Statistical Science
2021, Vol. 36, No. 1, 109–123
https://doi.org/10.1214/19-STS749

Bipartite Causal Inference with Interference

Corwin M. Zigler and Georgia Papadogeorgou

Abstract. Statistical methods to evaluate the effectiveness of interventions are increasingly challenged by the inherent interconnectedness of units. Specifically, a recent flurry of methods research has addressed the problem of *interference* between observations, which arises when one observational unit's outcome depends not only on its treatment but also the treatment assigned to other units. We introduce the setting of *bipartite causal inference with interference*, which arises when (1) treatments are defined on observational units that are distinct from those at which outcomes are measured and (2) there is *interference* between units in the sense that outcomes for some units depend on the treatments assigned to many other units. The focus of this work is to formulate definitions and several possible causal estimands for this setting, highlighting similarities and differences with more commonly considered settings of causal inference with interference. Toward an empirical illustration, an inverse probability of treatment weighted estimator is adapted from existing literature to estimate a subset of simplified, but interesting, estimands. The estimators are deployed to evaluate how interventions to reduce air pollution from 473 power plants in the U.S. causally affect cardiovascular hospitalization among Medicare beneficiaries residing at 18,807 zip code locations.

Key words and phrases: Air pollution, causal inference, interference, network dependence, power plants.

1. INTRODUCTION

Consider evaluating the causal effect of an intervention in a context with the following features: (1) the intervention is defined and measured on one type of observational unit, but (2) outcomes of interest are defined and measured on a second, distinct type of unit. Common examples include educational interventions applied to teachers with outcomes of interest defined on students, social interventions applied at neighborhoods with outcomes defined at the level of the resident, or as will be the focus of the present discussion, interventions applied at sources of air pollution (e.g., power plants) and health outcomes measured among people at specific population locations (e.g., zip codes). We refer to such a setting as one of *bipartite causal inference*, reminiscent of the two types of nodes in a bipartite graph. Such bipartite structures are commonplace in many fields where interest lies in evaluating the causal effects of an intervention.

Consider a setting of bipartite causal inference augmented with the complexity that interconnectedness among the two types of units gives rise to what has been termed in the causal inference literature *interference*, where outcomes for a particular unit depend upon treatments assigned to (possibly many) other units. We term the combination of these two features as the setting of *bipartite causal inference with interference*, which has not, to our knowledge, been previously considered.

Most existing work on causal inference with interference is formalized in the familiar setting with one level of observational unit [7, 8, 22, 9, 20, 10, 23, 26, 24, 15, 1, 4, 2, 13, 16, 19, 14]. The most well-studied examples are studies of infectious diseases where vaccinating a person will also reduce the infection risk of others who come into contact with that person [8, 10, 13, 16, 19] and the analysis of social networks where interventions can affect a unit directly and also through impact on an individuals' peers. Various estimands have been introduced to describe the effect on a particular unit's outcome due to treatments applied to other units, with terminology including *indirect effects*, *spillover effects*, *contamination*

Corwin M. Zigler is Associate Professor of Statistics and Data Sciences, Dell Medical School, University of Texas at Austin, 2317 Speedway D9800, Austin, Texas 78712-1823, USA (e-mail: cory.zigler@austin.utexas.edu). Georgia Papadogeorgou is Postdoctoral Associate, Department of Statistical Science, Duke University, 206 Old Chem Bldg, Durham, North Carolina 27708, USA (e-mail: gp118@duke.edu).

effects and *peer effects*, but the common theme is that interference typically arises because unit-to-unit interactions lead outcomes of some to depend on outcomes (and, by extension, treatments) of others. Methods for estimation and inference in such settings have considered both randomized and observational settings, with emphasis on settings of so-called *partial interference* that leverage assumptions of interference within, but not between, distinct and nonoverlapping clusters of units [22, 9, 10, 23, 13, 19, 11].

Similar formalization of interference problems in the bipartite setting presents challenges that have not been previously considered. One reason is the required technical distinctions relating to the two types of observational unit; defining estimands and corresponding estimators requires maintenance of the distinction between units where interventions occur and those where outcomes are measured. What's more, settings of bipartite causal inference with interference likely arise due to underlying scientific phenomena that cannot be described by the type of unit-to-unit outcome dependencies common to the study of infectious diseases or social networks. In the bipartite setting, interference is more likely a consequence of complex exposure dependencies that describe how the impact of a particular treatment propagates across units. Settings of interference due to complex exposure dependencies have been considered in the setting of one observational unit, albeit with much less focus than settings of unit-to-unit outcome dependencies [20, 24, 6].

The goal of this paper is to formalize the development of potential-outcomes methods relevant to settings of bipartite causal inference with interference. We define potential outcomes in this setting and introduce *interference mappings* describing the network of interconnectedness between units. From here, we formalize alternatives to the commonly-invoked stable unit treatment value assumption and propose several causal estimands unique to the bipartite setting. The discussion of estimands is intentionally general in order to introduce new types of causal quantities that could potentially be of interest in the bipartite setting. Toward a simple empirical illustration, we invoke several simplifying assumptions, including a bipartite version of partial interference, to focus on a subset of relevant estimands for which corresponding estimators can be derived from existing inverse probability weighted estimators. Throughout, we highlight similarities and differences with existing estimands and methods for causal inference with interference in settings with one level of observational unit.

For illustration, we frame the discussion in the context of evaluating interventions designed to reduce pollution-related health burden by limiting harmful emissions from power plants in the U.S.. The features defining the bipartite structure are that interventions are defined and implemented at the level of the power plant, but key questions for regulatory policy pertain to health outcomes (e.g., cardiovascular hospitalizations) measured at population locations across the country. Unlike in most existing literature on causal inference with interference, the interference in the power plant case is not due to dependent outcomes among locations or people (e.g., one person's hospitalization does not affect another person's risk). Rather, interference in this case is due to the nature of pollution exposure, which derives from complex processes that render an individual location subject to actions at many power plants and many power plants impacting common sets of locations.

Ultimately, the development in this paper is designed as a framework for addressing problems and data structures that have not been previously considered alongside the formalization of causal inference with interference. Explicitly targeting the complexities of interference due to air pollution transport presents the first step toward statistical tools for evaluating air quality control policies that have to date relied on deterministic physical-chemical air quality models that are not validated with observed data.

2. MOTIVATING SETTING: POWER PLANT REGULATORY POLICIES

Various compounds emitted from power plants undergo complex chemical and physical processes to form harmful air pollution that is transported across space. This phenomenon is known as *pollution transport*. In light of this phenomenon, existing regulatory assessments use deterministic models of pollution transport to simulate regulatory impacts. From a statistical perspective, the phenomenon of pollution transport manifests as interference between units, since outcomes at one location are dependent on treatments at many pollution sources located "upwind" (although note that pollution transport is generally more complex than just the direction of the wind). Development of new methods for interference can enhance current regulatory assessments by combining rigorous statistical methodology with state-of-the-art knowledge of pollution transport.

For example, consider a specific intervention that may or may not be implemented at a power plant, namely, the installation of selective catalytic reduction or selective noncatalytic reduction (SnCR) system, a technology known to reduce emissions of nitrous oxides (NO_x), important precursors to the formation of various types of air pollution known to be associated with adverse health outcomes [17, 3]. We aim to characterize the extent to which installation of such a SnCR system causally impacts hospitalization rates for cardiovascular disease (CVD) among Medicare beneficiaries. This setting fits the description of bipartite causal inference with interference because: (1) SnCR systems are installed (or not) at individual power plants; (2) CVD hospitalizations are measured at

zip codes; (3) CVD hospitalizations at a given zip code depend on the constellation of SnCR systems installed at many upwind power plants and; (4) a given power plant may impact the CVD hospitalizations at multiple zip codes.

3. POTENTIAL OUTCOMES FOR BIPARTITE CAUSAL INFERENCE WITH INTERFERENCE

The defining feature of the bipartite structure is the presence of two distinct types of observational units. First, define the set of *interventional units*, $\mathcal{P} = \{p_1, p_2, \ldots, p_P\}$ to be the available observational units upon which interventions either occur or not. In the motivating example, $\mathcal{P}$ is a set of $P = 473$ power plants located across the U.S. For each $p_i \in \mathcal{P}$, let $A_i = 1, 0$ denote the presence, absence of an intervention at the ith interventional unit, for example, an indicator of whether a power plant installs a SnCR system. Let $\mathbf{A} = (A_1, A_2, \ldots, A_P)$ denote a vector of possible treatment assignments to each of the interventional units in $\mathcal{P}$, with $\mathbf{a} \in \mathcal{A}(P)$ representing one vector of 2^P possible treatment allocations. Denote covariates measured at the level of the interventional units with $\mathbf{W}_i$, for $i = 1, 2, \ldots, P$.

Let $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$, denote a set of M units of a second type, termed *outcome units*. In the motivating example, $\mathcal{M}$ consists of $M = 18{,}807$ zip codes located across the U.S. Let Y_j denote a measured outcome at each of the $j = 1, 2, \ldots, M$ outcome units, for example, the number of CVD hospitalizations among Medicare beneficiaries residing at zip code m_j. Similarly, $\mathbf{X}_j$ could denote covariates measured at the outcome units, for example, zip code level population demographics. The salient feature of the bipartite structure is that, without further restrictions or assumptions, interventions are not defined on the outcome units (e.g., a zip code cannot be "treated" with a SnCR system), yet outcomes of interest are not defined on the interventional units (e.g., a power plant does not have a hospitalization count).

