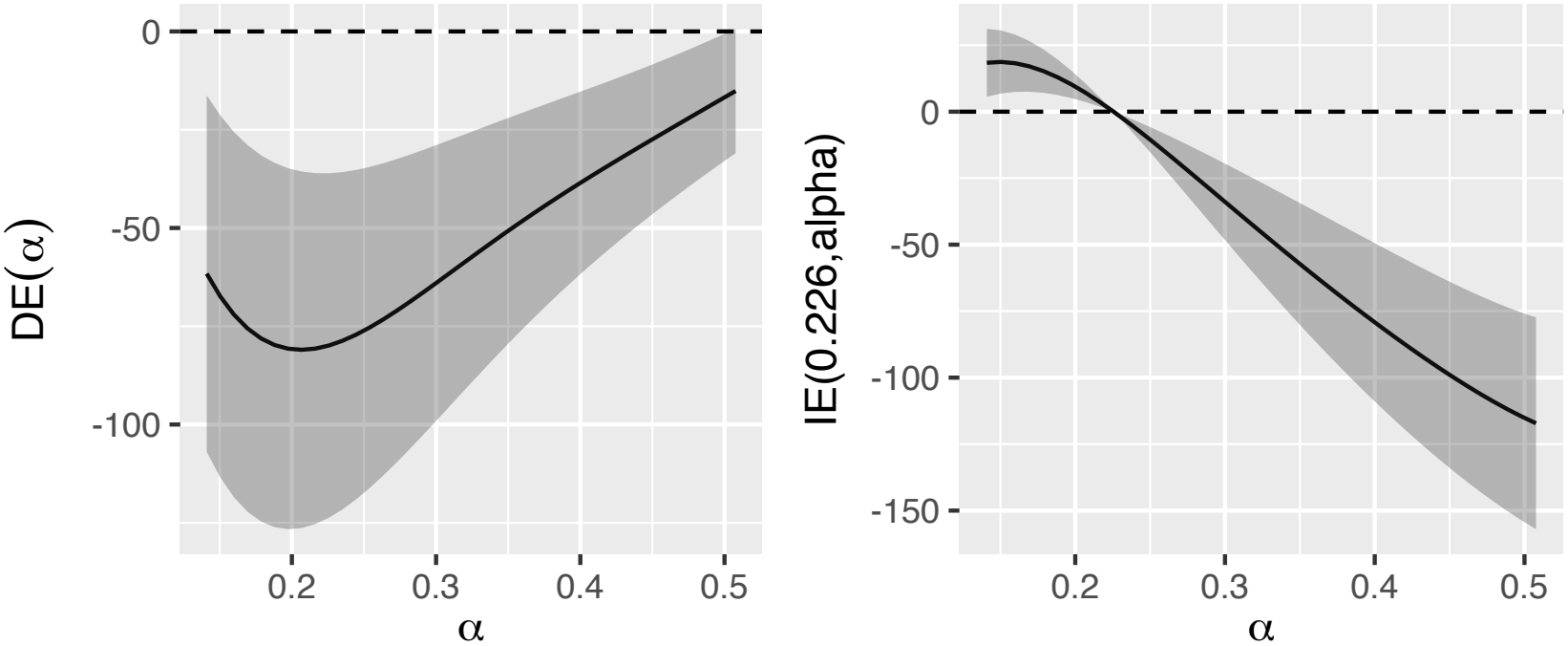
Effects of SCR on Medicare IHD Hospitalizations



- $P = 473$ power plants
  - Install (or not) (SCR) system to reduce NO_x emissions
  - Clustered into 40 groups
- $M = 21,350$ zip codes
  - Medicare CVD hospitalizations
  - And other confounders (weather, demographcis, etc.)

1. What is the average direct effect of installing SCR on the closest power plant?
2. What is the average indirect effect of increasing SCR installation on “upwind” plants in the same cluster?

