Statistical Science
2014, Vol. 29, No. 4, 559–578
DOI: 10.1214/14-STS501

Causal Diagrams for Interference

Elizabeth L. Ogburn and Tyler J. VanderWeele

Abstract. The term "interference" has been used to describe any setting in which one subject's exposure may affect another subject's outcome. We use causal diagrams to distinguish among three causal mechanisms that give rise to interference. The first causal mechanism by which interference can operate is a direct causal effect of one individual's treatment on another individual's outcome; we call this *direct interference*. *Interference by contagion* is present when one individual's outcome may affect the outcomes of other individuals with whom he comes into contact. Then giving treatment to the first individual could have an indirect effect on others through the treated individual's outcome. The third pathway by which interference may operate is *allocational interference*. Treatment in this case allocates individuals to groups; through interactions within a group, individuals may affect one another's outcomes in any number of ways. In many settings, more than one type of interference will be present simultaneously. The causal effects of interest differ according to which types of interference are present, as do the conditions under which causal effects are identifiable. Using causal diagrams for interference, we describe these differences, give criteria for the identification of important causal effects, and discuss applications to infectious diseases.

Key words and phrases: Causal diagrams, causal inference, contagion, DAGs, graphical models, infectiousness, interference, nonparametric identification, social networks, spillover effects.

Traditionally, causal inference has relied on the assumption of no interference, that is, the assumption that any subject's outcome depends only on his own treatment and not on the treatment of any other subject. This assumption is often implausible; for example, it is violated when the outcome is an infectious disease and treating one individual may have a protective effect on others in the population. Recent work in statistics has focused on relaxing the assumption of no interference (Graham, Imbens and Ridder, 2010; Halloran and Struchiner, 1995; Hudgens and Halloran, 2008; Manski, 2013; Rosenbaum, 2007; Tchetgen Tchetgen and VanderWeele, 2012; Vansteelandt, 2007). Much of this work has been motivated by the study of infectious diseases (Halloran and Struchiner, 1995; Tchetgen Tchetgen and VanderWeele, 2012; VanderWeele and Tchetgen Tchetgen, 2011a, 2011b; Halloran and Hudgens, 2012). Researchers have also explored the implications of interference on residents of neighborhoods when some residents are given housing vouchers to move (Sobel, 2006) or when new resources are introduced to the neighborhood (VanderWeele, 2010). Others have written about the interference that arises from assigning children to classrooms and assigning classrooms to educational interventions (Graham, Imbens and Ridder, 2010; Hong and Raudenbush, 2008; VanderWeele et al., 2013). The rising prominence of social networks in public health research underscores the need for methods that take into account the interconnections among individuals' treatments and outcomes (Christakis and Fowler, 2007; Cohen-Cole and Fletcher, 2008; Mulvaney-Day and Womack, 2009).

Elizabeth L. Ogburn is Assistant Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room E3620, Baltimore, Maryland 21205, USA (e-mail: eogburn@jhsph.edu). Tyler J. VanderWeele is Professor, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Kresge Building, Boston, Massachusetts 02115, USA (e-mail: tvanderw@hsph.harvard.edu).

Graphical models have shed light on the identification of causal effects in many settings (Dahlhaus and Eichler, 2003; Didelez, Kreiner and Keiding, 2010; Freedman, 2004; Greenland, Pearl and Robins, 1999; Pearl, 1995, 1997, 2000; Robins, 2003; Tian and Pearl, 2002a; Vansteelandt, 2007) but have not yet been applied to settings with interference. In this paper, we describe how to draw causal diagrams representing the complex interdependencies among individuals in the presence of interference, and how to use those diagrams to determine what variables must be measured in order to identify different causal effects of interest. We review the literature on causal diagrams and identification of causal effects in the absence of interference in Section 1, and recent work on the estimation of causal effects in the presence of interference in Section 2. In Section 3, we discuss which covariates must be measured and controlled for in order to identify causal effects in the presence of interference. Section 4 introduces the three distinct types of interference, provides causal diagrams to help explicate their structure, and describes some of the causal effects we would wish to estimate and the assumptions required to identify them. In Section 5, we use the concepts introduced in Section 4 to elucidate the nature of interference in social networks. Section 6 concludes the paper.

1. REVIEW OF IDENTIFICATION OF CAUSAL EFFECTS IN THE ABSENCE OF INTERFERENCE

Suppose that we wish to estimate the average causal effect of a treatment A on an outcome Y from observational data on n individuals for whom we have also measured a vector of confounders C. For simplicity, we will assume in this section and the next that A is binary and Y is continuous, but our remarks apply equally to A and Y discrete or continuous. Under the assumptions of no interference and a single version of treatment (we will not discuss the latter assumption here; see VanderWeele and Hernan, 2013, for discussion), $Y_i(a), a = 0, 1$ is defined as the counterfactual outcome we would have observed if, possibly contrary to fact, subject i had received treatment a. The average causal effect of A on Y is equal to $E[Y(1)] - E[Y(0)]$, and it is identified under the three additional assumptions of consistency,

$$Y_i(a) = Y_i \quad \text{if } A_i = a, \tag{1}$$

conditional exchangeability,

$$Y_i(a) \amalg A_i | C_i, \tag{2}$$

and positivity,

$$P(A_i = a | C_i = c) > 0 \quad \text{for all } a \text{ in the support of } A \text{ and for all } c \text{ in the support of } C \text{ such that } P(C = c) > 0. \tag{3}$$

We refer the reader to Hernán and Robins (2006) for discussion of these assumptions.

The conditional exchangeability assumption is sometimes referred to as the "no unmeasured confounding assumption." Identifying the variables that must be included in C can be assessed with the aid of causal diagrams (e.g., Greenland and Robins, 1986; Greenland, Pearl and Robins, 1999; Pearl, 2003).

Causal diagrams, or causal directed acyclic graphs (DAGs) consist of nodes, representing the variables in a study, and arrows, representing causal effects. In a slight abuse of terminology, we will not distinguish between nodes on a DAG and the variables they represent. A DAG is a collection of nodes and arrows in which no variable is connected to itself by a sequence of arrows aligned head-to-tail. A *causal* DAG is a DAG on which arrows represent causal effects and that includes all common causes of any pair of variables on the graph. The causal DAG in Figure 1 represents the scenario in which the effect of A on Y is confounded by a single confounder C. The three arrows encode the causal effects of C on A, C on Y, and A on Y. We briefly introduce terminology and results for DAGs but refer the reader to Pearl (2000, 2003) for details and discussion. Recently, Richardson and Robins (2013) introduced a new class of causal diagrams called single world intervention graphs (SWIGs). This work can be immediately and fruitfully applied to the interference settings we discuss below; however, in the interest of space we restrict our attention to DAGs.

A *path* on a DAG is any unbroken, nonrepeating sequence of arrows connecting one variable to another. A *directed path* (or a *causal path* on a causal DAG) is a path that follows arrows from tail to head. A variable X is an *ancestor* (or *cause*, if the DAG is causal) of Z if there is a directed path from X to Z. Equivalently, Z is a *descendent* of X. If the directed path from X to Z consists of a single arrow, then X is a *parent* of Z and Z is a *child* of X. On a causal DAG, we would say that X has a *direct effect* on Z. If a path

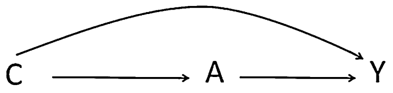

FIG. 1.

includes X, W and Z and if there are arrows from both X and Z into W, then W is a *collider* on the path. A collider is a path-specific concept. For example, in Figure 1, Y is a collider on one path from C to A (the path $C \to Y \leftarrow A$) but not on another (the path $C \to A \to Y$). A path can be *unblocked*, meaning roughly that information can flow from one end to the other, or *blocked*, meaning roughly that the flow of information is interrupted at some point along the path. If all paths between two variables are blocked, then the variables are *d-separated*, and if two variables are d-separated on a causal DAG then they are statistically independent. A path is blocked if there is a collider on the path such that neither the collider itself nor any of its descendants is conditioned on. An unblocked path can be blocked by conditioning on any noncollider along the path. Two variables are d-separated by a set of variables if conditioning on the variables in the set suffices to block all paths between them, and if two variables are d-separated by a third variable or a set of variables then they are independent conditional on the third variable or set of variables (Pearl, 1995, 2000).

A *backdoor path* from X to Z is one that begins with an arrow pointing into, rather than out of, X. For example, the path $A \leftarrow C \to Y$ in Figure 1 is a backdoor path from A to Y. Pearl (1995) proved that conditioning on a set of nondescendants of A that block all backdoor paths from A to Y suffices for exchangeability to hold for the effect of A on Y. This set need not be unique.

Identification of effects, other than total effects, often requires assumptions beyond (1), (2), and (3). *Path-specific effects* quantify the causal effect of one variable on another via specific causal pathways. Consider the DAG in Figure 2, which adds a mediator M to the path from A to Y. Now there are two different causal pathways from A to Y, namely $A \to Y$ and $A \to M \to Y$. Causal effects of the form $E[Y(a)] - E[Y(a')]$ capture all causal pathways from A to Y without distinguishing among them, but we may be interested specifically in direct effects, which bypass the mediator, and indirect effects, which go through the mediator. Define $M_i(a)$ to be the counterfactual value we would have observed for M_i if A_i had been set to a, and $Y_i(a, m)$ to be the counterfactual value of Y_i that we would have observed if M_i had been set to m and A_i to a. We make the additional consistency assumptions that $M_i(a) = M_i$ when $A_i = a$, that $Y_i(a, m) = Y_i$ when $A_i = a$ and $M_i = m$, and that $Y_i(a, M_i(a)) = Y_i(a)$. Then the *natural direct effect* is defined as $E[Y(a, M(a))] - E[Y(a', M(a))]$; it measures the expected change in Y due to a change in A, holding M fixed at $M(a)$. A direct path from X to Z is said to be *deactivated* in a particular causal contrast if X is set to the same value in the counterfactual for Z in both terms of the contrast. A path is deactivated if any arrow on the path is deactivated. In the natural direct effect, A is set to the same value in the counterfactual for M in both terms of the contrast; therefore the natural direct effect can be conceptualized as the effect of A on Y with the path $A \to M$ deactivated (Pearl, 2001). The *natural indirect effect*, defined as $E[Y(a', M(a))] - E[Y(a', M(a'))]$, measures the expected change in Y when A is held fixed but M changes from $M(a)$ to $M(a')$. This is the effect of A on Y with the arrow from A to Y deactivated (A is set to a' in the counterfactual for Y in both terms of the contrast). The natural direct and indirect effects sum to the total effect of A on Y: $E[Y(a)] - E[Y(a')] = E[Y(a, M(a))] - E[Y(a', M(a'))] = \{E[Y(a, M(a))] - E[Y(a', M(a))]\} + \{E[Y(a', M(a))] - E[Y(a', M(a'))]\}$. The *controlled direct effect* of A on Y, given by $E[Y(a, m)] - E[Y(a', m)]$ fixes M at a specific value m and compares the counterfactual outcomes under two different values of A. This is the effect of A on Y with the path $M \to Y$ deactivated.