Defining potential outcomes for the bipartite setting is notationally analogous to settings of one level of observational unit. Let $Y_j(\mathbf{a})$ denote the potential outcome that would be observed at outcome unit m_j under treatment allocation $\mathbf{a}$, for example, the number of CVD hospitalizations that would occur at the jth zip code under a specific allocation of SnCR systems on power plants. In the most general setting, a unique $Y_j(\mathbf{a})$ is defined for every possible $\mathbf{a} \in \mathcal{A}(P)$. The unique feature of these definitions in the bipartite setting is that $Y_j(\mathbf{a})$ are defined for $j = 1, 2, \ldots, M$, but $\mathbf{a}$ is a vector of length P.

3.1 Distinction with Clustered Experiments and Common Simplifications to Bipartite Structures

Many conventional instances of data structures with different levels of observational unit and interference can be appropriately cast in terms of (approximate) clustered experiments, where interventions are applied directly to outcome units, with a second type of "clustering unit" available to hierarchically cluster outcome units. One classical example appears in [9], where interventions and outcomes defined on students are available within clusters defined by schools. The key distinctions between this clustering setting and the bipartite setting considered here are: (1) distinct interventional and outcome units; (2) the possibility of a clustering unit that is distinct from the interventional unit and (3) the possibility that each cluster may consist of *more than one* interventional unit. One example of bipartite causal inference with interference within a clustered experiment would be investigation of a teacher intervention (interventional units) for its effect on student outcomes (outcome units) with both students and teachers clustered within schools (clustering units). A similar such clustered instance of bipartite causal inference with interference will be motivated in the power plant example in Section 6.

Furthermore, many bipartite settings permit simplification of the bipartite structure of the data by projecting onto the space of one type of observational unit. Projecting to the space of $\mathcal{M}$ could follow from linking each $m_j \in \mathcal{M}$ to exactly one $p_i \in \mathcal{P}$ by, for example, assuming that each m_j adopts the treatment status of the closest p_i. Such a reduction would extend the definition of the treatment (originally defined on $\mathcal{P}$) to the level of $\mathcal{M}$, and subsequent development could proceed as though $\mathcal{M}$ were the only observational units. A similar projection to the space of $\mathcal{P}$ could follow by aggregating measures originally defined at the level of $\mathcal{M}$. For example, one could consider the CVD hospitalizations among all zip codes within a certain distance of each $p_i \in \mathcal{P}$, and proceed as though $\mathcal{P}$ were the only observational units [17, 18, 12]. Such simplifications might be appropriate in settings for which it is self-evident which single interventional unit corresponds to a given outcome unit, such as the students-within-schools example where observations are hierarchically clustered.

Other simplifications to the bipartite structure could follow from changing the definition of the intervention (and the subsequent question of interest). For example, the intervention could be redefined to pertain to each m_j as some function of the interventions on $\mathcal{P}$. One such possibility in the power plant example would be defining a zip-code level treatment as a function of the treatment statuses of several power plants, such as the proportion of upwind power plants that installed a SnCR system. This would be similar to a so-called "exposure mapping," [2] which in this case would transform the goal of estimating causal effects of the intervention inherently defined at the level of $\mathcal{P}$ to estimating effects of the new redefined treatment at the level of $\mathcal{M}$, which may not correspond to any practicable intervention.

The development herein is designed to formulate causal estimands when no such simplification is appropriate, as in the power plant example where assigning each zip code to one power plant (or vice versa) would be too simplistic in light of the realities of air pollution transport and interest lies in the effects of specific interventions on individual power plants.

3.2 Extending to the Bipartite Setting: Interference Mappings and Structured SUTVA

Formalizing potential outcomes and causal estimands in the bipartite setting requires a reformulation of common assumptions about potential outcomes. In particular, we consider settings that constitute interference in the sense that the stable unit treatment value assumption (SUTVA), which is typically formalized to state that a unit's potential outcome depends only on that unit's treatment, no longer holds [21]. Toward this goal, we cast the bipartite data structure as a network with two different types of nodes, $p_i \in \mathcal{P}$ and $m_j \in \mathcal{M}$, where edges between p_i and m_j denote that interventions applied at p_i have some bearing on outcomes measured at m_j. We use the term *interference mapping* to denote such a network structure. In the power plant setting, this structure is governed by atmospheric and climatological conditions that transport power plant emissions across space as they transform into population pollution exposure.

For each outcome unit, let the *interference set* be the set of interventional units for which the presence or absence of the intervention may affect outcomes [14], a notion that will be made formal with a reformulated statement of SUTVA. Let $t_j^\top = (t_{j1}, t_{j2}, \ldots, t_{jP})$, where $t_{ji} = 1(0)$ if p_i is in the interference set for m_j. Define the *interference mapping* as $T = (t_1, t_2, \ldots, t_M)^\top$, where T is a $M \times P$ matrix denoting the interference sets for all $m_j \in \mathcal{M}$. This definition of T essentially amounts to what is often considered as an "adjacency matrix," even though the entries of T in this case can encode more complex relationships between the p_i, m_j than spatial adjacency. For notational simplicity, we will let $p_i \in T_j$ denote all i such that $t_{ji} = 1$ and use this to refer to all interventional units in the interference set for a given m_j. In the power plant example, the set of $p_i \in T_j$ can be thought of as the set of power plants that are "upwind" from the jth location, and we will refer to it as such. Similarly, we will let $m_j \in T_i^\top$ denote all j such that $t_{ji} = 1$ and use this to refer to all outcome units that contain p_i in their interference set. In the power plant example, this can be thought of as all locations that are "downwind" from the ith power plant.

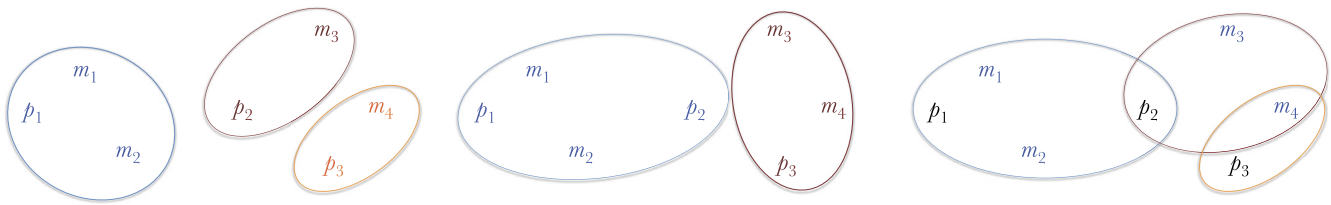
Let $\mathbf{A}_{\{T_j=1\}}$ denote the subvector of treatment assignments for the interventional units in the interference set for unit m_j, that is, the elements A_i corresponding to $p_i \in T_j$. Let $\mathbf{A}_{\{T_j\neq 1\}}$ be the treatment assignment subvector for interventional units *not* in the interference set for m_j. We reformulate the usual SUTVA as follows to formalize the meaning of the interference mapping:

ASSUMPTION (Structured SUTVA). For a specified interference mapping, T:

(i) $Y_j(\mathbf{A}) = Y_j(\mathbf{A}')$ for all j if $\mathbf{A} = \mathbf{A}'$,
(ii) $Y_j(\mathbf{A}) = Y_j(\mathbf{A}')$ for all j when $\mathbf{A}_{\{T_j=1\}} = \mathbf{A}'_{\{T_j=1\}}$ or equivalently, $Y_j(\mathbf{A}_{\{T_j=1\}}, \mathbf{A}_{\{T_j\neq 1\}}) = Y_j(\mathbf{A}_{\{T_j=1\}})$.

Part (ii) of structured SUTVA clarifies that potential outcomes for unit m_j need only be considered in terms of the treatment assignment vector of the $p_i \in T_j$. To simplify notation in the subsequent, we will use the subscript $(-i)$ denoting "not i" to implicitly refer to all interventional units in a given interference set except for p_i. For example, $Y_j(A_i = a, A_{(-i)} = \mathbf{a}_{(-i)})$ will refer to the potential outcome at m_j if p_i receives treatment a and the remainder of interventional units in T_j, denoted with $p_{k\neq i} \in T_j$, receive treatment vector $\mathbf{a}_{(-i)}$.

Several familiar settings can be formulated via T and structured SUTVA. To aid illustration, Figure 1 schematically depicts three bipartite interference mappings for a simple setting with $M = 4$ and $P = 3$, where ovals surrounding units represent membership in interference sets. A setting where outcome units are clustered hierarchically such that each $m_j \in \mathcal{M}$ is subject to exactly one A_i (e.g., students grouped within classrooms) and there is no interference is pictured in Figure 1(a). The mapping in this



(a) Clusters with SUTVA (b) Partial Interference (c) General Interference

FIG. 1. *Illustrations of interference mappings in simplified setting with* ($\mathcal{M} = \{m_1, m_2, m_3, m_4\}$) *and* ($\mathcal{P} = \{p_1, p_2, p_3\}$). *Potential outcomes at* m_j *depend upon treatments at all* p_i *in the same oval.*

setting is

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

More generally, this type of setting corresponds to T_j having exactly one element equal to 1, with every $T_j = T_{j'}$ when m_j and $m_{j'}$ are in the same cluster and, otherwise, $T_j^\top T_{j'} = 0$. The structure depicted in Figure 1(b) corresponds to a bipartite version of the so-called *partial interference* assumption [10, 22], where: (1) units are divided into nonoverlapping clusters consisting of ≥ 1 outcome unit and ≥ 1 interventional unit; and (2) outcome-unit potential outcomes are allowed to depend only on the treatments assigned to interventional units within the same cluster. Figure 1(b) corresponds to

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

and this setting is generally defined by specifying T_j as a P-vector with ith element equal to 1 only for p_i in the same cluster, maintaining the feature that $T_j^\top T_{j'} = 0$ when m_j and $m_{j'}$ are not in the same cluster. Figure 1(c) depicts a more general interference structure that cannot be described by nonoverlapping clusters as in the partial interference case, corresponding to

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

Using Figure 1(c) as an example, the set of upwind power plants for unit m_1 is $T_1 = (1, 1, 0)$, and the set of downwind zip codes for unit p_2 is $T_2^\top = (1, 1, 1, 1)$. Note that formulation of interference mappings in the standard single-unit setting could proceed analogously, but with T as $M \times M$ (or $P \times P$). This would include the most standard setting of no interference, corresponding to $T = \text{diag}\{1\}_{M \times M}$.