Estimated Effects of SCR on Medicare CVD Hospitalizations



- For a fixed "upwind" SCR propensity (α), installing SCR on the closest plant decrease IHD hospitalizations. This effect wanes as there are more SCRs "upwind."
- Installing more SCRs "upwind" reduces IHD hospitalization when the closest plant does not have SCR

Generalized Propensity Score Estimation for Bipartite Causal Inference with Interference

An Alternative to the Clustering/IPTW Approach

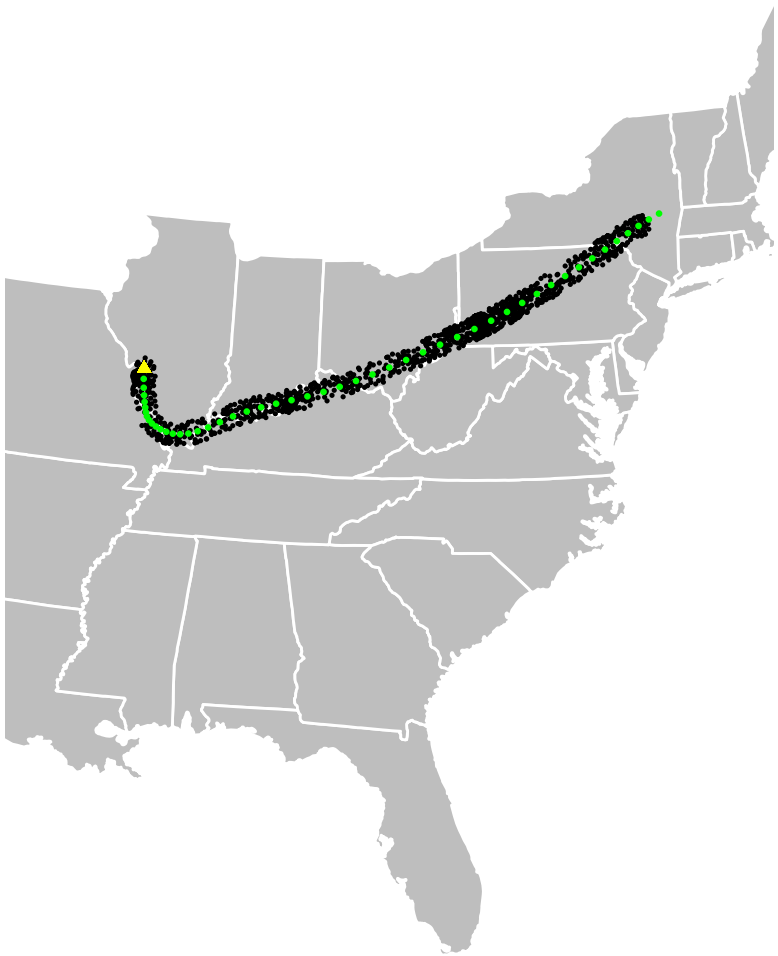
Instead of reducing the number of potential outcomes by clustering...

An Alternative to the Clustering/IPTW Approach

Instead of reducing the number of potential outcomes by clustering...

- Formulate a “bivariate treatment”
  1. “Individual”-level treatment of the closest power plant
  2. “Neighborhood”-level treatment of the “upwind” power plants
- Make smarter use of atmospheric science of air pollution
- Define slightly different causal estimands
- Different estimation strategy
- Apply to investigate the effects of scrubbers on IHD hospitalization

Atmospheric Modeling of Pollution Transport



- HYSPLIT air trajectory/dispersion
  - Simulate movement of air parcel through space
  - Based on historical weather data
- Align particles in time/space with zip codes $\rightarrow$ link to health outcomes (and other data)

Atmospheric Data Science Made Easy and Scalable

1. Simulate 100Ks of HYSPLIT dispersions

2. Link trajectories 10Ks population locations

Atmospheric Data Science Made Easy and Scalable

1. Simulate 100Ks of HYSPLIT dispersions
   - Previously web-based implementation hosted by NOAA
     - Hard to use (at least for us) parallel implementation
   - Incorporated parallelization into existing R package `splitR`
   - 100Ks trajectories in $\approx$ 3 hours