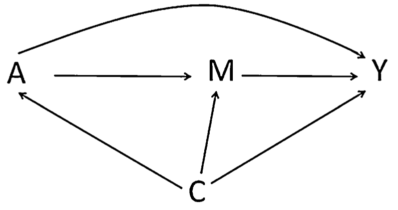

FIG. 2.

In order to identify the controlled direct effect, the following assumptions are sufficient:

$$Y_i(a, m) \amalg A_i | C_i \tag{4}$$

and

$$Y_i(a, m) \amalg M_i | A_i, C_i. \tag{5}$$

These correspond respectively to the absence of unblocked backdoor paths from A_i to Y_i (except possibly through M_i) conditional on C_i and from M_i to Y_i conditional on A_i and C_i. Avin, Shpitser and Pearl (2005) proved that in most settings the following is a necessary assumption for the identification of the average natural direct and indirect effects of A_i on Y_i mediated by M_i: there is no variable W_i such that (i) there is an activated directed path from A_i to W_i, (ii) there is a deactivated directed path from W_i to Y_i and (iii) there is an activated directed path from W_i to Y_i. A variable that satisfies conditions (i), (ii) and (iii) is known as a *recanting witness*, and we call the assumption of no variable satisfying these conditions the *recanting witness criterion*. In the context of natural direct and indirect effects, the recanting witness criterion is met if there is no confounder of the mediator–outcome relation that is caused by treatment, or (Pearl, 2001)

$$Y_i(a,m) \amalg M_i(a')|C_i. \tag{6}$$

Assumptions (5), (4), (6) and

$$M_i(a) \amalg A_i|C_i, \tag{7}$$

that is, the absence of unblocked backdoor paths from A_i to M_i conditional on C_i, suffice to identify the natural direct and indirect effects.

2. REVIEW OF IDENTIFICATION OF CAUSAL EFFECTS IN THE PRESENCE OF INTERFERENCE

Interference is present when one subject's outcome may depend on other subjects' treatments (Rosenbaum, 2007). It is often reasonable to make a *partial interference* assumption that interference can only occur within subgroups or *blocks* of subjects. This may be justified if the blocks are separated by time or space (Hudgens and Halloran, 2008; Sobel, 2006; Tchetgen Tchetgen and VanderWeele, 2012; Rosenbaum, 2007). Under interference, $Y_i(a)$ is not well-defined, since the value of Y that would have been observed for subject i had he received treatment a may depend on the treatments received by other subjects. We define counterfactual notation for interference following Hudgens and Halloran (2008), Tchetgen Tchetgen and VanderWeele (2012), Rubin (1990) and Halloran and Struchiner (1995). Suppose that n individuals fall into N blocks, indexed by k, with $m = n/N$ individuals in each block. If $N = 1$, so that interference may occur between any two subjects in the population, then we say that there is full interference. If $N = n$, then an individual's treatment can only affect his own outcome and there is no interference. Let $\mathbf{A}_k \equiv (A_{k1}, \dots, A_{km})$ be the vector of treatment assignments for individuals in block k and let $\mathbf{a}_k$ denote an m-dimensional vector in the support of $\mathbf{A}_k$. Let $\mathbf{Y}_k \equiv (Y_{k1}, \dots, Y_{km})$ and $\mathbf{C}_k \equiv (C_{k1}, \dots, C_{km})$ be the vector of outcomes and array of covariates, respectively, for individuals in block k. In what follows, we reserve boldface letters for vectors or arrays of length m in which the ith entry corresponds to the ith individual in block k, and we omit the subscript k when taking expectations over blocks. Define $Y_{ki}(\mathbf{a}_k)$ to be the counterfactual outcome that we would have observed for individual i in block k under an intervention that set $\mathbf{A}_k$ to $\mathbf{a}_k$. Following Tchetgen Tchetgen and VanderWeele (2012) we replace assumption (1) above with a new assumption of consistency under interference:

$$Y_{ki}(\mathbf{a}_k) = Y_{ki} \quad \text{when } \mathbf{A}_k = \mathbf{a}_k. \tag{8}$$

We also require modified positivity and exchangeability assumptions in order to identify causal effects under interference: we assume that we have measured a set of pretreatment covariates C for each individual such that (Tchetgen Tchetgen and VanderWeele, 2012)

$$Y_{ki}(\mathbf{a}_k) \amalg \mathbf{A}_k|\mathbf{C}_k \tag{9}$$

and

$$\begin{aligned} &P(\mathbf{A}_k = \mathbf{a}_k|\mathbf{C}_k = \mathbf{c}_k) > 0 \\ &\text{for all } \mathbf{a}_k \text{ in the support of } \mathbf{A}_k \\ &\text{and for all } \mathbf{c}_k \text{ in the support of } \mathbf{C}_k. \end{aligned} \tag{10}$$

These assumptions suffice to identify the expectations of counterfactuals of the form $Y_{ki}(\mathbf{a}_k)$ whenever $\mathbf{a}_k$ is an instance of a well-defined intervention $\mathbf{a}$ and, therefore, to identify causal effects that are contrasts of such expectations. An intervention will be well-defined if it uniquely determines which subjects in block receive treatment. Well-defined interventions are possible, for example, if all blocks are of the same size, if the individuals in each block are distinguishable from one another, and if the individuals are ordered in the same way across blocks. Suppose interference occurs within blocks comprised of a father (subject 1), a mother (subject 2) and a child (subject 3). Then an intervention $(1, 0, 1)$ indicates that the father and child receive treatment but the mother does not. If the blocks are of different sizes or if there is no natural way to distinguish among the individuals in each block, some interventions may be well-defined under assumptions that the effects of treatment are the same for different members of the block and do not depend on the size of the block, for example, the intervention that assigns treatment to every individual in every block. We assume throughout that all interventions are well-defined. For simplicity, we assume that the blocks are of the same size and that there is a natural ordering of the subjects in each block, but most of our comments and results extend to more general settings. (In the absence of well-defined interventions, some causal effects can still be defined, identified and estimated under two-stage randomization; see Hudgens and Halloran, 2008; Vanderweele, Tchetgen and Halloran, 2012; VanderWeele and Tchetgen Tchetgen, 2011b; Halloran and Hudgens, 2012.)

In Section 3, we use graphical models to determine which variables must be included in the conditioning set in order for exchangeability to hold. This gives the identification criteria under interference for causal effects that are contrasts of expectations of counterfactuals of the form $Y_{ki}(\mathbf{a}_k)$. Sometimes we may wish to identify path-specific effects; these require additional assumptions for identification that we discuss below.

In this paper, we focus on identification, rather than estimation, of causal effects. We merely note here that, for the purposes of estimation and inference, the effective sample size is N and thus observation of multiple blocks may be required.

2.1 Causal Effects of $\mathbf{A}$ on Y

Although recognition of the fact that interference may occur in certain settings dates at least as far back as Ross (1916), it is only recently that progress has been made on identifying causal effects in the presence of interference. Halloran and Struchiner (1995) defined four effects that do not depend on understanding the mechanisms underlying interference and that are identifiable under assumptions (8), (9) and (10).

The overall effect of intervention $\mathbf{a}$ compared to intervention $\mathbf{a}'$ on subject i is defined as $OE_i(\mathbf{a}, \mathbf{a}') \equiv E[Y_i(\mathbf{a})] - E[Y_i(\mathbf{a}')]$. We use the index i to indicate that the expectations do not average over individuals within a block but rather over blocks for a particular individual i. For example, if the blocks are comprised of a father (subject 1), a mother (subject 2) and a child (subject 3), then $OE_3(\mathbf{a}, \mathbf{a}') = E[Y_3(\mathbf{a})] - E[Y_3(\mathbf{a}')]$ is the overall effect on a child of intervention $\mathbf{a}$ compared to intervention $\mathbf{a}'$. The average overall effect $OE(\mathbf{a}, \mathbf{a}') \equiv E[Y(\mathbf{a})] - E[Y(\mathbf{a}')]$, where $E[Y(\mathbf{a})] \equiv \frac{1}{m}\sum_{i=1}^{m} E[Y_i(\mathbf{a})]$, averages over the empirical mean of the counterfactual outcomes for each block. The unit-level effect of treatment on subject i fixes the treatment assignments for all subjects in each block except i, and compares the counterfactual outcomes for subject i under two different treatment assignments. Let $\mathbf{a}_{k,-i} = (a_{k,1}, \dots, a_{k,i-1}, a_{k,i+1}, \dots, a_{k,m})$ be a vector of length $m - 1$ of treatment values for all subjects in block k except for subject i. Then $UE_i(\mathbf{a}; \tilde{a}, \bar{a}) \equiv E[Y_i(\mathbf{a}_{-i}, \tilde{a})] - E[Y_i(\mathbf{a}_{-i}, \bar{a})]$, where $Y_{ki}(\mathbf{a}_{k,-i}, \tilde{a})$ is subject i's counterfactual outcome under the intervention in which the subjects in block k except for subject i receive treatments $\mathbf{a}_{k,-i}$ and subject i receives treatment $\tilde{a}$. The spillover effect of intervention $\mathbf{a}$ compared to intervention $\mathbf{a}'$ on subject i fixes i's treatment level and compares his counterfactual outcomes under the two different interventions. That is, $SE_i(\mathbf{a}, \mathbf{a}'; \tilde{a}) \equiv E[Y_i(\mathbf{a}_{-i}, \tilde{a})] - E[Y_i(\mathbf{a}'_{-i}, \tilde{a})]$. (The unit-level effect is often referred to as the direct effect and the spillover effect as the indirect effect of an intervention, but in order to avoid confusion with the direct effect for DAGs defined in Section 2 and the natural direct and indirect effects defined in Section 2.2, we will use different terminology. See Tchetgen Tchetgen and VanderWeele, 2012, and Vanderweele, Tchetgen and Halloran, 2012, for further discussion of terminology.) We can also average these effects over individuals within a block. The average unit-level effect is $UE(\mathbf{a}; \tilde{a}, \bar{a}) \equiv E[Y(\mathbf{a}_-, \tilde{a})] - E[Y(\mathbf{a}_-, \bar{a})]$ and the average spillover effect is $SE(\mathbf{a}, \mathbf{a}'; \tilde{a}) \equiv E[Y(\mathbf{a}_-, \tilde{a})] - E[Y(\mathbf{a}'_-, \tilde{a})]$, where $E[Y(\mathbf{a}_-, \tilde{a})] \equiv \frac{1}{m}\sum_{i=1}^{m} E[Y_i(\mathbf{a}_{-i}, \tilde{a})]$. The total effect compares an individual's counterfactual at one treatment level in a block that receives one intervention to his counterfactual at a different treatment level in a block that receives the another intervention: $TE_i(\mathbf{a}, \mathbf{a}'; \tilde{a}, \bar{a}) \equiv E[Y_i(\mathbf{a}_{-i}, \tilde{a})] - E[Y_i(\mathbf{a}'_{-i}, \bar{a})]$. The average total effect is defined analogously to the other average effects above. Hudgens and Halloran (2008) showed that the total effect can be decomposed into a sum of unit-level and spillover effects: $TE_i(\mathbf{a}, \mathbf{a}'; \tilde{a}, \bar{a}) = E[Y_i(\mathbf{a}_{-i}, \tilde{a})] - E[Y_i(\mathbf{a}_{-i}, \bar{a})] + E[Y_i(\mathbf{a}_{-i}, \bar{a})] - E[Y_i(\mathbf{a}'_{-i}, \bar{a})] = DE_i(\mathbf{a}; \tilde{a}, \bar{a}) + IE_i(\mathbf{a}, \mathbf{a}'; \bar{a})$.