4. ESTIMANDS FOR BIPARTITE CAUSAL INFERENCE WITH INTERFERENCE

As with other settings of causal inference with interference, the interconnectedness between units may not only complicate inference for familiar causal estimands, but may also introduce new causal estimands of interest. Among causal estimands of frequent interest in the presence of interference with one level of observational unit are so-called "total" and "overall" effects. We focus in particular on other estimands akin to "direct effects," which capture the effect of changing the treatment status of a given interventional unit while holding fixed the treatment statuses of other interventional units in the interference set, and "indirect" effects, which capture the effect of holding one interventional unit's treatment status fixed but changing the treatment statuses of others. We focus on these estimands in particular to explicate complications arising in the bipartite setting due to the fact that treatment is not directly applied or withheld from outcome units, as in studies with one level of observational unit, and notions of "direct" and "indirect" take on a somewhat different meaning.

Recall that, for a specified interference mapping, T, the $(-i)$ subscript denotes all interventional units but p_i within the interference set for a given m_j, that is, $p_{k \neq i} \in T_j$. In principle, causal effects can be defined as comparisons between $Y_j(\mathbf{a})$, $Y_j(\mathbf{a}')$ for any two $\{\mathbf{a}, \mathbf{a}'\} \in \mathcal{A}(P)$, that is, any two intervention allocations in the space of possible allocations. As a starting point for development, denote the most primitive individual-level causal effects as

$$\begin{aligned} &Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{a}_{(-i)}) \\ &\quad - Y_j\big(A_i = a', \mathbf{A}_{(-i)} = \mathbf{a}'_{(-i)}\big), \end{aligned} \tag{1}$$

which denotes the causal effect on outcome unit m_j of treatment allocation $\mathbf{a}$ with $a_i = a$ *versus* treatment allocation $\mathbf{a}'$ with $a'_i = a'$. A key feature of the bipartite setting highlighted in (1) is the natural definition of individual effect for every (p_i, m_j) pair for $m_j \in \mathcal{M}$ and $p_i \in \mathcal{P}$. For example, setting $a = 0$, $a' = 1$ and $\mathbf{a}_{(-i)} = \mathbf{a}'_{(-i)}$ in (1) yields a quantity akin to a "direct" effect on outcome unit m_j of treating (vs. not treating) interventional unit p_i while holding the treatment status of all other $p_{k \neq i} \in T_j$ fixed at $\mathbf{a}_{(-i)}$. P such "direct" effects could be defined for outcome unit m_j.

4.1 Individual-Level Estimands Based on Average Potential Outcomes Under Classes of Treatment Allocations

While development of causal estimands with interference has followed along several lines of development, we adopt a perspective analogous to [10, 18], where estimands are defined based on average individual-level potential outcomes, averaged over many possible treatment allocations. For example, much work has focused on "allocation strategies" representing values of $\mathbf{a}$ that adhere to a certain probability (or proportion) of treated units, typically denoted with α [23, 13, 18, 19].

We extend this convention and define α to denote a counterfactual treatment allocation strategy where the propensity of interventional units in an interference set to receive treatment $A_i = 1$ is set to α. In the bipartite setting, we refer to the definition of α as "$\mathcal{M}$-centric" in that it refers to the allocation to all units in the interference

set for a particular m_j, for example, all power plants "upwind" from a specific zip code. The set of possible treatment allocations adhering to α is denoted with $\mathcal{A}(|T_j|)$, where $|T_j|$ denotes the number of interventional units in the interference set for m_j.

In the bipartite setting, individual average potential outcomes that average over all treatment allocations fixing $A_i = a$ for a $p_i \in T_j$ and having treatment propensity of the interference set fixed to α are defined as

$$\overline{Y}_j(A_i = a, \alpha) = \sum_{s \in \mathcal{A}(|T_j|-1)} Y_j(A_i = a, \mathbf{A}_{(-i)} = s) \times \pi(s|A_i = a; \alpha), \tag{2}$$

where $s \in \mathcal{A}(|T_j| - 1)$ } denotes the set of possible $\mathbf{a}_{(-i)}$ that, along with $a_i = a$, lie in $\mathcal{A}(|T_j|)$. Here, $\pi(s|A_i = a; \alpha)$ denotes the probability of each such allocation, conditional on A_i being fixed at a, which is specified by the researcher to, for example, represent independent Bernoulli allocation of treatments to units or realistic interventions dependent on covariates [23, 18]. Average potential outcomes of the form (2) will be used to construct causal estimands of interest.

Using (2), we define a bipartite version of an individual-level "direct effect," where "direct" is used to refer to the effect of treating (vs. not) a specific $p_i \in T_j$, while holding the treatment allocation strategy fixed at α:

$$DE_{(i,j)}(\alpha) = \overline{Y}_j(A_i = 1; \alpha) - \overline{Y}_j(A_i = 0; \alpha). \tag{3}$$

For example, $DE_{(i,j)}(\alpha)$ would be the direct effect on outcome unit m_j of treating (vs. not) the ith power plant, when all upwind plants are assigned treatment according to α.

Similarly, we define a bipartite version of an individual-level "indirect effect," where "indirect" is used to refer to the effect of holding the treatment status of a specific p_i fixed, while changing the allocation to other $p_{k \neq i} \in T_j$:

$$IE^a_{(i,j)}(\alpha, \alpha') = \overline{Y}_j(A_i = a; \alpha) - \overline{Y}_j(A_i = a; \alpha'). \tag{4}$$

For example, $IE^a_{(i,j)}(\alpha, \alpha')$ would be the indirect effect on outcome unit m_j of holding the treatment status of power plant p_i to $A_i = a$ and changing treatment allocations of other upwind power plants from α to α'.

In addition to expanded notation relative to settings with one level of unit, the salient feature of individual-level effects such as (3) and (4) is that they are defined, in full generality, for every (p_i, m_j) pair of $p_i \in \mathcal{P}$ and $m_j \in \mathcal{M}$. This is because, unlike in the single-unit setting, there is no automatic or self-evident notion of which treatment "directly" applies to each unit; interest could lie, at least in principle, in the effect of intervening at any power plant on any zip code location. This introduces different strategies for defining the types of average causal effects that will be discussed in Section 4.2.

4.2 $\mathcal{M}$-Indexed Average Causal Effects

Recall that, unlike in standard settings of interference where treatments are given directly to one level of unit, the bipartite setting entails no automatic or self-evident notion of which treatment applies directly to which unit. Thus, it may be of interest to average individual-average potential outcomes for each outcome unit over all interventional units in the interference set. We introduce the term $\mathcal{M}$-indexed average potential outcomes to refer to average potential outcomes for a given $m_j \in \mathcal{M}$, averaged over $p_i \in T_j$:

$$\overline{Y}_j(a, \alpha) = \frac{1}{|T_j|} \sum_{i \in T_j} \overline{Y}_j(A_i = a, \alpha), \tag{5}$$

which are defined to represent the average potential outcome under $A_i = a$ and allocation program α across all interventional units in the interference set for m_j.

The $\mathcal{M}$-indexed average potential outcomes in (5) can be used to define average causal effects paralleling those defined in Section 4.1. Define the $\mathcal{M}$-indexed average direct effect as

$$DE_j(\alpha) = \overline{Y}_j(1, \alpha) - \overline{Y}_j(0, \alpha) = \frac{1}{|T_j|} \sum_{i \in T_j} DE_{(i,j)}(\alpha) \tag{6}$$

to denote the average effect on outcome unit m_j of treating a single $p_i \in T_j$ while holding fixed the treatment propensities for all $p_{k \neq i} \in T_j$, averaged over all $p_i \in T_j$. The population-average $\mathcal{M}$-indexed direct effect could be defined as $\overline{DE}_{\mathcal{M}} = \frac{1}{M} \sum DE_j(\alpha)$ representing, for example, the average effect on hospitalizations of installing (vs. not) SnCR on a single upwind power plant while holding the treatment probability of all upwind plants fixed at α.

Similarly, the $\mathcal{M}$-indexed indirect effect is defined as

$$IE^a_j(\alpha, \alpha') = \overline{Y}_j(a, \alpha) - \overline{Y}_j(a, \alpha') = \frac{1}{|T_j|} \sum_{i \in T_j} IE^a_{(i,j)}(\alpha, \alpha') \tag{7}$$

to represent the average effect on outcome unit m_j of holding treatment at a single $p_i \in T_j$ fixed at a while changing the treatment allocation for the interference set from α to α', averaged over all $p_i \in T_j$. The population-average $\mathcal{M}$-indexed indirect effect could be defined as $\overline{IE}_{\mathcal{M}} = \frac{1}{M} \sum IE_j(\alpha)$ representing, for example, the average effect of holding the SnCR status fixed at an upwind power plant while changing the SnCR allocation of all other upwind plants from α to α'.