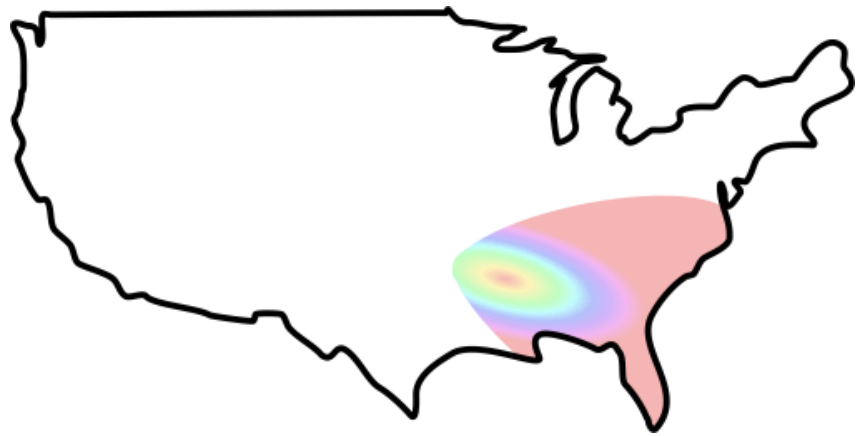
2. Link trajectories 10Ks population locations
   - PostGIS/PostgreSQL
     - Open-source, industry standard
   - Link 100Ks trajectories to 10Ks zip codes in $\approx$ 2 hours on personal machine

All developed with “front-end” R interface

HYSPLIT Average Dispersion (HyADS)

Use **HYSPLIT** to model dispersion of emissions $\rightarrow$ populations

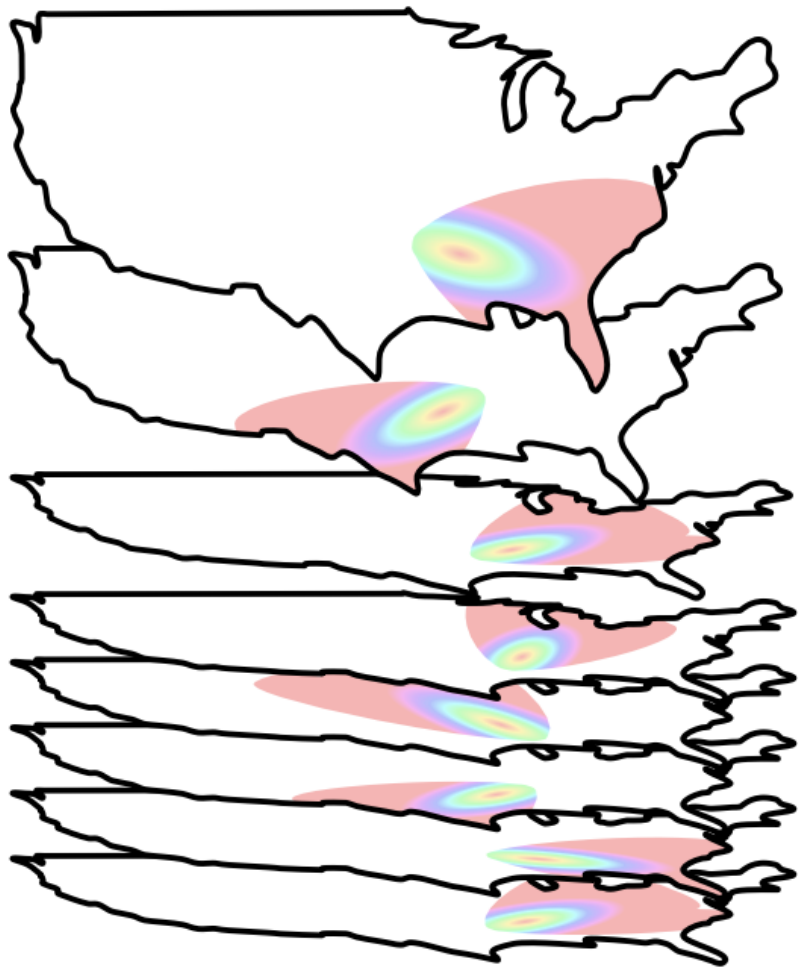
- Simulate dispersion of 100 parcels starting at individual stack
- Parcels tracked for 10 days and locations aggregated to ZIP codes
- Repeat at 6 hour intervals daily
- $\Rightarrow a_{ij}$ to denote "influence" of p_i on m_j



HYSPLIT Average Dispersion (HyADS)

Use **HYSPLIT** to model dispersion of emissions $\rightarrow$ populations

- Simulate dispersion of 100 parcels starting at individual stack
- Parcels tracked for 10 days and locations aggregated to ZIP codes
- Repeat at 6 hour intervals daily
- $\Rightarrow a_{ij}$ to denote "influence" of p_i on m_j



"Individual" and "Neighborhood" Treatment

- For $j = 1, 2, \ldots, M$ =26,149 zip codes define:
  1. "Individual" Treatment: Scrubber status of the closest power plant: A_{i^*}
     - i^* indexes the power plant closest to zip code j