Sobel (2006), Hudgens and Halloran (2008), VanderWeele and Tchetgen Tchetgen (2011b) and Tchetgen Tchetgen and VanderWeele (2012) proposed ways to estimate and extend unit-level, spillover, total and overall effects. We will not discuss these extensions here except to note that they require the same three identifying assumptions (8), (9) and (10).

2.2 Path Specific Effects of $\mathbf{A}$ on Y

In Sections 4.2 and 4.3, we describe path-specific effects that may be of interest in certain interference contexts; here we review and extend the literature on path-specific effects under interference. Let $\mathbf{M}_k \equiv (M_{k1}, \dots, M_{km})$ be a vector of variables that may lie on a causal pathway from $\mathbf{A}_k$ to Y_{ki}. VanderWeele (2010) provided identifying assumptions and expressions for mediated effects with cluster-level treatments. These effects are applicable to the context of partial interference where there is interference by the mediator but not by the treatment (M_{ki} may have an effect on Y_{kj} for $i \neq j$, but $\mathbf{A}_k$ is set at the cluster level, and thus is the same for all individuals in block k). We adapt them here to accommodate interference by the treatment in addition to the mediator. We make the consistency assumptions that $\mathbf{M}_k(\mathbf{a}_k) = \mathbf{M}_k$ when $\mathbf{A}_k = \mathbf{a}_k$, that $Y_{kj}(\mathbf{a}_k, \mathbf{m}_k) = Y_{kj}$ when $\mathbf{A}_k = \mathbf{a}_k$ and $\mathbf{M}_k = \mathbf{m}_k$, and that $Y_{kj}(\mathbf{a}_k, \mathbf{M}_k(\mathbf{a}_k)) = Y_{kj}(\mathbf{a}_k)$. The expected controlled direct effect of a block-level treatment $\mathbf{A}_k$ on individual i's outcome, not through $\mathbf{M}_k$, is defined as $E[Y_i(\mathbf{a}, \mathbf{m})] - E[Y_i(\mathbf{a}', \mathbf{m})]$; it measures the expected change in Y_{ki} due to a change in $\mathbf{A}_k$, intervening to set $\mathbf{M}_k$ to $\mathbf{m}_k$. The expected natural direct effect is $E[Y_i(\mathbf{a}, \mathbf{M}(\mathbf{a}))] - E[Y_i(\mathbf{a}', \mathbf{M}(\mathbf{a}))]$; it measures the expected change in Y_{ki} due to a change in $\mathbf{A}_k$, holding $\mathbf{M}_k$ fixed at $\mathbf{M}_k(\mathbf{a}_k)$. The expected natural indirect effect of $\mathbf{A}_k$ on Y_{ki} through $\mathbf{M}_k$, given by $E[Y_i(\mathbf{a}', \mathbf{M}(\mathbf{a}))] - E[Y_i(\mathbf{a}', \mathbf{M}(\mathbf{a}'))]$, measures the expected change in Y_{ki} when $\mathbf{A}_k$ is fixed but $\mathbf{M}_k$ changes from $\mathbf{M}_k(\mathbf{a}_k)$ to $\mathbf{M}_k(\mathbf{a}_k')$. Average controlled direct, natural direct, and natural indirect effects are defined similarly to the average effects in Section 2.1: we average the counterfactuals within each block before taking the expectations over blocks. These natural direct and indirect effects are identifiable under the following four assumptions:

$$Y_{ki}(\mathbf{a}_k, \mathbf{m}_k) \perp\!\!\!\perp \mathbf{A}_k | \mathbf{C}_k, \tag{11}$$

$$Y_{ki}(\mathbf{a}_k, \mathbf{m}_k) \perp\!\!\!\perp \mathbf{M}_k | \mathbf{A}_k, \mathbf{C}_k, \tag{12}$$

$$\mathbf{M}_k(\mathbf{a}_k) \perp\!\!\!\perp \mathbf{A}_k | \mathbf{C}_k \tag{13}$$

and

$$Y_{ki}(\mathbf{a}_k, \mathbf{m}_k) \perp\!\!\!\perp \mathbf{M}_k(\mathbf{a}_k') | \mathbf{C}_k. \tag{14}$$

Assumptions (11), (12) and (13) correspond, respectively, to the absence of unblocked backdoor paths from $\mathbf{A}_k$ to Y_{ki} (except possibly through $\mathbf{M}_k$) conditional on $\mathbf{C}_k$, from $\mathbf{M}_k$ to Y_{ki} conditional on $\mathbf{A}_k$ and $\mathbf{C}_k$, and from $\mathbf{A}_k$ to $\mathbf{M}_k$ conditional on $\mathbf{C}_k$. Assumption (14), similar to (6), corresponds to the recanting witness criterion. Under these assumptions, counterfactual expectations of the form $E[Y_i(\mathbf{a}', \mathbf{M}(\mathbf{a}))]$ are identified and, therefore, so are the natural direct and indirect effects, which are contrasts of such expectations. Specifically, $E[Y_i(\mathbf{a}', \mathbf{M}(\mathbf{a}))]$ is identified by

$$\sum_{\mathbf{c}} \sum_{\mathbf{m}} E[Y_i | \mathbf{A} = \mathbf{a}', \mathbf{M} = \mathbf{m}, \mathbf{C} = \mathbf{c}] \\ \cdot P(\mathbf{M} = \mathbf{m} | \mathbf{A} = \mathbf{a}, \mathbf{C} = \mathbf{c}) P(\mathbf{C} = \mathbf{c}).$$

Assumptions (11) and (12) suffice to identify the controlled direct effect.

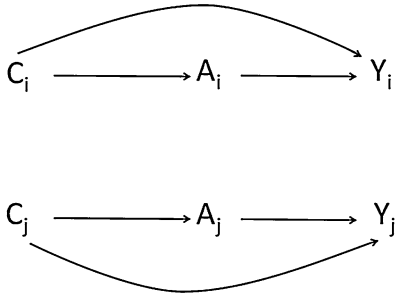

FIG. 3.

3. COVARIATE CONTROL

Although the subscripts are usually suppressed, under the assumption of no interference the standard DAG for the effect of a treatment A on an outcome Y with confounders C is drawn to show the relationships among Y_i, A_i and C_i for subject i. Under interference, however, it is not sufficient to consider causal pathways at the individual level; a causal DAG must depict an entire block. For simplicity, we will focus on blocks of the smallest size that preserves the essential structure of interference, which for our purposes will be two or three. The principles extend to groups of any size, but the DAGs become considerably more complex as the blocks grow. The DAG for the effect of A on Y in a group of size two with no interference is depicted in Figure 3. In what follows, we represent a single block of subjects on each DAG, and we therefore suppress the subscript k indicating membership in block k.

Interference can be represented by a DAG like the one given in Figure 4. The arrows from A_i to Y_j for $i \neq j$ represent the effect that one individual's treatment has on another's outcome. This representation suffices whenever contrasts of counterfactuals of the form $Y(\mathbf{a})$, such as the effects described in Section 2.1, are the only effects of interest. However, as we will see below, when contagion or allocational interference are present, such a diagram does not represent information about how the effect of A_i on Y_j operates. In Section 4, we describe how to represent this additional

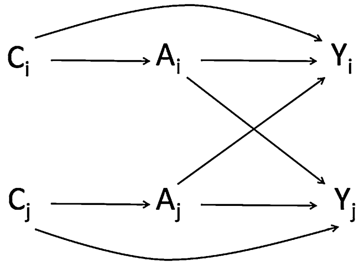

FIG. 4.

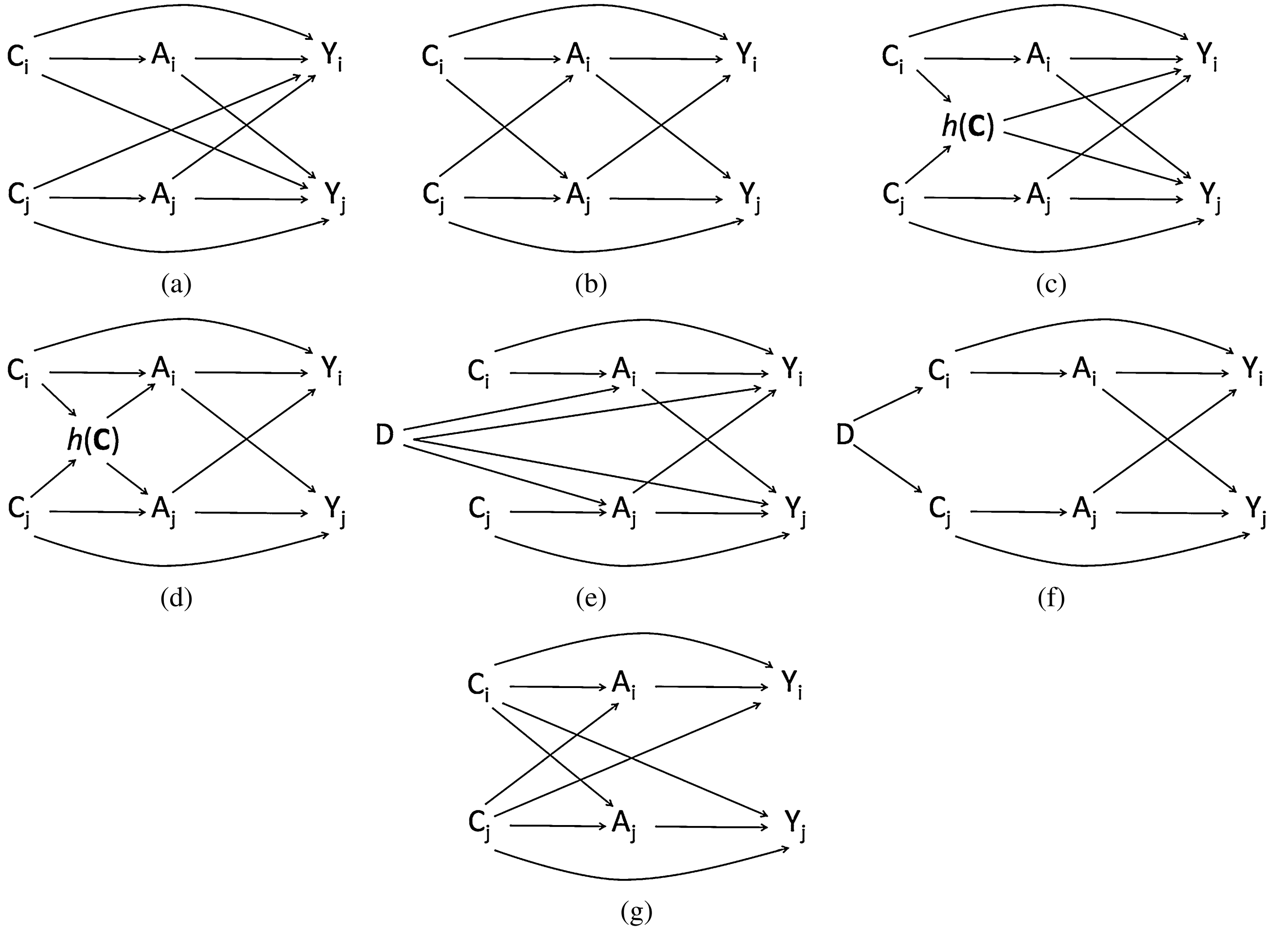

FIG. 5.