4.3 $\mathcal{P}$-Indexed Average Causal Effects

The indexing of individual-level potential outcomes in (2) (and their corresponding individual-level estimands in

(3) and (4)) by both the outcome unit j and interventional unit i invites averaging potential outcomes over $p_i \in T_j$, as in the $\mathcal{M}$-indexed quantities in Section 4.2, or averaging potential outcomes over $m_j \in T_i^\top$, which might be referred to as "$\mathcal{P}$-indexed" quantities. $\mathcal{P}$-indexed effects analogous to (5), (6) and (7) could be defined for a particular p_i based on averaging potential outcomes over $m_j \in T_i^\top$ representing, for example, the average impact of a treatment decision at a particular power plant, averaged across all downwind zip codes. A main complication with such quantities under the present framework relates to α which, recall, is inherently $\mathcal{M}$-centric in that it refers to the allocation of treatments to interventional units in the interference set for m_j, $p_i \in T_j$. Thus, while calculating $\mathcal{M}$-centric average potential outcomes involves fixing α to a single interference set (T_j), calculating a $\mathcal{P}$-indexed average potential outcome would correspond to averaging over potential outcomes for outcome units $m_j \in T_i^\top$ with potentially different interference sets, that is, to fixing the treatment allocation of $p_i \in T_j$ for all $m_j \in T_i^\top$. For example, one could define a $\mathcal{P}$-indexed direct effect analogous to (6) to characterize how installing an SnCR system at power plant p_i affects hospitalization outcomes, on average across all downwind zip codes, with each downwind zip code having propensity of SnCR installation among its respective upwind plants fixed to α. Such $\mathcal{P}$-indexed effects, while potentially of interest and an important topic for future work, are not pursued here in favor of exploration of a subset of $\mathcal{M}$-indexed effects for which estimators can be derived from existing work.

4.4 Key-Associated Average $\mathcal{M}$-Indexed Causal Effects

The fundamental feature that individual-level causal effects can be naturally defined for every (p_i, m_j) pair in the bipartite setting may be simplified in settings where each outcome unit can be associated with a single $p_i \in T_j$ at which intervening is of particular interest. Denote such an interventional unit with $p^*_{i(j)}$, defined for every $m_j \in \mathcal{P}$. We will refer to $p^*_{i(j)}$ as the "key associated" interventional unit for outcome unit m_j. In practice, criteria for determining the relevant $p^*_{i(j)}$ for every $m_j \in \mathcal{M}$ will undoubtedly vary, but examples in the power plant setting include the closest or largest power plant located upwind from a given location. When indexing other quantities defined for $p^*_{i(j)}$, we will simplify notation and use the subscript i^*. For example, A_{i^*} will be used to denote the treatment assignment of $p^*_{i(j)}$.

The potential outcomes and estimands in Section 4.2 averaged over all interventional units for each m_j, owing to the fact that the bipartite setting does not inherently contain a notion of which p_i corresponds "directly" to each m_j. However, definition of a $p^*_{i(j)}$ for every $m_j \in \mathcal{M}$, invites focus on only a subset of the individual-level causal effects of types (3) and (4), specifically those corresponding to the intrinsic interest in the key-associated interventional unit. Rather than consider every (p_i, m_j) pair, interest is confined to exactly one individual-level direct effect ($DE_{(i^*,j)}(\alpha)$) and exactly one individual-level indirect effect ($IE^a_{(i^*,j)}(\alpha, \alpha')$) for each $m_j \in \mathcal{M}$.

Population-average analogs of these effects can be defined as

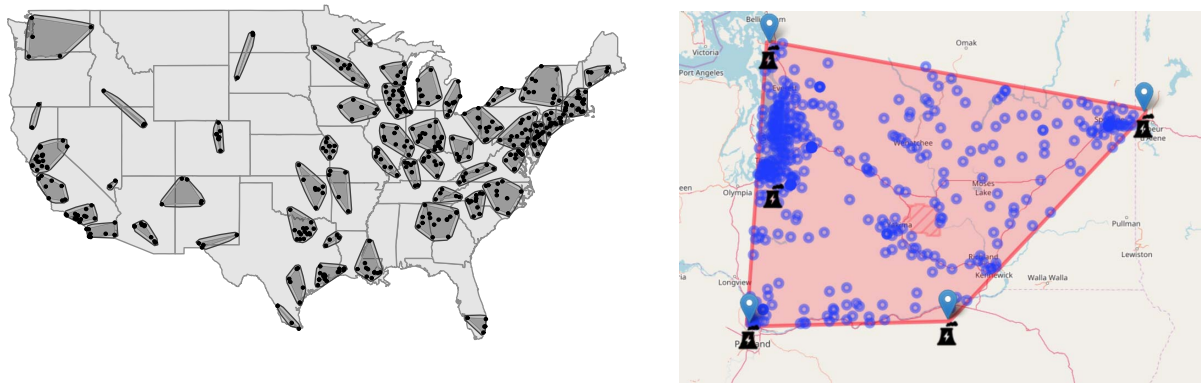
$$\overline{DE}^*(\alpha) = \frac{1}{M}\sum_{j=1}^{M} DE_{(i^*,j)}(\alpha), \tag{8}$$

$$\overline{IE}^{*a}(\alpha, \alpha') = \frac{1}{M}\sum_{j=1}^{M} IE^a_{(i^*,j)}(\alpha, \alpha'). \tag{9}$$

The estimand (8) corresponds to the average effect on outcome units of treating (vs. not) the key-associated unit while holding fixed the allocation program to other interventional units in the interference set. In the power plant example, this could correspond, for example, to the average effect on hospitalizations of installing an SnCR system on the closest power plant while holding fixed the allocation of SnCR systems to other upwind plants. The estimand (9) corresponds to the average effect on outcome units of holding the treatment at the key-associated unit fixed while varying the allocation program to other interventional units in the interference set from α to α'. In the power plant example, this could correspond to the average effect on hospitalizations of holding the SnCR status of the closest power plant fixed while changing the allocation to other upwind plants.

5. ESTIMATORS UNDER BIPARTITE PARTIAL INTERFERENCE IN OBSERVATIONAL STUDIES

While the development in Section 4 pertains to a general form of interference mappings, T, we illustrate the development of bipartite estimators for the simplified setting of *partial interference*, for which existing estimators in the one unit setting extend in a relatively straightforward way. Consider a partition of $\mathcal{P}$ into K nonoverlapping clusters of interventional units: $\{P^1, P^2, \ldots, P^K\}$, each of size $|P^k|$. For example, power plants could be clustered according to geographic proximity. Consider a corresponding grouping of $\mathcal{M}$ into exactly K nonoverlapping clusters $\{M^1, M^2, \ldots, M^K\}$, where each M^k consists of $|M^k|$ outcome units that are linked in some fashion to the interventional units in P^k. For example, M^k could consist of all of zip code locations within a certain distance of at least one of the power plants in P^k. Partial interference in this case assumes that potential outcomes at $m_j \in M^k$ depend only on the treatments assigned to $p_i \in P^k$. In the terminology of Section 3.2, this amounts to an interference mapping where T has a block structure such that, for $k = 1, 2, \ldots, K$, T_j is the same for all



(a) (b)

FIG. 2. (a) *Grouping of power plants in interference clusters and assignment of zip codes to clusters. Each cluster is depicted with two polygons, the inner polygon corresponds to the convex hull of the power plants, and the outer polygon corresponds to the convex hull of zip code centroids in that cluster.* (b) *One cluster of power plants and corresponding zip codes with zip codes' centroids depicted in blue.*

$m_j \in M^k$, with $t_{ji} = 1$ for all $p_i \in P^k$ and $t_{ji} = 0$ otherwise. Recall that this implies $T_j^\top T_l = 0$ for all $m_j \in M^k$, $m_l \in M^{k'}$, denoting no common interventional units in the interference sets for two outcome units in different clusters. For simplicity, assume that for each $k = 1, 2, \ldots, K$, both M^k and P^k contain at least one unit of their respective type. Figure 2 illustrates one such clustering in the power plant example.

5.1 Cluster-Level Average Potential Outcomes Under Partial Interference

The partial interference assumption invites definition of cluster-specific analogs to the average effects proposed in Section 4.2. The expressions $p_i \in T_j$ and $m_j \in T_i^\top$ become equivalent to $p_i \in P^k$ and $m_j \in M^k$, and α takes on the familiar meaning of the cluster-level treatment propensity referring to all $p_i \in P^k$. $\mathcal{M}$-indexed effects such as those in (6) and (7) could be averaged over all $j \in M^k$ for all $k = 1, 2, \ldots, K$ to create cluster averages. However, we focus on developing estimators for the power plant setting that correspond to analogs to the key-associated $\mathcal{M}$-indexed estimands defined in Section 4.4.

Specifically, based on (2) we define cluster-level average potential outcomes of the form:

$$\overline{Y}^k(A_{i^*} = a, \alpha) = \frac{1}{|M^k|} \sum_{j \in M^k} \overline{Y}_j(A_{i^*} = a, \alpha) \tag{10}$$

to denote the cluster-level average potential outcome when $p^*_{i(j)}$ receives treatment a and all other p_i in the cluster receive allocation program α, averaged over all outcome units in the cluster. Population average potential outcomes can be subsequently defined with $\overline{Y}(A_{i^*} = a, \alpha) = \sum_k \overline{Y}^k(A_{i^*} = a, \alpha)/K$.

Formulation of cluster-average potential outcomes leads to the following expressions for cluster-level average direct and indirect effects:

$$\begin{aligned} DE^{k*}(\alpha) &= \overline{Y}^k(A_{i^*} = 1, \alpha) - \overline{Y}^k(A_{i^*} = 0, \alpha) \\ &= \frac{1}{|M^k|} \sum_{j \in M^k} DE_{(i^*, j)}(\alpha), \end{aligned} \tag{11}$$

$$\begin{aligned} IE^{k*a}(\alpha, \alpha') &= \overline{Y}^k(A_{i^*} = a, \alpha) - \overline{Y}^k(A_{i^*} = a, \alpha') \\ &= \frac{1}{|M^k|} \sum_{j \in M^k} IE^a_{(i^*, j)}(\alpha). \end{aligned} \tag{12}$$

Population-level effects defined in (8) and (9) can be constructed from (11) and (12) as

$$DE^*(\alpha) = \frac{1}{K} \sum_k DE^{k*}(\alpha), \quad \text{and} \tag{13}$$

$$IE^{a*}(\alpha, \alpha') = \frac{1}{K} \sum_k IE^{ka*}(\alpha). \tag{14}$$

5.2 IPTW Estimator for Average Potential Outcomes

Here we illustrate that, among all the estimands defined for the bipartie setting in Section 4, existing estimators in [23, 19, 14, 18] are essentially directly applicable to estimands that rely on: (1) clusters of units and partial interference and (2) a relevant key-associated $p^*_{i(j)}$ defined for each $m_j \in \mathcal{M}$. Technical development follows from previous work, ensuring that population (cluster) quantities related to treatment assignment are confined to $i = 1, 2, \ldots, P(P^k)$ while population (cluster) quantities related to outcomes are confined to $j = 1, 2, \ldots, M(M^k)$. Otherwise, theoretical underpinnings of the estimators extend trivially.