"Individual" and "Neighborhood" Treatment

- For $j = 1, 2, \ldots, M$ =26,149 zip codes define:
  1. "Individual" Treatment: Scrubber status of the closest power plant: A_{i^*}
     - i^* indexes the power plant closest to zip code j
  2. "Neighborhood" Treatment: A function of the scrubber statuses of all other "linked" power plants
     - $G_j = \sum_{i \neq i^*} a_{ji} A_i$
     - HYSPLIT-weighted treatment rate
     - "Neighborhood" $\rightarrow$ "upwind"

"Individual" and "Neighborhood" Treatment

- For $j = 1, 2, \ldots, M$ =26,149 zip codes define:
  1. "Individual" Treatment: Scrubber status of the closest power plant: A_{i^*}
     - i^* indexes the power plant closest to zip code j
  2. "Neighborhood" Treatment: A function of the scrubber statuses of all other "linked" power plants
     - $G_j = \sum_{i \neq i^*} a_{ji} A_i$
     - HYSPLIT-weighted treatment rate
     - "Neighborhood" $\rightarrow$ "upwind"

- **Consequence:** $Y_j(A_1, \ldots A_P) \rightarrow Y_j(A_{i^*}, G_j)$
  - Hospitalizations are the same under different **A** if the treatment of the closest power plant and the "upwind" treatment rate are the same
  - Formalized via adaptation of SUTVA

"Direct" Effects in the Bipartite Setting

No self-evident meaning of treatment applying "directly" to a unit

The "direct" effect of treating the closest plant while holding fixed the "upwind" treatments:

$$\tau(g) = E\left[\, Y_j(A_{i*} = 1, G_j = g) - Y_i(A_{i*} = 0, G_j = g) \,\right]$$

"Overall" average over the distribution of "upwind" treatments:

$$\tau = \sum_{g \in \mathcal{G}} \tau(g) P(G_j = g)$$

"The average effect on IHD hospitalizations of installing a scrubber on the closest power plant"

"Indirect" or "Spillover" Effects in the Bipartite Setting

The "indirect" effect of installing scrubbers on "upwind" power plants without changing the closest plant:

$$\delta(g; a) = E\left[\, Y_j(A_{i*} = a, G_j = g) - Y_j(A_{i*} = a, G_j = 0) \,\right]$$

"Overall" average over the distribution of "upwind" treatments:

$$\Delta(a) = \sum_{g \in \mathcal{G}} \delta(g; a) P(G_j = g)$$

"The average effect on IHD hospitalizations of installing more scrubbers on 'upwind' plants"

Individual and Neighborhood Propensity Scores

Extension of Forastiere et al. (2016)

View "individual" and "upwind" treatments as bivariate treatment and estimate a **joint propensity score**:

$$\psi(a; g; x) = P(A_{i*} = a, G_j = g | \mathbf{X}_j = x)$$

Further decompose into:

$$\begin{aligned} \psi(a; g; x) &= P(A_{i*} = a, G_j = g | \mathbf{X}_j = x) && (1) \\ &= P(G_j = g | A_{i*} = a, \mathbf{X} = x) P(A_{i*} = a | \mathbf{X}_{i*} = x) && (2) \end{aligned}$$

$(1) \equiv \phi(a, x) \equiv$ "individual propensity sore"

$(2) \equiv \lambda(g; z; x) \equiv$ "upwind propensity score"

Outline of Estimation Strategy

Extension of Forastiere et al. (2016)

1. Estimate the individual propensity score for M =26,149 zip codes
2. Sub-classify zip codes into 5 subclasses
3. Within each subclass:
   (a) Estimate the "upwind" propensity score, G_j
   (b) Estimate a Poisson regression model for $Y_j|G_j$
   (c) Predict unobserved potential outcomes over a grid of values for G_j
4. Estimate an average dose-response function for the relationship between $Y(A_{i*}, G_j = g)$ and values of $G_j = g$
5. Calculate estimates of "direct" and "upwind" effects
6. Calculate standard errors and confidence intervals using bootstrap

Other variations on this theme . . .