information on a DAG. We describe covariate control in the general cases depicted in the DAGs in Figures 4 and 5 before moving on in Section 4 to tease apart the structures that make direct interference, interference by contagion, and allocational interference distinct. The principles of covariate control in the presence of interference are straightforward: like in the case of no interference, they follow from the fact that all backdoor paths from treatment to outcome must be blocked by a measured set of covariates. However, without taking the time to draw the operative causal DAG with interference it is easy to make mistakes, like controlling only for individual-level covariates when block-level covariates are necessary to identify the causal effect of interest. Below we will consider a number of different settings and causal structures, and discuss in each whether control for only an individual's covariates suffices to identify causal effects or whether control for the covariates of the entire block (or of some summary) is needed.

If the individuals in the block share no common causes of A or Y, as in the DAG in Figure 4, then C_i suffices to block the backdoor paths from A_i to Y_i and from A_j to Y_i and, therefore, exchangeability for the effect of $\mathbf{A}$ on Y_i holds conditional on C_i. That is, $Y_i(a_i, a_j) \amalg \mathbf{A}|C_i$ for all i. If C_j is a direct cause of Y_i for $i \neq j$, as in Figure 5(a), then exchangeability for the effect of $\mathbf{A}$ on Y_i necessitates block- and not just individual-level covariates. Even if each individual's treatment is randomized conditional on his own covariates (this corresponds to the absence of arrows C_j to A_i for $i \neq j$ on the DAG), there is still a backdoor path from A_j to Y_i via C_j and, somewhat counterintuitively, it is necessary to control for C_j in addition to C_i in any model for the effect of $\mathbf{A}$ on Y_i. On the other hand, if C_j directly affects A_i but not Y_i for $i \neq j$, as in Figure 5(b), then for exchangeability for the effect of $\mathbf{A}$ on Y_i it suffices to condition only on C_i. If, in addition to C_i, a function $h(\mathbf{C})$ of the vector of covariates influences outcome (Figure 5(c); for example, the mean value of **C** for the block), then C_i and either $h(\mathbf{C})$ or C_j need to be conditioned on in order to achieve exchangeability for the effect of **A** on Y_i. If $h(\mathbf{C})$ only influences treatment assignment (Figure 5(d)), then only C_i must be conditioned on. If a block-level characteristic D is a common cause of A and Y (Figure 5(e)), then C_i and D must be conditioned on in order to achieve exchangeability for the effect of **A** on Y_i. If C_i and C_j share a common cause (Figure 5(f)), then exchangeability for the effect of **A** on Y_i holds conditional on C_i.

Even in the absence of interference for the effect of **A** on **Y**, there are scenarios in which individual-level covariates do not suffice to control for the effect of an individual's treatment on his own outcome. For example, the DAG in Figure 5(g) depicts a scenario in which one individual's covariates affect another individual's treatment and outcome (represented by the arrows $C_i \to A_j$ and $C_i \to Y_j$), but there is no effect of one individual's treatment on another's outcome (no directed path from A_i to Y_j). In other words, there is interference for the effect of **C** on **Y** but not for the effect of **A** on **Y**. Vansteelandt (2007) noted that in this setting it is necessary to condition on **C** to achieve exchangeability for the effect of A_i on Y_i.

Consider the DAG in Figure 4, but now suppose that **C** is unobserved. As we discussed above, C_i is a confounder of the effect of A_i on Y_i, but in this DAG the effect of A_j on Y_i is unconfounded (there is no backdoor path from A_j to Y_i). If a researcher hypothesizes that the DAG in Figure 4 represents the underlying causal structure in a particular setting but he does not have access to data on the confounders **C**, then the effect of **A** on Y_i is not identified. However, the unconfounded effect of A_j on Y_i is identified by $E[Y_i|A_j = a_j]$. This quantity has an interpretation in the interference setting as the weighted average of expected counterfactuals within strata of C.

$$\begin{aligned}
&E[Y_i|A_j = a_j] \\
&\quad = \sum_{a_i}\sum_{c} E[Y_i|A_i = a_i, A_j = a_j, C_i = c] \\
&\qquad\qquad \cdot P(A_i = a_i|C_i = c)P(C_i = c) \\
&\quad = \sum_{a_i}\sum_{c} E\bigl[Y_i(a_i, a_j)|A_i = a_i, A_j = a_j, C_i = c\bigr] \\
&\qquad\qquad \cdot P(A_i = a_i|C_i = c)P(C_i = c) \\
&\quad = \sum_{a_i}\sum_{c} E\bigl[Y_i(a_i, a_j)|C_i = c\bigr] \\
&\qquad\qquad \cdot P(A_i = a_i|C_i = c)P(C_i = c),
\end{aligned}$$

where the first equality relies on the facts that $A_i \amalg A_j|C_i$ and $C_i \amalg A_j$, the second relies on consistency, and the third on conditional exchangeability. Alternatively, this quantity has the interpretation of $E[Y_i(a_j)]$ in an experiment where Y_i is considered to be the only outcome, A_j is the treatment of interest and is intervened on, and A_i is randomly assigned according to the actual distribution $P(A_i|C_i)$ in the population.

The hypothetical experiment described above points toward a possible strategy for estimating the effect of one component of **A** on Y_i when the confounders of the effect of **A** on Y_i are not fully observed. The researcher can analyze each block of subjects as a single observation, with a single treatment and outcome. This strategy discards data on others' treatments and outcomes but may allow for progress even if the full set of covariates needed to identify $Y_i(\mathbf{a})$ are not observed.

In some of the DAGs in Figure 5, identification of the effect of A_j on Y_i requires fewer covariates than the effect of **A** on Y_i. In Figure 5(c), C_j suffices to control for confounding of the effect of A_j on Y_i even though it does not suffice for the effect of **A** on Y_i. For the DAG in Figure 5(e), D suffices to control for confounding of the effect of A_j on Y_i.

In some cases, we can identify the effect of A_i on Y_i with fewer covariates than are required to identify the effect of **A** on Y_i. In Figures 5(a) and 5(c), we can identify the effect of A_i on Y_i if only C_i is observed, even though we cannot identify the effect of **A** on Y_i without also observing C_j (or $h(\mathbf{C})$). In the Appendix, we give the identifying expressions for these effects.

For the DAGs in Figures 5(b)–5(f), the entire vector **C** is not necessary to identify the effect of **A** on Y_i, though it would in general be necessary in order to jointly identify the effects of **A** on Y_i and on Y_j (i.e., the effect of **A** on **Y**). This is because in order to identify the effect of **A** on **Y** we require conditional exchangeability to hold for all subjects. In these settings, if **C** is not fully observed but certain components or functions of **C** required to identify the effect of **A** on Y_i are, then we can proceed by considering Y_i to be the only outcome in each block. We can still consider the full vector of treatments **A**, and therefore identify unit-level, spillover, total and overall effects of **A** on Y_i.

4. THREE DISTINCT TYPES OF INTERFERENCE

By understanding the causal mechanisms underlying interference, we can more precisely target effects of interest. There are three distinct causal pathways by which one individual's treatment may affect another's outcome. All fall under the rubric of interference. The distinction among them has generally not been made, but they differ in causal structure, effects of interest and requirements for identification of effects. Often more than one type of interference will be present simultaneously.

The first pathway by which interference may operate is a direct causal effect of one individual's treatment on another individual's outcome, unmediated with respect to the first individual's outcome. We call this *direct interference*. As an example, suppose that the outcome is obesity and the treatment is dietary counseling from a nutritionist. An individual who receives treatment can in turn "treat" his associates by imparting to them the information gained from the nutritionist; therefore, if individual i receives treatment and individual j does not, individual j may be nevertheless be exposed to the treatment of individual i and his or her outcome will be affected accordingly.

A second pathway by which one individual's treatment may affect another individual's outcome is via the first individual's outcome. For example, if the outcome is an infectious disease and the treatment is a prophylactic measure designed to prevent disease, then the treatment of individual i may affect the outcome of individual j by preventing individual i from contracting the disease and thereby from passing it on. We call this type of interference *interference by contagion*. It is differentiated from direct interference by the fact that it does not represent a direct causal pathway from the exposed individual to another individual's outcome, but rather a pathway mediated by the outcome of the exposed individual.

The third pathway for interference is *allocational interference*. Treatment in this setting allocates individuals to groups; through interactions within a group individuals' characteristics may affect one another. An example that often arises in the social science literature is the allocation of children to schools or of children to classrooms within schools (Angrist and Lang, 2004; Graham, Imbens and Ridder, 2010; Hong and Raudenbush, 2008). The performance and behavior of student i may affect the performance and behavior of student j in the same class, for example, by distracting or motivating student j or by occupying the teacher's attention. Another example that can be seen as allocational interference is the effect of college enrollment on wage differences for college- versus high-school-educated workers, where the wage difference depends on the proportion of workers in each education category (Heckman, Lochner and Taber, 1998).

4.1 Direct Interference

Direct interference is present when there is a causal pathway from one individual's treatment to another individual's outcome, not mediated by the first individual's outcome. Interference can be direct with respect to a particular outcome but not another. Consider two individuals living in the same household, each randomized to an intervention designed to prevent high cholesterol. Suppose the intervention consists of cooking classes, nutritional counseling and coupons that can be redeemed for fresh produce, and consider a household in which one individual is treated and one untreated. The treated individual could bring fresh produce into the household, prepare healthy meals, and talk about the nutritionist's counsel, thereby exposing the other individual to a healthier diet. If the outcome of interest is a measure of blood cholesterol level, then this is an example direct interference: the untreated individual is exposed to the treated individual's diet and that exposure reduces the untreated individual's cholesterol. On the other hand, if the outcome is a measure of healthy diet and behavior, then the same story depicts contagion rather than direct interference: the treated individual adopts a healthier diet which results in the untreated individual also adopting a healthier diet. Diet may spread by contagion; cholesterol presumably would not.