Specifically, we propose a refinement (to reflect the bipartite setting) of the simple estimator proposed in [23] for the cluster-level average potential outcomes in (10). A corresponding estimator for the population-level average potential outcome follows immediately, with asymptotically normal distribution as the number of clusters K

increases to infinity. This development follows existing work in [18, 19, 14, 23], leading directly to estimators for the population-level key-associated $\mathcal{M}$-indexed direct and indirect effects in (13) and (14) with known asymptotic distributions.

The estimator for the cluster-level average potential outcome has the familiar form:

$$\begin{aligned}&\widehat{Y}^k(A_{i^*}=a;\alpha)\\ &\quad=\frac{1}{|M^k|}\sum_{j\in M^k}\frac{\pi(\mathbf{A}^k_{(-i^*)}|A_{i^*}=a,\alpha)}{f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k|\mathbf{W}^k,\mathbf{X}^k)}\\ &\qquad\times I(A_{i^*}=a)Y_j,\end{aligned}\tag{15}$$

with corresponding estimator for the population-average potential outcome:

$$\widehat{Y}(A_{i^*}=a;\alpha)=\frac{1}{K}\sum_{k=1}^{K}\widehat{Y}^k(A_{i^*}=a;\alpha).\tag{16}$$

The term $f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k|\mathbf{W}^k,\mathbf{X}^k)$ in the denominator of (15) represents the cluster-level propensity score for the probability that the $p_i \in P^k$ receive the observed treatment vector $\mathbf{A}^k$, conditional on the interventional-unit and outcome unit covariates in the cluster, $\mathbf{W}^k$ and $\mathbf{X}^k$. The term $\pi(\mathbf{A}^k_{(-i^*)}|A_{i^*}=a,\alpha)$ in the numerator of (15) represents the user-specified probability distribution of different cluster-level treatment allocations adhering to the program α (specified in accordance with (2)).

Under the following assumptions and following work in [23, 14, 18]: $\widehat{Y}^k(A_{i^*}=a;\alpha)$ in (15) is unbiased for $\overline{Y}^k(A_{i^*}=a,\alpha)$ in (10) for the known cluster propensity score; unbiasedness of $\widehat{Y}(A_{i^*}=a;\alpha)$ in (16) for $\overline{Y}(A_{i^*}=a,\alpha)$ follows trivially.

ASSUMPTION 1 (Positivity). For $k\in\{1,2,\ldots,K\}$, the probability of observing cluster treatment vector $\mathbf{A}^k=\mathbf{a}^k$ given cluster covariates $\mathbf{W}^k$, $\mathbf{X}^k$ is denoted by $f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k=\mathbf{a}^k|\mathbf{W}^k,\mathbf{X}^k)$ and is positive for all $\mathbf{a}^k\in\mathcal{A}(|P^k|)$.

ASSUMPTION 2 (Ignorability). For $k\in\{1,2,\ldots,K\}$, the observed cluster treatment $\mathbf{A}^k$ is conditionally independent of the set of cluster potential outcomes $\mathbf{Y}^k(\cdot)$ given the cluster covariates $\mathbf{W}^k$, $\mathbf{X}^k$, denoted as $\mathbf{A}^k \perp\!\!\!\perp \mathbf{Y}^k(\cdot)|\mathbf{W}^k,\mathbf{X}^k$.

Under superpopulation (of clusters) versions of Assumption 1 and Assumption 2 as stated in [18], $\widehat{Y}(A_{i^*}=a,\alpha)$ is consistent and asymptotically normal for the superpopulation counterpart to the above estimands for a known or correctly specified and estimated parametric propensity score model ($f_{\mathbf{A}|\mathbf{W},\mathbf{X},k}(\mathbf{A}^k|\mathbf{W}^k,\mathbf{X}^k)$). Appendix A presents a simulation study evaluating the performance of the above estimators for the bipartite setting.

6. EVALUATING SNCR SYSTEMS ON MEDICARE HOSPITALIZATIONS IN THE PRESENCE OF POLLUTION TRANSPORT

A previous analysis that simplified the bipartite structure by projecting to the level of $\mathcal{P}$ showed that SnCR systems at coal- or gas-fired power plants reduce ambient air pollution in the areas immediately surrounding power plants and in other "downwind" areas [18]. Here, we conduct an analysis of SnCR on hospitalizations with a more complete regard for the bipartite structure of the problem. Specifically, we estimate direct and indirect effects (13) and (14) of SnCR installation on zip code hospitalizations for CVD among Medicare beneficiaries.

The set of interventional units consists of 473 coal or natural gas burning power plants operating in the continental U.S. during the summer months (June–August) of 2004. These power plants, with their characteristics and important aggregate area-level characteristics (i.e., $\mathbf{W}$) have been previously described in detail [17]. Power plants are partitioned into 50 clusters as in [18] using Ward's agglomerative hierarchical clustering [25] with an objective function based solely on geographic closeness of power plants within a cluster. Clustering nearby power plants is meant to be a rough approximation to the phenomenon of pollution transport that dictates interference in this setting.

The initial set of outcome units considered for this analysis corresponds to 37,240 U.S. zip codes, each with a measured number of hospitalizations for cardiovascular disease (codes ICD-9 390 to 459) among Medicare fee-for-service beneficiaries in 2005 (no outcome-unit covariates, $\mathbf{X}$, are included in the analysis). Zip codes were assigned to a cluster of power plants if the zip code centroid was located within the area defined by the power plant locations' convex hull and a buffer zone of 30 km. If a zip code belonged to more than one cluster based on this definition, it was assigned to the cluster that included the closest power plant. If a zip code was not within 30 km of the buffer zone of any power plant cluster, it was excluded from the analysis. This resulted in a total of 18,807 zip codes representing the population of interest of areas of the U.S. that are considered likely to be impacted by interventions at power plants (See Figure 2). Figure 3 shows the observed distribution of the hospitalization outcome over the 18,807 zip codes.

For each zip code, m_j, the key-associated power plant, $p^*_{i(j)}$, is defined to be the power plant located closest to the centroid of the zip code. Corresponding key-indexed direct and indirect effects thus cohere to notions of intervening to control local pollution (e.g., from the closest plant) versus those to control long-range transported pollution from more distant upwind plants, which are important distinctions for development of local and interstate (or regional) regulatory policies. Key-indexed direct and

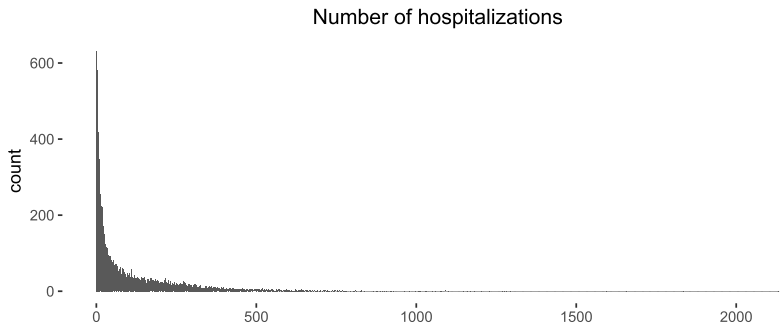


FIG. 3. *Distribution of observed* 2005 *number of cardiovascular hospitalizations for the* 18,807 *zip codes included in the analysis.*

indirect effects were estimated using the IPW estimator defined in Section 5.2 for values of α ranging from the 20th to the 80th percentile of the observed proportion of treated power plants across the clusters. The propensity score model was specified as a logistic regression adjusting linearly for power plant, weather and demographic covariates and including a cluster-specific random effect to match the previous analysis of [18]. Power plant characteristics include the percent of total capacity at which a plant typically operates, the amount of fuel energy burned, an indicator of Phase 2 participation in the Acid Rain Program, an indicator for whether a plant burns mostly gas fuel, and indicators for the size of the plant in terms of number of generating units. Area-level characteristics include ambient temperature, median household income, median house value, population per square mile and population percentages of high school graduates, residence in urban areas, white, black and hispanic populations, housing occupancy, poverty and migration to the area within 5 years.

The numerator specifying counterfactual treatment allocation probabilities was specified as independent Bernoulli assignments to treatment $\pi(\mathbf{A}^k_{(-i^*)}|A_{i^*} = a, \alpha) = \prod_{p_k \neq i \in T_j} \alpha^{A_k}(1-\alpha)^{1-A_k}$ as in [23]. Results are depicted in Figure 4. The direct effect is estimated to be negative for all values of α (achieving statistical significance at the 0.05 level for all $\alpha \geq 0.1$), implying that installation of SnCR at a zip code's closest power plant leads to a significant reduction in number of cardiovascular hospitalizations at that location. Note that the direct effect becomes less pronounced as α increases, indicating that installing SnCR on a zip code's closest power plant has a smaller impact on CVD hospitalizations when more upwind power plants also have SnCR installed. The other two plots in Figure 4 depict estimates of the indirect effect $IE^{0*}(\alpha_1, \alpha_2)$ for values $\alpha_1 \in \{0.1, 0.4\}$. These results represent expected changes in hospitalizations when the closest power plant does not have SnCR and the propensity of upwind power plants to install SnCR shifts from α_1 to α_2. The decreasing trend in both plots indicates that a higher proportion of SnCR among upwind plants leads to decreased CVD hospitalizations when the closest power plant remains without SnCR. For example, a change in the propensity of upwind units to install SnCR from 10% to 45.8% would lead to 56.4 (95% CI: 25.8–87.1) fewer hospitalizations on average per zip code when the closest plant remains without SnCR.