Bipartite Causal Inference with Interference

Analysis of Scrubbers on Medicare IHD Hospitalizations

Estimated Effects of Scrubbers IHD Hospitalization

Very preliminary results

- Average "direct" effect of installing a scrubber on the closest power plant (τ)
  - -18.10 (-26.91, -7.33) hospitalizations per 10K person-years

Estimated Effects of Scrubbers IHD Hospitalization

Very preliminary results

- Average "direct" effect of installing a scrubber on the closest power plant (τ)
  - -18.10 (-26.91, -7.33) hospitalizations per 10K person-years
- Average "indirect" effect of installing more scrubbers on "upwind" plants ($\Delta(a)$):
  - If the closest plant has a scrubber: -35.46 (-59.94, -19.30)
  - If the closest plant does not have a scrubber: -32.93 (-56.31, -16.87)

Estimated Effects of Scrubbers IHD Hospitalization

Very preliminary results

- Average "direct" effect of installing a scrubber on the closest power plant (τ)
  - -18.10 (-26.91, -7.33) hospitalizations per 10K person-years
- Average "indirect" effect of installing more scrubbers on "upwind" plants ($\Delta(a)$):
  - If the closest plant has a scrubber: -35.46 (-59.94, -19.30)
  - If the closest plant does not have a scrubber: -32.93 (-56.31, -16.87)

Conclusion: Scrubbers reduce IHD hospitalizations, with more gains coming from reduced pollution transport from "upwind" plants than from nearby plants.

Bipartite Causal Inference with Interference

Summary and Conclusions

Summary

- Causal inference with spatial data my often confront *interference* and *spillover*
- Introduce new types of causal questions
- Budding research in studies on networks
- Not as much focus on explicitly *spatial* data
- Clear role in the study of air quality interventions
- Role in many studies of interventions that “spread” spatially
- Lots of work to do!

Thank You!

Key contributors to this work:

- Biostatistics
  - Laura Forastiere (Yale/Univ. of Florence)
  - Fabrizia Mealli (Univ. of Florence)
  - Georgia Papadogeorgou (HSPH)
  - Chanmin Kim (BU)
  - Kevin Cummiskey (HSPH)
- Environmental Engineering/Atmospheric Science
  - Lucas Henneman (HSPH)
- Data Science
  - Christine Choirat (HSPH)
- Funding: NIH R01ES026217, USEPA RD-83479801

This work's contents contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M. Zigler. Causal inference for interfering units with cluster and population level treatment allocation programs. arXiv:1711.01280 [math, stat], November 2017. arXiv: 1711.01280.

L. Liu, M. G. Hudgens, and S. Becker-Dreps. On inverse probability-weighted estimators in the presence of interference. Biometrika, 103(4):829-842, December 2016.

Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On causal inference in the presence of interference. Statistical Methods in Medical Research, 21(1):55-75, February 2012.

Carolina Perez-Heydrich, Michael G. Hudgens, M. Elizabeth Halloran, John D. Clemens, Mohammad Ali, and Michael E. Emch. Assessing effects of cholera vaccination in the presence of interference. Biometrics, May 2014.

Daniel J. Graham, Emma J. McCoy, and David A. Stephens. Quantifying the effect of area deprivation on child pedestrian casualties by using longitudinal mixed models to adjust for confounding, interference and spatial dependence. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(4):931- 950, 2013.

Bipartite Causal Inference with Interference

Extra Slides

Linking Data Sources

One example data set

Power Plants*

- $P = 409$ Coal-fired power plants
- Daily pollution emissions (e.g., SO_2)
- Plant features (size, operating capacity, stack features, etc.)
- Whether each had a **scrubber** installed

*Data from EPA Air Markets Division and Energy Information Administration

Linking Data Sources

One example data set

Power Plants*

- $P = 409$ Coal-fired power plants
- Daily pollution emissions (e.g., SO_2)
- Plant features (size, operating capacity, stack features, etc.)
- Whether each had a **scrubber** installed

*Data from EPA Air Markets Division and Energy Information Administration

Populations+

- $M = 26,149$ zip codes in Eastern US
- Hospitalization records on $\approx$ 19million Medicare beneficiaries
- Demographics
- Weather/climate
- Smoking rates

+ Data from CMS, US Census, NCDC, CDC

Linking Data Sources

One example data set

Power Plants*

- $P = 409$ Coal-fired power plants
- Daily pollution emissions (e.g., SO_2)
- Plant features (size, operating capacity, stack features, etc.)
- Whether each had a **scrubber** installed

*Data from EPA Air Markets Division and Energy Information Administration

Atmospheric Science Tools

$\Longleftrightarrow$

Knowledge of pollution transport to link pollution **sources** to exposed **populations**