In many settings, direct interference and contagion will be present simultaneously for the same outcome. For example, suppose that in the story above the outcome were weight change. Then it is possible that the treated individual's family member could lose weight both because of exposure to healthier foods (direct interference) and because he was motivated by the weight loss of his relative (contagion).

Direct interference has the simplest causal structure of the three types of interference. In addition to a direct causal path from A_i to Y_i, there is also a direct path from A_i to Y_j for all pairs (i, j) such that subjects i and j are in the same block. Direct interference in a block of size two is depicted in the DAGs in Figures 4 and 5, with the exception of Figure 5(g). Because there is only a single path from A_i to Y_j for any pair i, j, differences between counterfactuals of the form $Y_i(\mathbf{a})$ capture all of the causal effects of $\mathbf{A}$ on Y_i and, therefore, effects like the total, unit-level, spillover and overall effects described in Section 2.1 summarize the causal effects of $\mathbf{A}$ on Y_i.

4.2 Interference by Contagion

Interference by contagion often has a complex causal structure, because it can involve feedback among different individuals' outcomes over time. The causal structure of the effect of A_i on Y_i is straightforward: A_i has a direct protective effect on Y_i, represented by a direct arrow from A_i to Y_i on the DAG. The effect of A_i on Y_j is considerably more complex. It is tempting to represent the effect of A_i on Y_j as a mediated effect through Y_i, but this cannot be correct, as Y_i and Y_j are contemporaneous and, therefore, one cannot cause the other. The effect of A_i on Y_j is mediated through the evolution of the outcome of individual i; this complicated structure is depicted in the DAG in Figure 6, where Y_i^t represents the outcome of individual i at time t, T is the time of the end of follow-up, and the dashed arrows represent times 4 through $T - 1$, which do not fit on the DAG (but which we assume were observed). The unit of time required to capture the causal structure depends on the nature of transmission of the outcome; it should be the case that the probability of one individual's outcome affecting another's is unaltered by differences in timing on a scale smaller than the units used.

In order to further explicate the structure, we consider the case of an infectious disease like the flu. Infectious diseases are paradigmatic examples of contagion. Halloran and Struchiner (1995), Hudgens and Halloran (2008) and VanderWeele and Tchetgen Tchetgen (2011a) have written about identification and estimation of overall, unit-level, spillover and total effects for vaccinations against infectious diseases, and we follow up with this literature in Section 4.2.1. Although we illustrate the principles of interference by contagion through the lens of infectious diseases, this type of interference can occur in many and diverse settings: an educational intervention assigned to one student could affect that student's performance, which in turn might affect the performance of her classmates; a get-out-the-vote mailing could motivate its recipients to decide to vote, and communicating that decision to friends could change the friends' voting behavior. The principles that we discuss below apply to any of these settings.

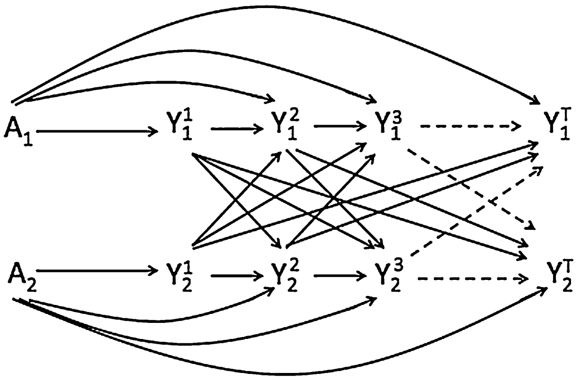

FIG. 6.

Suppose that the flu vaccine has a protective effect against the flu by preventing or shortening the duration of episodes of the flu for some individuals. Let A be an indicator of getting the flu vaccine before the start of a six-month long flu season, and let Y be the total number of days spent infectious with the flu over the course of the season. In the DAG in Figure 6, Y_i^t represents the flu status of individual i at time t (measured in days), $T \equiv 180$ is the day of the end of flu season, and the dashed arrows represent days 4 through $T - 1$, which do not fit on the DAG (but which we assume were observed). Let Y_i^t be the total number of days spent infectious up to and including day t. (Note that, when Y_i^t is observed for all t, this is equivalent to coding it as an indicator of individual i being infectious at time t.) In choosing days as the unit of time, we are making the assumption that the probability of one individual infecting another is not affected by a difference of a fraction of a day in flu duration.

We will rarely have fine-grained information on the evolution of the outcome over time. In the rest of this section, we describe how to draw the appropriate causal DAGs and how to identify causal effects in such cases. Drawing a causal DAG using only a subset of relevant variables has been extensively studied in the graphical models literature and involves an operation known as projection (Pearl and Verma, 1994; Tian and Pearl, 2002b; Verma, 1993). Projection algorithms are somewhat technical; below we provide an intuitive discussion of the construction of causal DAGs when not all variables relevant to contagion are observed.

If we only observe the outcome (cumulative days of the flu) at the end of the season, then, as in the DAG in Figure 7, we replace the collection of all unobserved variables (i.e., Y_i^t for $t < T$) with U. Without additional assumptions, we cannot replace the two diagonal pathways through U with direct arrows from A_1 to Y_2^T and from A_2 to Y_1^T; that would imply that

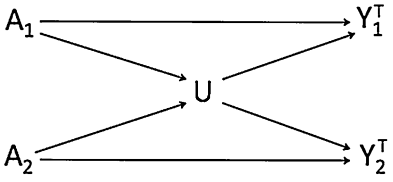

FIG. 7.

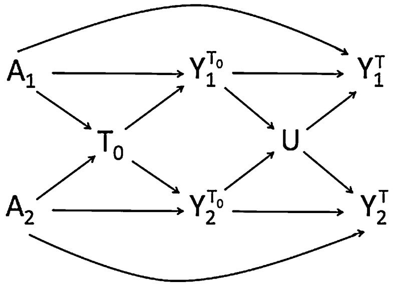

FIG. 8.

$Y_1^T \amalg Y_2^T | A_1, A_2$, which is shown to be false by the DAG in Figure 6. If we know who gets the flu first, then the DAG in Figure 8 represents the causal relations among the observed variables, where T_0 is the time of the first case of the flu. The unmeasured variable U cannot be omitted because Y_1^T is not independent of Y_2^T conditional on $\{Y_1^{T_0}, Y_2^{T_0}\}$; Y_1^T depends on the number and timing of individual 2's illnesses between time T_0 and time T. It might seem as though $Y_1^{T_0}$ should be independent of $Y_2^{T_0}$ conditional on $\{A_1, A_2\}$ in this scenario, because there can be no contagion before the first case of the flu. But this is not the case: $Y_i^{T_0}$ can be thought of as an indicator that individual i gets the flu before or at the same time as individual j, and this is dependent on individual j remaining healthy through time $T_0 - 1$. On the other hand, conditioning on the time of the first case of the flu renders $Y_1^{T_0}$ and $Y_2^{T_0}$ independent, because conditioning on T_0 is tantamount to conditioning on both individuals remaining healthy until the time of the first case, that is, on their entire flu histories up to time T_0.

Suppose information on the number but not duration of cases of the flu is available. Then we could define Y^t to be the number of distinct cases initiated by time t. However, this outcome fails to capture all of the relevant information about an individual's flu status, because a case of the flu that lasts ten days may be more contagious than one that lasts five days. Therefore, there may be an effect of A_i on Y_j^t that is not mediated by Y_j^s, $s < t$, being instead mediated by the duration of individual i's flu incidents. This is represented on the DAG in Figure 9 by an arrow from A_i to Y_j^t. These arrows encode the fact that Y_j^t is dependent on Y_i^t even conditional on $\{Y_i^s, Y_j^s\}$ for all $s < t$. Similarly, if the outcome on the DAG in Figure 8 were the total number of flu episodes by the end of the season instead of the cumulative days of flu we would add arrows from A_i to $Y_j^{T_0}$ and to Y_j^T.

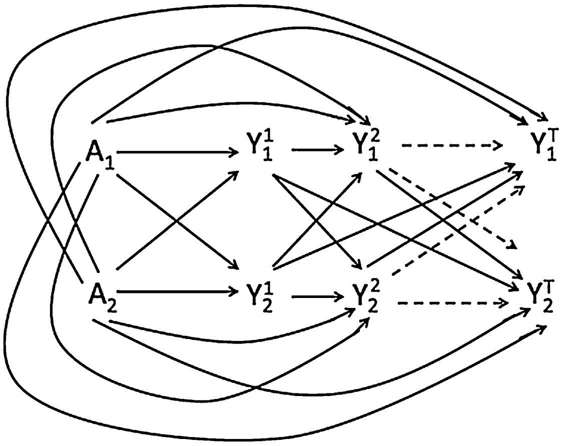

FIG. 9.

If there are common causes C of treatment and outcome for each individual, as in the DAG in Figure 10, then exchangeability for the effect of $\mathbf{A}$ on Y_i^T will hold conditional on C_i. If there are common causes of treatments for different individuals, or of treatments and outcomes across individuals, then exchangeability requires conditioning on them. The same conclusions about exchangeability hold if we observe the outcome only at select time points.

The overall, unit-level, spillover and total effects defined in Section 2.1 do not distinguish among the multiple distinct causal pathways from A_i to Y_j^T. We discuss estimation of path-specific effects below.