Overall, the results of the analysis indicate the benefit of installing SnCR for reducing CVD hospitalizations among Medicare beneficiaries with a careful account of how the effectiveness of controls installed at nearby power plants interacts with interventions at upwind plants. For illustration, Appendix B presents alternative analyses

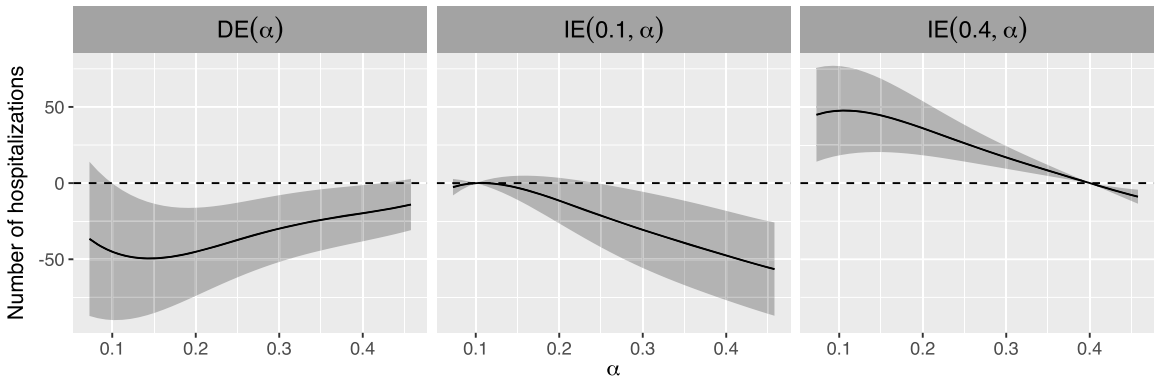


FIG. 4. *Direct effect and indirect effect estimates and confidence intervals. For the indirect effects* ($IE(\alpha_1, \alpha_2)$), *values of* α_1 *are fixed to* 0.1 (*middle panel*) *and* 0.4 (*right panel*) *and the* x*-axis corresponds to varying values for* α_2.

of SnCR for reducing CVD hospitalizations, intentionally simplified to reduce the bipartite structure and rely on more familiar estimation estimation strategies.

7. DISCUSSION

We have introduced the new setting of bipartite causal inference with interference, which arises in a variety of settings where interventions are enacted at one type of observational unit, outcomes of interest are defined and measured at a distinct type of unit, and the complexities of exposure patterns lead outcomes to depend on the treatments of many interventional units. The setting is particularly relevant to the study of air pollution regulatory policy, where interventions occur at pollution sources (e.g., power plants), health outcomes are measured at population locations (e.g., zip codes) and complex atmospheric processes and long-range pollution transport lead to interference. Formalization of this setting represents an important added dimension to recent work on interference that extends the formality of potential outcomes methods to settings that do not cohere to the oft-considered setting of one level of observational unit and unit-to-unit outcome dependencies (e.g., infectious diseases or social networks).

Potential outcomes and causal estimands were formulated generally, drawing commonalities and distinctions with existing work for interference. While the general development of estimands was designed to introduce the possibilities of formalizing the bipartite interference problem, estimators were developed for only a subset of possible estimands. For illustration and to motivate the use of the bipartite framework in an applied problem, we ultimately employed estimators that rely on the assumption of partial interference and require that interest lies in a single key-associated interventional unit for each outcome unit. The proposed estimators relied heavily on existing work developed in the one-unit setting, and made use of simplistic clustering methods to form partial interference clusters based only on a rough approximation of the true interference process. Future research to expand beyond these simplified estimators is clearly warranted, including formulations that acknowledge much richer structures of interference beyond simple clustering and those that go beyond the perspective of individual-average potential outcomes averaged over specified allocation programs (e.g., as in [5, 11]).

Even with the simplifications that led to the proposed estimators, the formalization of bipartite interference and application of the simplified estimators in the context of the power plant evaluation represents an important step in air pollution policy research that, to our knowledge, has only previously been considered in [18]. Long-range pollution transport according to atmospheric processes is ubiquitous to the study of pollution interventions at point sources (e.g., power plants or factories), and formal methods for statistical evaluation are lacking for such interventions. Despite the progress herein, the clustering and partial interference assumption employed in the analysis of SnCR systems is a nontrivial simplification of actual pollution transport, and produces only an approximate analysis of the effect of SnCR on Medicare CVD hospitalizations. Extensions to more general interference patterns are essential, and a topic of ongoing work.

APPENDIX A: SIMULATION STUDY

Here, we provide a simulation study to evaluate the operating characteristics of the inverse probability of treatment weighting estimator in (15) and (16), with $\pi(\mathbf{A}^k_{(-i^*)}|A_{i^*} = a, \alpha)$ specified to denote independent Bernoulli assignments to treatment among units within a cluster. Note that this simulation study closely resembles a bipartite version of the simulation studies in [18] and [19].

A.1 Data Generation

2000 clusters were simulated to represent the superpopulation of interest. Each cluster included 3, 4, 5 or 6 interventional units. The number of outcome units per cluster was simulated from a Poisson distribution with mean 20. The key-associated interventional unit for each outcome unit was chosen randomly over the interventional units in the cluster.

For each interventional unit, we assumed the presence of two covariates $\mathbf{W} = (W_1, W_2)$ generated as independent $N(0, 0.2^2)$ random variables. For each outcome unit, we simulated potential outcomes $Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{a}_{(-i)})$ from Bernoulli$(p(\mathbf{a}))$ for

$$\begin{aligned}\text{logit}\, p(\mathbf{a}) &= 0.5 - 0.6a - 1.4\alpha - 0.098W_{1i} \\ &\quad - 0.145W_{2i} + 0.351a\alpha,\end{aligned} \tag{17}$$

where $\alpha = \mathbf{1}'\mathbf{a}/\mathbf{1}'\mathbf{1}$ is the proportion of treated interventional units in the cluster. This way of generating potential outcomes assumes that, when $a = a'$ and $\mathbf{1}'\mathbf{a}_{(-i)} = \mathbf{1}'\mathbf{a}'_{(-i)}$, then $Y_j(A_i = a, \mathbf{A}_{(-i)} = \mathbf{a}_{(-i)}) = Y_j(A_i = a', \mathbf{A}_{(-i)} = \mathbf{a}'_{(-i)})$. Note that the IPW estimator is agnostic with regard to the model for generating potential outcomes, and for that reason we would expect that the performance of the IPW estimator would be comparable under different and more complicated potential outcome generative models (17).

Potential outcomes generated according to (17) are fixed across replicated data sets. These potential outcomes are used to calculate the superpopulation average potential outcomes, direct and indirect effects. These quantities are the target for estimation and used to evaluate operating characteristics of the estimator applied to samples from this superpopulation.

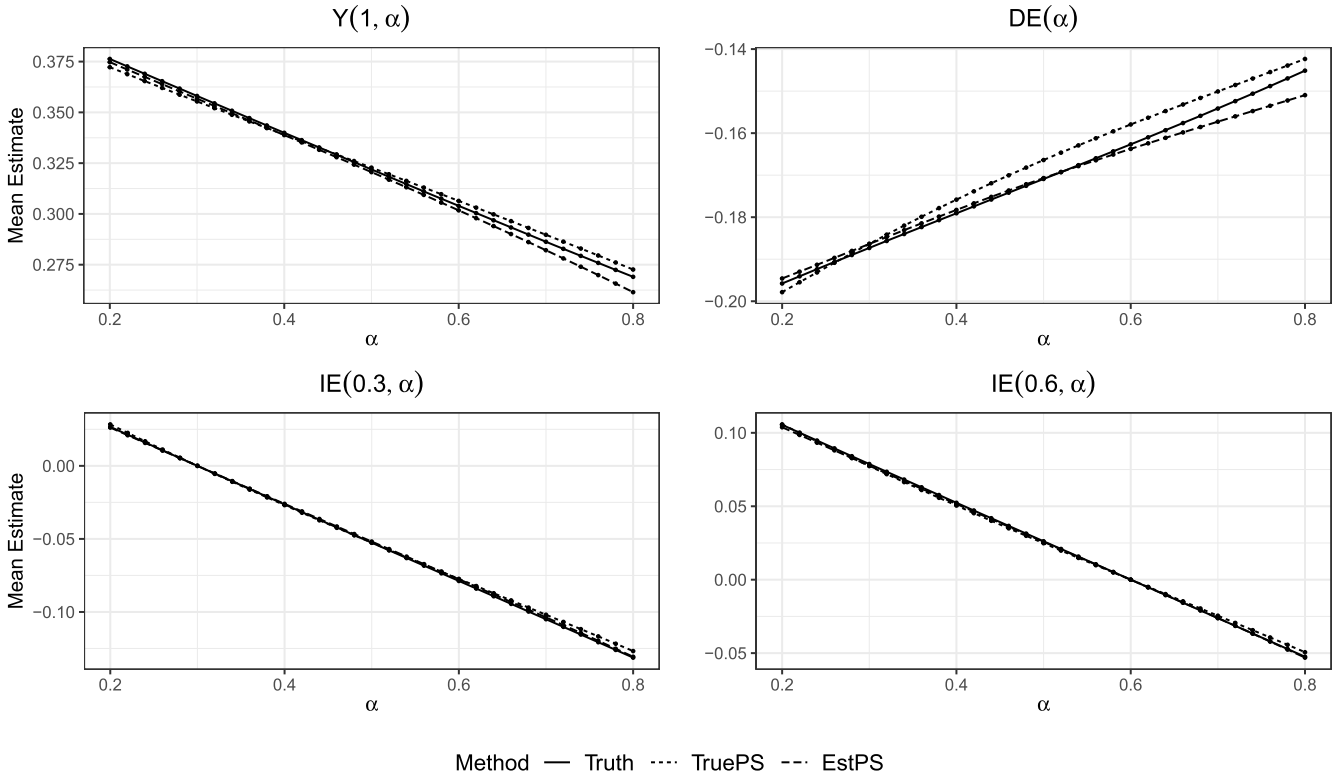


FIG. 5. *Mean estimate of the population average potential outcome as a function of* α*, direct effect as a function of* α *and indirect effect* $IE(\alpha_1, \alpha_2)$ *as a function of* α_2 *for* $\alpha_1 \in \{0.3, 0.6\}$.