$\Longleftrightarrow$

Populations+

- $M = 26,149$ zip codes in Eastern US
- Hospitalization records on $\approx$ 19million Medicare beneficiaries
- Demographics
- Weather/climate
- Smoking rates

+ Data from CMS, US Census, NCDC, CDC

Common Simplifications

1. Assume each $m_j \in \mathcal{M}$ adopts treatment status of exactly one $p_i \in \mathcal{P} \rightarrow$ Proceed as though $\mathcal{M}$ were the only units
   - E.g., link each student to his/her teacher; link each zip code to its closest power plant
2. Aggregate measures on $\mathcal{M}$ to the level of $\mathcal{P} \rightarrow$ Proceed as though $\mathcal{P}$ were the only units
   - E.g., calculate the average student outcomes for a teacher; calculate average hospitalization rate for zip codes near a power plant
3. Re-define the intervention to apply to $\mathcal{M}$
   - E.g., define a zip-code level treatment as a function of treatment status of nearby power plants
   - Similar to an "exposure mapping" in Aronow and Samii (2015)

Many Types of Average Effects

1. $\mathcal{P}$-indexed effects
   - Effects of treatments on $p_i \in \mathcal{P}$, averaged over locations $m_j \in \mathcal{M}$ that contain p_i in their interference set
   - E.g., the "typical" effect across all connected locations of treating a single power plant
2. $\mathcal{M}$-indexed effects
   - Effects on $m_j \in \mathcal{M}$, averaged over treatments at $p_i \in \mathcal{P}$ in the interference set
   - E.g., the "typical" effect on a single location of treating a single power plant
3. $\mathcal{M}^*$-indexed effects
   - Do not average over all (i, j)
   - Effects on $m_j \in \mathcal{M}$, defined according to a single i^* in the interference set
   - E.g., the "typical" effect of treating the closest power plant, accounting for connectedness

$\mathcal{P}$-indexed Average Direct Effects

$$DE_i(\alpha) = \frac{1}{n_{t_i'}} \sum_{j \in t_i'} DE_{(i,j)}(\alpha)$$

average effect of treating unit p_i, averaged across all m_j having p_i in their interference set.

- Assumes allocation strategy for each $m_j \in t_i'$ is held fixed at α.
- Average DE: $\bar{DE}_{\mathcal{P}}(\alpha) = \frac{1}{P} \sum_i DE_i(\alpha)$
  - Average effect of treating a single power plant while holding allocation at "connected" monitors at α.
  - The "typical" effect of treating a single power plant, accounting for connectedness
  - Does this really make sense to hold α fixed at all $m_j \in t_i'$?
    - Maybe only in clustered/partial interference setting

$\mathcal{P}$-indexed Average Indirect Effects

$$IE_i^a(\alpha, \alpha') = \frac{1}{n_{t_i'}} \sum_{j \in t_j'} IE_{(i,j)}^a(\alpha, \alpha')$$

average effect of holding treatment at p_i fixed but changing allocations to all m_j having p_i in their interference set

- Average IE: $\bar{IE}_{\mathcal{P}}^a(\alpha, \alpha') = \frac{1}{P} \sum_i IE_i^a(\alpha, \alpha')$
  - Average effect of changing allocations among all "connected" monitors while holding a single power plant fixed
  - The typical effect of changing treatments to other plants
- Changing allocation to all m_j at the same time?
- Make sense/useful without clustering?

$\mathcal{M}$-indexed Average Direct Effects

$$DE_j(\alpha) = \frac{1}{n_t} \sum_{i \in t_j} DE_{(i,j)}(\alpha)$$

Average effect of treating a single power plant in the interference mapping for m_j

- Average DE: $\bar{DE}_{\mathcal{M}} = \frac{1}{M} \sum DE_j(\alpha)$
- Typical effect on a location of treating a single power plant in the mapping
- Not useful

$\mathcal{M}$-indexed Average Indirect Effects

$$IE^a_j(\alpha, \alpha') = \frac{1}{n_t} \sum_{i \in t_j} IE^a_{(i,j)}(\alpha, \alpha')$$

- $\bar{IE}^a_{\mathcal{M}}(\alpha, \alpha') = \frac{1}{M} \sum IE^a_j(\alpha, \alpha')$
- Typical effect of changing allocation while holding a single plant fixed, accounting for connectedness
- Not useful