4.2.1 *Contagion and infectiousness*. Recently, Vanderweele, Tchetgen and Halloran (2012) described the decomposition of the spillover effect, that is the ef-

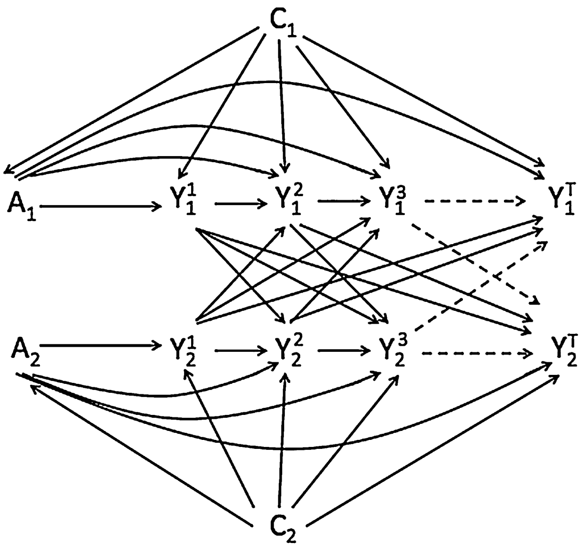

FIG. 10.

fect of A_i on Y_j, into a "contagion effect" and an "infectiousness effect." The contagion effect is the protective effect that vaccinating one individual has on another's disease status by preventing the vaccinated individual from getting the disease and thereby from transmitting it, similar to the effect we discussed above. In order to illustrate the infectiousness effect, consider the following refinement to our example. Suppose that a vaccine exists for the flu that prevents the disease for some individuals and also makes some cases of the disease among vaccinated individuals less likely to be transmitted. Let A be an indicator of vaccination before the start of flu season and let Y^t be the total number of episodes of flu up to and including day t. Then A_i may have an effect on Y_j^t even if it has no effect on Y_i^t, that is, even if individual i would get the flu whether vaccinated or not, by preventing individual i from transmitting the flu to individual j. This infectiousness effect represents a pathway that is distinct from the contagion effect because it does not operate through the infection status of the vaccinated individual. The infectiousness effect has the structure of a direct effect by which individual i's vaccination confers improved protection against individual i's flu on individual j; it represents a type of direct interference. This is similar to the example above in which the duration of flu episodes was unobserved: infectiousness, like flu duration, is a property of the infected individual's disease state, but if it is not captured by the outcome measure then it has the structure of a direct effect of A_i on Y_j. The contagion and infectiousness effects are not identifiable without strong assumptions or carefully conceived data collection. When they are identifiable their sum (or product if they are defined on the multiplicative scale) is equal to the total spillover effect of A_i on Y_j.

Vanderweele, Tchetgen and Halloran (2012) defined the contagion and infectiousness effects as the natural indirect and direct effects, respectively, of A_1 on Y_2^T with $Y_1^{T_0}$ as the mediator, where T_0 is the day of the first flu infection and T is the day of the end of follow-up, for example, the last day of the flu season. That is, the contagion effect is defined as $E[Y_2^T(0, Y_1^{T_0}(1))] - E[Y_2^T(0, Y_1^{T_0}(0))]$ and infectiousness by $E[Y_2^T(1, Y_1^{T_0}(1))] - E[Y_2^T(0, Y_1^{T_0}(1))]$. For simplicity and consistency with the existing literature we adopt the setting used in VanderWeele and Tchetgen Tchetgen (2011a), Vanderweele, Tchetgen and Halloran (2012) and Halloran and Hudgens (2012): each block is a group of size two who share a household, and in each pair individual 1 is randomized to vaccine while individual 2 is always unvaccinated. If, as those authors assumed, disease status is observed only at the end of follow-up, Figure 11 depicts the operative DAG. Although there is an unmeasured variable on the path from A_1 to Y_2^T, it is not a confounder and we can identify the effect of A_1 on Y_2^T. However, we cannot identify the component contagion and infectiousness effects without observing the mediator $Y_1^{T_0}$. In order to circumvent the problem of the unobserved mediator, Vanderweele, Tchetgen and Halloran (2012) assumed that each individual can be infected only once and that individual 2 can only be infected by individual 1, as would be the case if individual 2 were homebound. These assumptions dramatically simplify the causal structure by ensuring that individual 1 is infected first and that there is no feedback between the time of first infection and the end of follow up. Then Y_1^T must be equal to $Y_1^{T_0}$, and thus $Y_1^{T_0}$ is observed. These assumptions are encoded by the DAG in Figure 12. Now the contagion and infectiousness effects can be identified as the natural indirect and direct effects of A_1 on Y_2^T mediated by $Y_1^T = Y_1^{T_0}$, as long as assumptions (4) through (7) are met. Three of these assumptions correspond to the absence of any unmeasured confounding of the relationships between A_1 and Y_1^T, between Y_1^T and Y_2^T conditional on A_1, and between A_1 and Y_2^T, respectively. These assumptions can be made conditional on measured covariates $\mathbf{C}$. Assumption (6) is the recanting witness criterion.

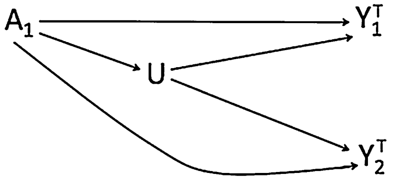

FIG. 11. *This DAG corresponds to a household of size two in which individual* 2 *is always unvaccinated and disease status is assessed at the end of follow-up. U represents unmeasured variables.*

The simplifying assumption that the outcome can occur at most once may be reasonable for many infectious diseases. More limiting is the assumption that individual 2 can only be infected by individual 1. Here, we

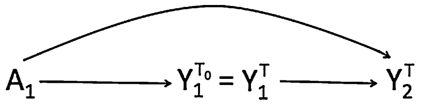

FIG. 12. *This DAG corresponds to the same setting as Figure* 11, *but under the assumptions that individual* 2 *can only be infected by individual* 1 *and that only one event is possible per subject during follow-up.*

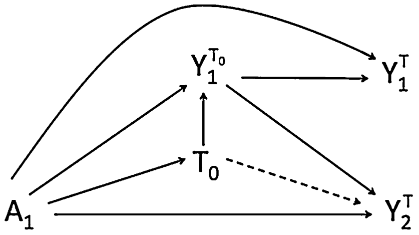

FIG. 13. *This DAG corresponds to the same setting as Figure 12, but without the assumption that individual 2 can only be infected by individual 1. The dashed arrow is present when T is defined as the end of the flu season; it is absent when T is defined as $T_0 + s$.*

will describe settings in which it may be possible to relax this assumption. Even if we observe the time and identity of the first case of the flu in addition to the outcome at the end of follow-up, relaxing this assumption makes identification of the contagion and infectiousness effects impossible. In the DAG in Figure 13, $Y_1^{T_0}$ is an indicator of whether individual 1 was infected first. Although it is not straightforward to imagine intervening on T_0, the time of the first infection, we include this variable in the DAG in order to ensure that the DAG encodes the true conditional independence statements. The presence of the arrow from T_0 to Y_2^T makes T_0 a recanting witness for the contagion and infectiousness effects: it is a confounder of the mediator–outcome relation that is caused by treatment. This arrow is necessitated by the fact that T_0 predicts Y_2^T even conditional on A_1 and $Y_1^{T_0}$. To see this, imagine two different pairs in which individual 1 is vaccinated and gets sick first ($A_1 = Y_1^{T_0} = 1$). Suppose that in one pair, the vaccinated individual gets sick at the very end of the follow-up period ($T_0 = T - 1$). Then the probability that the second individual gets the flu after time T_0 but before time T is very small. Suppose that in the other pair the vaccinated individual gets sick on the first day of the flu season ($T_0 = 1$). The probability that we observe the second individual in this pair to get sick before the end of follow-up is much higher.

One possible solution to the recanting witness problem is to let T_0 determine a new artificial end of follow-up, so that the amount of time between T_0 and T is constant over different values of T_0. In particular, if we know that an infected individual is infectious for up to s days after becoming symptomatic, then we can let $T = T_0 + s$ and collect data on $Y_1^{T_0}$, $Y_1^{T_0+s}$ and $Y_2^{T_0+s}$. If neither individual in the pair is observed to get the flu then $T_0 = T =$ *the last day of the flu season*. Setting the artificial end of follow-up to lag behind the time of first infection by s days ensures that we will observe $Y_1^{T_0} = Y_2^{T_0+s} = 1$ for any pair in which individual 2 catches the flu from individual 1. We throw away data on pairs for which the first infection occurs fewer than s days before the end of the flu season, but if s is small then it may be reasonable to assume that any resulting bias is negligible.

One further assumption is required in order for $Y_2^{T_0+s} \perp\!\!\!\perp T_0 | Y_1^{T_0}, A_1$, which is the conditional independence assumption that licenses the omission of an arrow from T_0 to $Y_2^{T_0+s}$. Suppose that cumulative exposure to the flu virus makes people, on average, less susceptible to infection as the flu season progresses due to acquired immunity. Then, for pairs in which individual 1 is vaccinated and gets sick first, individual 2 is less likely to catch individual 1's flu later in the season as compared to earlier (for larger values T_0 compared to smaller values). This violates $Y_2^{T_0+s} \perp\!\!\!\perp T_0 | Y_1^{T_0}, A_1$. If we assume that the probability of individual 2 catching the flu if exposed on day t is constant in t, then $Y_2^{T_0+s}$ is independent of T_0 conditional on A_1 and $Y_1^{T_0}$. Therefore, T_0 is not a recanting witness (it is still caused by treatment but is no longer a confounder of the mediator–outcome relation). Assuming that exchangeability assumptions (4), (5) and (7) hold (see Vanderweele, Tchetgen and Halloran, 2012 for discussion of their plausibility in this context), the contagion and infectiousness effects of A_1 on $Y_2^{T_0+s}$ are identifiable. The contagion effect is given by $E[Y_2^{T_0+s}(0, Y_1^{T_0}(1))] - E[Y_2^{T_0+s}(0, Y_1^{T_0}(0))]$ and infectiousness by $E[Y_2^{T_0+s}(1, Y_1^{T_0}(1))] - E[Y_2^{T_0+s}(0, Y_1^{T_0}(1))]$. The spillover effect of A_1 on $Y_2^{T_0+s}$ on the additive scale is the sum of the contagion and infectiousness effects.

If both individuals are randomized to vaccination, then A_2 is a confounder of the relationship between $Y_1^{T_0}$ and $Y_2^{T_0+s}$. Assuming A_2 is observed, this does not pose a problem for identification of the contagion and infectiousness effects. Generalizations of the contagion and infectiousness effects to blocks of size greater than two are also possible (Vanderweele, Tchetgen and Halloran, 2012).

An alternative definition of an infectiousness effect, proposed by Vanderweele, Tchetgen and Halloran (2012), is the controlled direct effect of A_1 on Y_2^T, holding $Y_1^{T_0}$ fixed at 1. Identification of this effect does not require the recanting witness criterion to hold and, therefore, it is identifiable when the end-of-follow up is fixed and does not depend on T_0. A disadvantage of this controlled direct infectiousness effect is that it does not admit a decomposition of the indirect effect of A_1 on Y_2^T. That is, if we subtract the controlled direct infectiousness effect from the total effect of A_1 on Y_2^T, the remainder cannot be interpreted as a contagion effect.

4.3 Allocational Interference

In allocational interference, an individual is allocated to a group and his outcome is affected by which individuals are allocated to the same group. In many real settings, group allocation is not random, rather individuals select their own group or are assigned based on previously observed characteristics. One example of random group allocation is the assignment of college freshman to dorms or dorm rooms (Carrell, Fullerton and West, 2009; Sacerdote, 2000). We differentiate between allocational interference and other scenarios in which individuals already in groups are assigned to receive individual or group-level treatments. In the former setting, an individual's outcome depends on the specific composition of his group, while in the latter it depends on treatment assignments for his group but not necessarily on group composition. Of course, these two phenomena often occur in tandem, as they would if children were assigned to classrooms which were then assigned different educational interventions.