To evaluate the estimator, we simulated draws from the superpopulation for various values of N, the number of clusters. For each $N \in \{200, 400, 600, 800, 1000\}$, we simulated 200 data sets, and for each data set N clusters were sampled randomly from the superpopulation. When a cluster was sampled, all of its units (interventional and outcome) are observed. For each simulated data set, we generate the observed treatment assignment from a Bernoulli(p) for

$$\text{logit}\, p = -0.2 + b_k + 0.3W_1 - 0.15W_2,$$

where $b_k \sim N(0, 0.2^2)$ is a cluster-specific random effect. The observed outcome for each outcome unit corresponds to the potential outcome for the observed level of the treatment.

A.2 Simulation Results

For each simulated data set, we calculate the IPW estimator in (15) and (16), direct and indirect effects and variance estimates based on asymptotic approximations, using both true and estimated propensity scores.

We present detailed results for $N = 200$ clusters, but results for other values of N were similar. In Figure 5, we show the mean estimate over 200 simulated data sets of the population average potential outcome for $a = 1$,

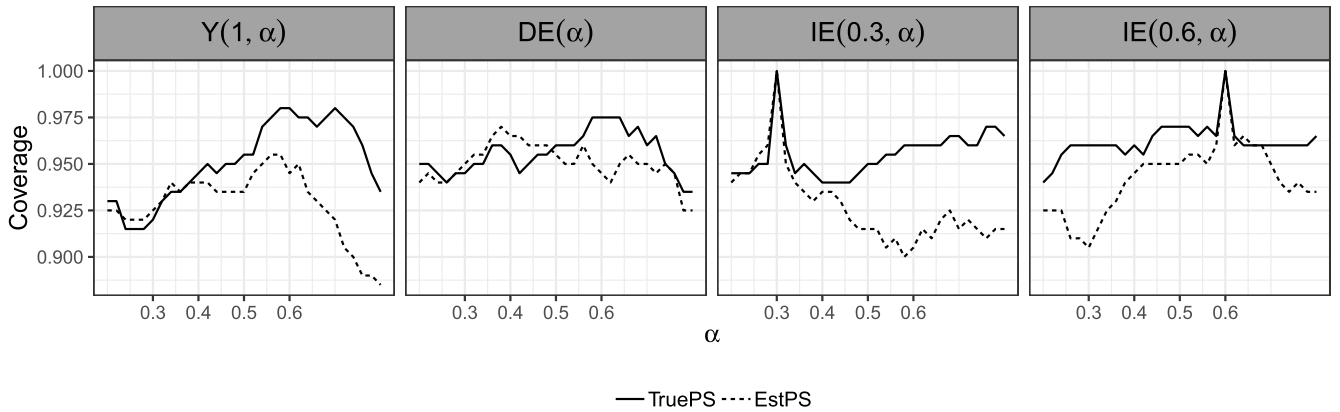


FIG. 6. *Coverage of* 95% *confidence intervals for the IPW estimator based on the true and estimated propensity score.*

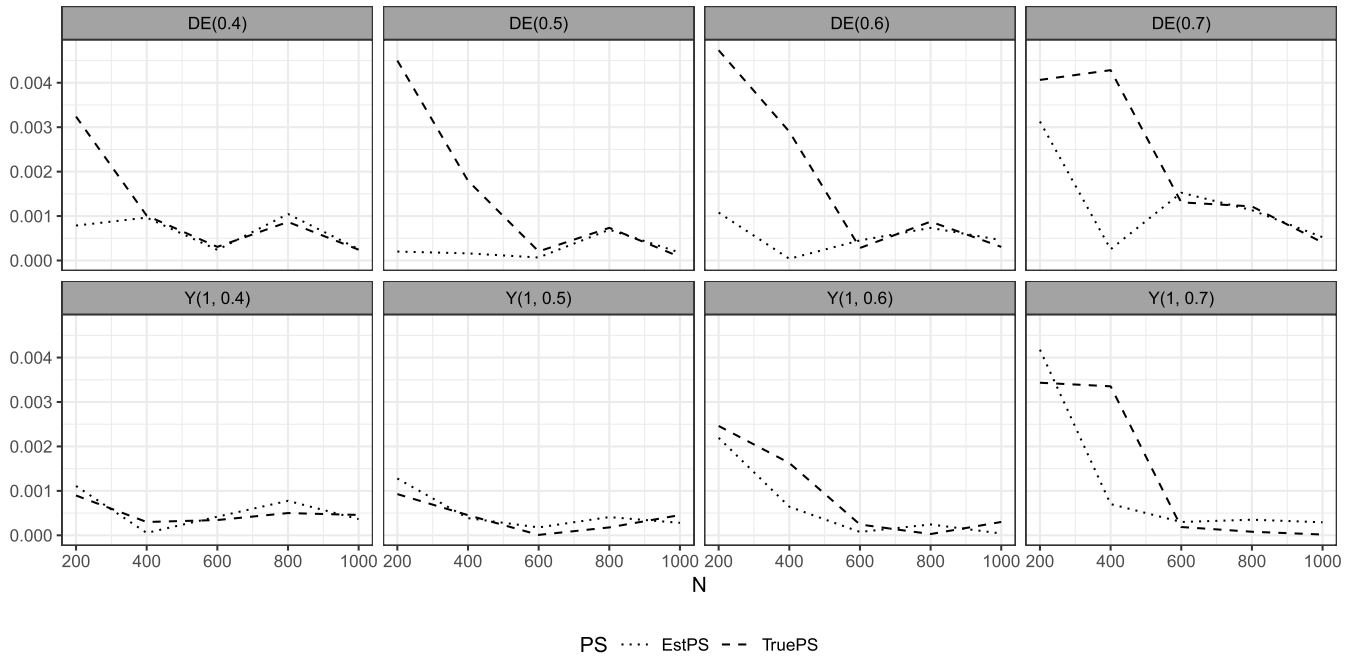


FIG. 7. *Absolute bias as a function of the number of clusters for the potential outcome and direct effect estimator employing the true or the correctly specified and estimated propensity score.*

the direct effect and indirect effect for $a = 0$ and $\alpha_1 \in \{0.3, 0.6\}$. Results for different values of a and α_1 are similar. Note that that the estimators are closed to unbiased for all quantities. The correctly specified and estimated propensity score estimator returns direct effect estimates slightly closer to the truth in comparison to the IPW estimator using the true propensity score. In Figure 6, we show that the IPW estimator using the true or the correctly-specified estimated propensity score achieves nominal coverage of the 95% confidence intervals.

Lastly, in Figure 7 we show the absolute bias of the IPW estimator for the true and the estimated propensity score as a function of the number of clusters. We see that the estimator employing the estimated propensity score has smaller bias compared to the estimator using the true propensity score. As expected, the bias of both estimators is declining in the number of clusters.

APPENDIX B: ALTERNATIVE SIMPLIFIED EVALUATIONS OF SNCR SYSTEMS ON MEDICARE HOSPITALIZATIONS

For comparison with the results analysis of Section 6, we performed two alternative analyses for illustration that maintain the same cluster structure as in Section 6, but simplify the bipartite structure of the data as described in Section 3.1 and rely on more familiar estimands and estimation procedures:

1. *Projection to interventional units*: We projected all data to the space of interventional units by linking each zip code to its closest power plant and then assigning each power plant an outcome defined as the total number of cardiovascular hospitalizations among linked zip codes. Covariates used for adjustment ($\boldsymbol{W}$) are exactly the same as those in Section 6, which were only measured at the interventional units. The resulting data set consists of clusters of power plants, with each power plant having a single treatment, a set of observed covariates and a single outcome representing health outcomes at nearby zip codes. Then the analysis was performed as in the more familiar setting of one level of observational unit (power plants) using a cluster propensity score model identical to that in Section 6 with the estimands and estimators in [23] and asymptotic variances in [19]. Note that this approach is the same as one employed in Web Appendix B of [18], but using hospitalizations (instead of ambient pollution) as the outcome.

2. *Projection to outcome units with outcome modeling*: We projected all data to the space of outcome units by assigning to each zip code two "treatment" quantities: (1) the treatment assigned to the closest power plant (the "key associated" treatment) and (2) the proportion of power plants within the zip code's cluster that were treated (the "neighborhood" treatment). Then we fit a Poisson model predicting the total number of hospitalizations as a function of the key associated treatment, the neighborhood treatment (linear, quadratic and cubic term), interaction between the two treatments (with neighborhood treatment linear and quadratic term) and aggregate cluster covariates defined as the average of covariates in $\boldsymbol{W}$ over power plants in the cluster. We used these models to predict potential outcomes under alternative specifications of (a, α), averaged within groups to estimate a

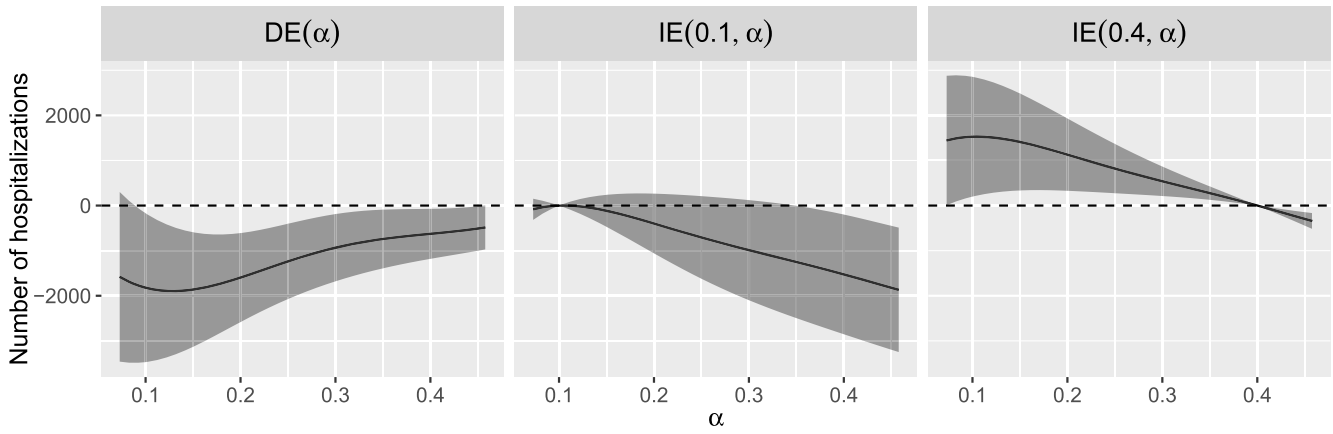


FIG. 8. *Direct and indirect effect estimates and* 95% *confidence intervals from Alternative Analysis* 1 *based on projection to the level of power plants. Estimates correspond to expected changes in the number of hospitalizations in the area surrounding a power plant.*

quantity resembling the group average potential outcome, and averaged over clusters to estimate a quantity resembling the population average potential outcome. Inference was acquired by employing the bootstrap over clusters, similarly as in [18]. Note that this analysis represents a simplified parametric version of methods outlined in [5, 11].

Figures 8 and 9 show the estimated direct and indirect effects based on these analyses. In Figure 8, describing the analysis of the data projected to the level of power plants, we see similar patterns as in the analysis in Figure 4, but with a smaller region of values for (α_1, α_2) for which indirect effects are estimated to be significantly different from zero. Note too the difference in interpretation (and scale) between the estimates in Figure 8 and those in Section 6, owing to the fact that the results in Figure 8 represent effects on cardiovascular hospitalizations across *all* zip codes linked to a power plant, whereas those in Figure 4 represent average effects in cardiovascular hospitalizations within a single zip code. The similar patterns in results from the analysis projected to power plants and those from the genuine bipartite analysis in Section 6 derive in part from the specific strategies for linking zip codes and defining the "key associated" interventional units, as well both analyses' reliance on only power-plant level covariates. Such a correspondence is not expected in general.

Results in Figure 9 from the analysis projecting to the level of zip code are substantially different. Point estimates are much closer to zero, and no indirect effect $IE(\alpha_1, \alpha_2)$ was identified as significantly different from zero. In addition to the different definition and interpretation of estimands with only zip codes as the observational unit, these differences in the effect estimates are likely driven by the reliance on a (relatively simple) parametric model for cardiovascular outcomes, which is expected to be more prone to model misspecification than analyses relying on propensity score weighting. Thus, we particularly stress the use of this alternative analysis as a simple

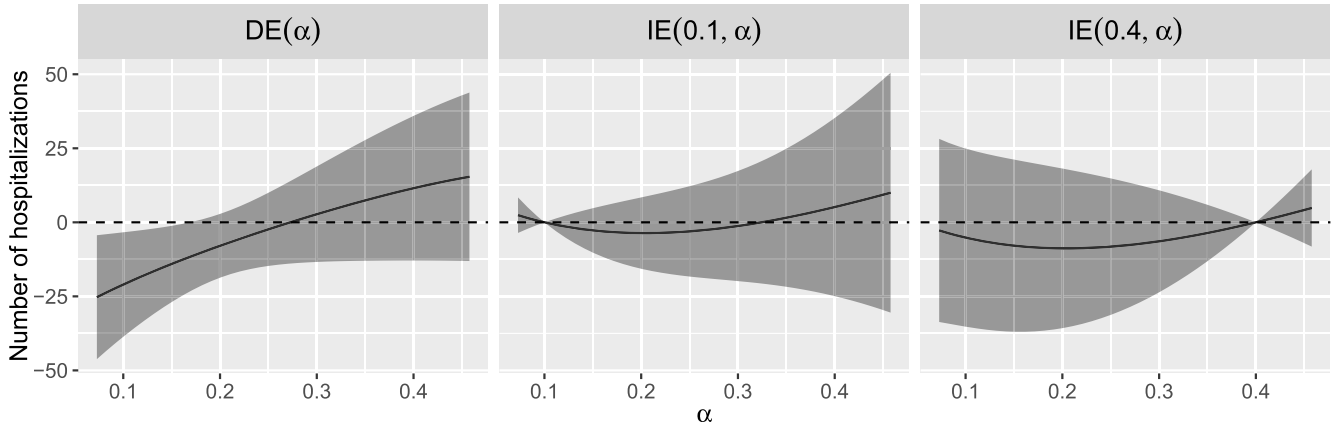


FIG. 9. *Direct and indirect effect estimates and* 95% *confidence intervals from Alternative Analysis* 2 *based on projection to the level of zip codes. Estimates correspond to expected changes in the number of hospitalizations in a zip code for a change in the treatment of the closest power plant or a change in the probability of treatment of other power plants in the cluster.*

illustration of what might occur in an analysis that simplified the bipartite structure in this way.

ACKNOWLEDGMENTS

The authors thank Dr. Chanmin Kim for helpful discussions in preliminary graphical illustrations and Dr. Christine Choirat for operationalizing the assignment of zip codes to clusters of power plants.

This work was supported by research funding from NIH R01ES026217 and EPA 83587201. Its 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

[1] ARONOW, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Methods Res.* **41** 3–16. MR3190698 https://doi.org/10.1177/0049124112437535

[2] ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* **11** 1912–1947. MR3743283 https://doi.org/10.1214/16-AOAS1005

[3] BELL, M. L., MCDERMOTT, A., ZEGER, S. L., SAMET, J. M. and DOMINICI, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987–2000. *JAMA J. Am. Med. Assoc.* **292** 2372–2378.

[4] BOWERS, J., FREDRICKSON, M. M. and PANAGOPOULOS, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* **21** 97–124.

[5] FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2016). Identification and estimation of treatment and interference effects in observational studies on networks. Preprint. Available at arXiv:1609.06245 [stat].

[6] GRAHAM, D. J., MCCOY, E. J. and STEPHENS, D. A. (2013). Quantifying the effect of area deprivation on child pedestrian casualties by using longitudinal mixed models to adjust for confounding, interference and spatial dependence. *J. Roy. Statist. Soc. Ser. A* **176** 931–950. MR3120956 https://doi.org/10.1111/j.1467-985X.2012.01071.x

[7] HALLORAN, M. E. and STRUCHINER, C. J. (1991). Study designs for dependent happenings. *Epidemiology* **2** 331–338.

[8] HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6** 142.

[9] HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

[10] HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

[11] KARWA, V. and AIROLDI, E. (2018). A systematic investigation of classical causal inference strategies under misspecification due to network interference. Preprint. Available at arXiv:1810.08259v1 [stat.ME].

[12] KIM, C., DANIELS, M. J., HOGAN, J. W., CHOIRAT, C. and ZIGLER, C. M. (2019). Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls. *Ann. Appl. Stat.* **13** 1927–1956. MR4019162 https://doi.org/10.1214/19-AOAS1260

[13] LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564 https://doi.org/10.1080/01621459.2013.844698

[14] LIU, L., HUDGENS, M. G. and BECKER-DREPS, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika* **103** 829–842. MR3620442 https://doi.org/10.1093/biomet/asw047

[15] LUO, X., SMALL, D. S., LI, C.-S. R. and ROSENBAUM, P. R. (2012). Inference with interference between units in an fMRI experiment of motor inhibition. *J. Amer. Statist. Assoc.* **107** 530–541. MR2980065 https://doi.org/10.1080/01621459.2012.655954

[16] OGBURN, E. L. and VANDERWEELE, T. J. (2017). Vaccines, contagion, and social networks. *Ann. Appl. Stat.* **11** 919–948. MR3693552 https://doi.org/10.1214/17-AOAS1023

[17] PAPADOGEORGOU, G., CHOIRAT, C. and ZIGLER, C. M. (2019). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics* **20** 256–272. MR3922132 https://doi.org/10.1093/biostatistics/kxx074

[18] PAPADOGEORGOU, G., MEALLI, F. and ZIGLER, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics* **75** 778–787.

[19] PEREZ-HEYDRICH, C., HUDGENS, M. G., HALLORAN, M. E., CLEMENS, J. D., ALI, M. and EMCH, M. E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics* **70** 734–744. MR3261790 https://doi.org/10.1111/biom.12184

[20] ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112

[21] RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [Ann. Agric. Sci. **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

[22] SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

[23] TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

[24] VERBITSKY-SAVITZ, N. and RAUDENBUSH, S. W. (2012). Causal inference under interference in spatial settings: A case study evaluating community policing program in Chicago. *Epidemiol. Methods* **1** 107–130.

[25] WARD, J. H. JR. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58** 236–244. MR0148188

[26] ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289–302.