Similarly, contagion is often present in conjunction with allocational interference. For example, in the allocation of children to classrooms there is likely to be feedback among children who are in the same classroom in terms of achievement, attitude and tendency to act out. Therefore, any measure of behavior or achievement that can evolve over time would likely be subject to contagion. An outcome like end-of-the-year test scores, on the other hand, would evince allocational interference without contagion as one student's test score cannot directly affect another's (though we would likely still envision contagion with respect to knowledge or learning).

Allocational interference is perhaps the most complicated of the three types of interference. We first describe how to represent basic allocational interference on a DAG. Then we introduce a toy example and use it to illustrate some additional DAG structures and to briefly discuss causal effects that may be of interest in the presence of allocational interference. This discussion is far from exhaustive, but we hope that this section will serve as a guide for how to think about allocational interference.

Allocational interference assigns subjects to groups within each block. Recall that blocks of individuals are independent from one another, and that interference is possible within but not between blocks. This is not the case for groups; interference will generally be present across groups within the same block. In the school example, blocks might be different schools and groups might be classrooms within each school. Suppose that within each block there are L possible groups to which an individual can be allocated. Then treatment is a categorical variable, A, which for each subject takes on the value $l \in (1, \dots, L)$ of the group to which that individual was assigned. As in Section 2, we let k index N blocks with m individuals in each block. Let $\mathbf{A}_k$ be the vector of group assignments in block k.

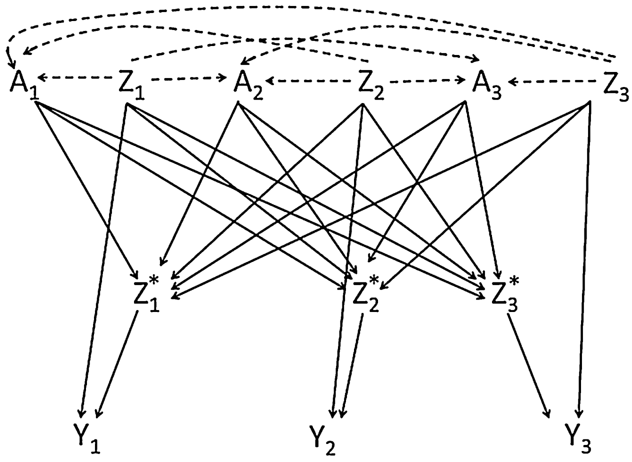

FIG. 14.

Figure 14 provides a DAG for a scenario with allocational interference and $m = 3$. Within each block, we allocate the subjects into two distinct groups and each individual's outcome may be affected by who is in each group. The DAG depicts a single block and we therefore suppress the subscript k. Let Y_i be the outcome for individual i and Z_i be a vector of all baseline characteristics that affect the outcome of individual i or the outcomes of the other individuals with whom he comes into contact. Define Z_i^* to be an $(m-1)$-dimensional array with jth element equal to $Z_j \times I(A_j = A_i)$, $j \neq i$, that is, the baseline covariates of subject j multiplied by the indicator that individual j is assigned to the same group as individual i. In addition to the individual level causal arrows from Z_i and Z_i^* into Y_i, we require arrows from Z_i to Z_j^* and from A_i to Z_j^* for all pairs (i, j), $i \neq j$. This is because Z_j^* is by definition a function of $\mathbf{A}$ and of Z_i for all $i \neq j$. Randomized allocation corresponds to the absence of the dashed arrows into $\mathbf{A}$. Otherwise, an individual's baseline covariates Z_i may affect his group assignment A_i. For Z_i to affect A_i and not A_j would entail an allocation rule that ignores balance across groups. In some settings, the vector of baseline covariates $\mathbf{Z}$ (or a function of $\mathbf{Z}$, e.g., its mean) would affect the allocation rule. This is represented by the presence of the dashed arrows into $\mathbf{A}$.

We now describe a toy example in which allocational interference operates and informs causal effects of interest. Suppose that runners enter a 5000 meter race, but the track is not wide enough for all of the competitors to race simultaneously. A race represents a single interference block; in order to perform statistical inference on the effects discussed below we would likely need to observe several independent races, indexed by k. The runners in each race are divided into smaller groups to race in successive heats. Number the subjects according to some composite measure of their recent performance, so that runner 1 is the fastest based on the composite measure and runner m is the slowest. Let Y_{ki} be the time in which runner i in block k finishes the race and Z_{ki} be a vector of all relevant baseline characteristics. A runner's speed will affect his own outcome. Moreover, his speed, confidence and sportsmanship may have an impact on the outcomes of the runners with whom he is grouped. These characteristics should all be included in Z_{ki}. The runners are divided into three heats, so $A_{ki} \in \{1, 2, 3\}$. For simplicity, we assume that m is divisible by 3 and each heat has $m/3$ runners, though heats of different sizes are possible. Consider the following two allocations: In allocation $\mathbf{a}$, runners 1 through $m/3$ are assigned to the first heat, runners $(m/3) + 1$ through $2m/3$ to the second heat and runners $(2m/3) + 1$ through m to the third heat. In allocation $\mathbf{a}'$, on the other hand, the first heat is comprised of runners $1, 4, 7, \dots, m - 2$; the second of runners $2, 5, 8, \dots, m - 1$, and the third of the remaining runners. Allocation $\mathbf{a}'$ results in a more balanced distribution of baseline speed across heats.

Are runners, on average, likely to run faster under one of these two allocations? This is a question about the overall causal effect $E[Y(\mathbf{a})] - E[Y(\mathbf{a}')]$. If the number of groups is the same under both allocations, as in our example, then the direct and indirect effects of allocations $\mathbf{a}$ and $\mathbf{a}'$ may also be of interest (see Section 2.1). The expected unit-level effect $UE_1(\mathbf{a}; 3, 1) \equiv E[Y_1(\mathbf{a}_{-1}, 3)] - E[Y_1(\mathbf{a}_{-1}, 1)]$ is the expected effect on runner 1 of racing in the fastest versus the slowest heat in allocation $\mathbf{a}$. The expected spillover effect $SE_1(\mathbf{a}, \mathbf{a}'; 1) \equiv E[Y_1(\mathbf{a}_{-1}, 1)] - E[Y_1(\mathbf{a}'_{-1}, 1)]$ is the expected effect on runner 1 of running in the first heat when that heat is comprised of the fastest runners versus running in the first heat when that heat is comprised of runners with a mix of speeds. (In both cases, the expectations are with respect to multiple independent races.) This might matter if running in the first heat was advantageous because the crowd was more enthusiastic earlier on, a point to which we return below.

As always, in order to identify expectations of counterfactuals of the form $Y_{ki}(\mathbf{a}_k)$ we require conditional exchangeability for the effect of $\mathbf{A}_k$ on Y_{ki}. If $\mathbf{A}_k$ and Z_{ki} share any common causes, as they would if, for example, heats were assigned based on the identity of the runners' coaches, then those common causes must be included in the conditioning set. If A_{ki} depends on any component of Z_{ki} then that component must be included in the conditioning set in order to achieve exchangeability. Similarly, if A_{ki} depends on a component of $\mathbf{Z}_k$, that is, there are arrows from Z_{ki} and from Z_{kj}, $j \neq i$, into A_{ki}, then that component of $\mathbf{Z}_k$ must be included in the conditioning set. We note that conditioning on a component of $\mathbf{Z}_k$ may block part of the effect of $\mathbf{A}_k$ on Y_{ki}; because Z^*_{ki} is a deterministic function of $\mathbf{Z}_k$ conditioning on the latter is effectively conditioning on the former. Z^*_{ki} lies on the causal pathway from $\mathbf{A}_k$ to Y_{ki} and, therefore, conditioning on it blocks part of the causal effect of interest.

Let T_{kl_i} be a group-level property of group l_i in block k, where l_i indexes the group to which individual i is assigned, and define $T_{ki} \equiv T_{kl_i}$ to be the group-level property to which individual i is exposed. If the order in which the heats are run makes a difference, because the weather changes throughout the day or because runners are tired later in the day, then T_{ki} could be the time at which subject i's heat is scheduled to race. This is an example of a preallocation group-level covariate that allows the groups to be distinguished from one another without reference to their composition. Other examples of this kind of group-level property are the teacher in charge of each classroom, the curricula to which classrooms are assigned, or different locations in which heats are assigned to run.

For the purposes of our race example, let T_{ki} be a measure of the crowd enthusiasm when runner i runs. The arrows from A_i into T_i on the DAG in Figure 15 are necessitated by the way we have defined T_{ki}, namely as a collection of properties of the group to which individual i is assigned. Properties that depend on group composition but are not captured by Z^*_{ki}, such as the number of runners in the heat, can also affect Y_{ki}. Unlike the time at which each heat runs, these properties arise after group allocation and, therefore, do not distinguish the groups from one another a priori. If T_{ki} is itself affected by group composition, because the size of the crowd is determined by who is in

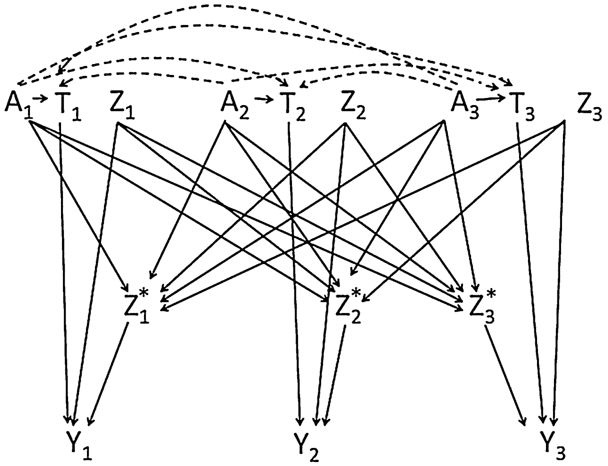

FIG. 15.

each heat, then we would also require arrows $A_j \to T_i$; these are the dashed arrows in Figure 15. Suppose that crowd enthusiasm is determined by the proportion of runners in the heat who are in the fastest quartile of all of the runners in the race, based on the baseline composite measure. Then T_{ki} is affected by $\mathbf{Z}_k$ (which includes a measure of each runner's previous performance), and we require arrows from each Z_j into T_i, as on the DAG in Figure 16. If data on $\mathbf{T}_k$ is not collected, or if it is not known whether any group-level properties affect Y_{ki}, then we would add arrows from each A_i into each Y_j on the DAG in Figure 15, to represent the residual effect of $\mathbf{A}_k$ on Y_{kj} due to unobserved group properties.

We also may be interested in whether the effect of an allocation is mediated by group attributes $\mathbf{T}_k$. In order to identify mediated effects through $\mathbf{T}_k$ it must be hypothetically possible to intervene on $\mathbf{T}_k$ without manipulating any other variable that may cause Y_{ki} not through $\mathbf{T}_k$. For example, if T_{ki} is the enthusiasm of the crowd when i's group runs the race, then we can imagine intervening on $\mathbf{T}_k$ by changing the composition of the crowd without changing the assignments of runners to heats or their covariates. On the other hand, if T_{ki} is the number of runners in i's group, then clearly any intervention on $\mathbf{T}_k$ must operate because of $\mathbf{A}_k$, which causes Y_{ki} not through $\mathbf{T}_k$. Natural direct, natural indirect and controlled direct effects are not coherently defined in this case. This is because counterfactuals of the form $Y_{ki}(\mathbf{a}_k, \mathbf{T}_k(\mathbf{a}'_k))$ are not well-defined if we cannot hypothetically simultaneously intervene on $\mathbf{A}_k$, setting it to $\mathbf{a}_k$, and on $\mathbf{T}_k$, setting it to its counterfactual value under allocation $\mathbf{a}'_k$. If the only way to set $\mathbf{T}_k$ to its counterfactual value under allocation $\mathbf{a}'_k$ is through an intervention that sets $\mathbf{A}_k$ to $\mathbf{a}'_k$, then this hypothetical joint intervention is not possible. The effects of $\mathbf{A}_k$ on Y_{ki} with Z^*_{ki} as a mediator are similarly incoherent, because it is impossible to imagine intervening on Z^*_{ki} without manipulating $\mathbf{A}_k$, $\mathbf{Z}_k$, or both. If we are interested in the role that Z^*_{ki} plays in the effect of group allocation on the outcome, we can instead estimate the effects of Z^*_{ki} on Y_{ki}.

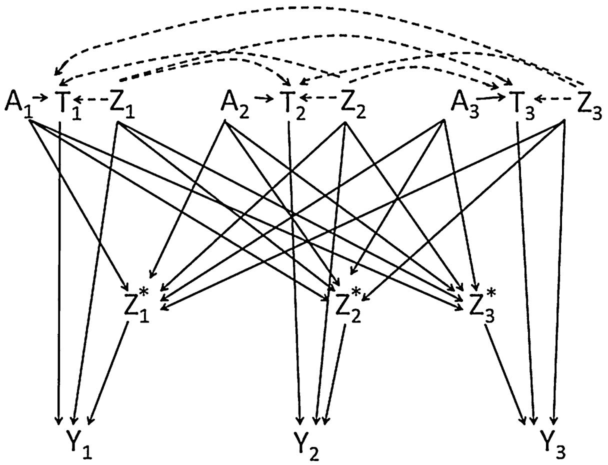

FIG. 16.

5. A NOTE ON INTERFERENCE AND SOCIAL NETWORKS

Our discussion thus far has focused on settings in which individuals are clustered into blocks and in which individuals in distinct blocks do not influence each other. In some contexts, it may be the case that there are no or few distinct independent blocks; social networks constitute one such setting. A social network is a collection of individuals and the social ties between them, for example ties of friendship, kinship or physical proximity. Social networks are of public health interest because certain health-related behaviors, beliefs and outcomes may propagate socially (Christakis and Fowler, 2007, 2008; Smith and Christakis, 2008), but of course they are rife with interference, making causal inference difficult.

Allocational interference essentially involves intervening on the network structure itself, creating new ties by assigning individuals to the same group, for example, assigning children to the same classroom, and possibly breaking old ties by assigning individuals to different groups. An intervention on classroom assignments within a school could be seen as creating a new network topology at the beginning of every school year. Because the network itself is manipulated in allocational interference, it may be a useful lens through which to understand interventions on ties in social network contexts.

Contagion and direct interference occur naturally and widely in social networks. Direct interference may be present whenever an exposure consists of ideas, beliefs, knowledge or physical goods which can be shared by an exposed individual with his associates. Contagion in social networks has been written about extensively in recent years (Christakis and Fowler, 2007, 2008; Mulvaney-Day and Womack, 2009; Smith and Christakis, 2008). Infectious diseases are more likely to spread between people closely connected in a social network (e.g., because they live together or spend time together), and in addition there is some recent evidence that traits and behaviors like obesity and smoking may be "socially contagious" (Christakis and Fowler, 2007, 2008). The precise mechanisms for these purported phenomena are unknown, but some researchers have hypothesized that latent outcomes related to the observed outcomes may be transmitted through social contact. For example, Christakis and Fowler (2007, 2008) have suggested that beliefs about the acceptability of different smoking behaviors and body types may be contagious. If this is in fact the mechanism underlying what appears to be contagion of smoking behavior or obesity, then the true structure is depicted by the DAG in Figure 17. O_i^t represents the observed characteristic of subject i at time t, for example his smoking behavior, and B_i^t represents his beliefs, for example, about the acceptability of smoking. We observe a phenomenon that resembles contagion, namely that O_i^t appears to have a causal effect on O_j^{t+1}. However, this apparent effect may not be due to a causal pathway but rather to the backdoor path $O_i^t \leftarrow B_i^{t-1} \rightarrow B_j^t \rightarrow O_j^{t+1}$ as in Figure 17. Another possible structure that would give rise to apparent contagion is presented in Figure 18. Here, O is indeed contagious, but the path by which contagion operates is mediated by B. Distinguishing between these different structures could have implications for interventions and policies. If the DAG in Figure 17 represents the true causal structure, then intervening on O or introducing a policy targeted at affecting O will not disrupt the contagious process; we should attempt to intervene on underlying beliefs instead. If the DAG in Figure 18 captures the true mechanisms at work, then intervening on either O or B can disrupt the contagious process.

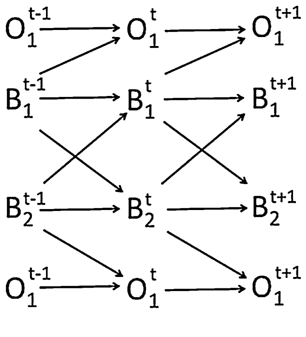

FIG. 17.

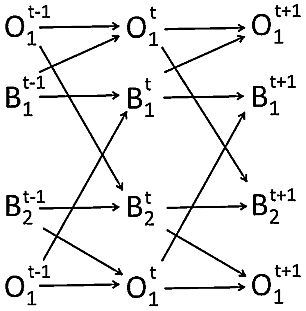

FIG. 18.

The discussion of contagion and direct interference in Section 4 may be useful for clarifying aspects of social network research. In many types of social networks and for many types of exposures and outcomes, both contagion and direct interference will be present. The results in Section 4.2.1 can sometimes be used to differentiate between the two types of interference. Contagion cannot be identified by cross-sectional network data without very strong assumptions about temporal and causal relationships, and some effects related to contagion require fine-grained information on outcomes in the network over time. Conversely, social network data can be used to refine assumptions about the structure of interference. We have assumed that interference occurred between all individuals in a block. This corresponds to a network in which each individual has a tie to every other individual in the same block. But if anything is known about the actual network topology, specifically about the absence of ties between certain individuals, then this information could be used to refine the causal structure of interference represented on the DAGs given in Section 3 and, therefore, the conditions under which causal effects are identifiable.

6. CONCLUSION

It is of paramount importance to carefully consider the specific causal structure whenever interference may operate on the relation between one individual's treatment and another's outcome. The possible structures are numerous and valid causal inference may require different assumptions in each one, depending on the effect of interest and the nature of confounding.

Also of great importance is the definition of the variables involved in the causal pathways under investigation. In some cases, the difference between interference by contagion and direct interference is contextual: depending on how we define the treatment and the outcome, some causal relationships can be seen as either one. Recall the example of direct interference that we presented in Section 4.1: the outcome is weight change and the treatment dietary counseling from a nutritionist. Direct interference occurs when a treated individual "treats" his associates by imparting to them the information gained from the nutritionist, thereby directly affecting their obesity status. Underlying this direct interference is a contagious process by which the treated individual transmits his understanding of how to adopt and maintain a healthy diet to his associates; "catching" this understanding causes the associates to lose weight. Defining variables precisely and narrowly is always a difficulty for causal inference; the challenge is to conduct valid inference when we do not observe the underlying causal mechanisms but instead have to base our analyses on constructs like weight, symptomatic flu, scholastic achievement, visits with a nutritionist, etc.

This paper scratches the surface of the enormous challenge of causal inference in the presence of interference. We have not, for example, touched upon estimation of causal effects or on inference, areas where some progress has been made in recent years (Aronow and Samii, 2013; Bowers, Fredrickson and Panagopoulos, 2013; Graham, Imbens and Ridder, 2010; Hudgens and Halloran, 2008; Manski, 2013; Rosenbaum, 2007; Tchetgen Tchetgen and VanderWeele, 2012) but much more is needed.

APPENDIX

We describe the identification of the effects of A_i on Y_j for the DAGs in Figure 5 when $\mathbf{C}$ is not fully observed.

In Figure 5(c), standardizing by C_i identifies the effect of A_i on Y_j:

$$\begin{aligned}
&\sum_{c_i} E[Y_j|A_i=a_i, C_i=c_i]P(C_i=c) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E[Y_j|A_i=a_i, \\
&\qquad\qquad C_i=c_i, A_j=a_j, C_j=c_j] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E\big[Y_j(a_i,a_j)|A_i=a_i, \\
&\qquad\qquad C_i=c_i, A_j=a_j, C_j=c_j\big] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E\big[Y_j(a_i,a_j)|C_i=c_i, C_j=c_j\big] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i).
\end{aligned}$$

This is a weighted average of $\mathbf{C}$-specific counterfactuals $Y_j(\mathbf{a})$. [Replacing C_i with D in the expressions above gives the identifying expression for the effect of A_i on Y_j in Figure 5(e).] Standardizing by C_i also identifies the effect of A_i on Y_i for the DAGs in Figures 5(a) and 5(c), similarly giving a weighted average of $\mathbf{C}$-specific counterfactuals:

$$\begin{aligned}
&\sum_{c_i} E[Y_i|A_i=a_i, C_i=c_i]P(C_i=c) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E[Y_i|A_i=a_i, \\
&\qquad\qquad C_i=c_i, A_j=a_j, C_j=c_j] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E\big[Y_i(a_i,a_j)|A_i=a_i, \\
&\qquad\qquad C_i=c_i, A_j=a_j, C_j=c_j\big] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i) \\
&\quad=\sum_{a_j}\sum_{c_j}\sum_{c_i} E\big[Y_i(a_i,a_j)|C_i=c_i, C_j=c_j\big] \\
&\qquad\cdot P(A_j=a_j, \\
&\qquad\qquad C_j=c_j|A_i=a_i, C_i=c_i) \\
&\qquad\cdot P(C_i=c_i).
\end{aligned}$$

ACKNOWLEDGMENT

This work was supported by NIH Grant ES017876.

REFERENCES

ANGRIST, J. D. and LANG, K. (2004). Does school integration generate peer effects? Evidence from Boston's METCO program. *American Economic Review* **94** 1613–1634.

ARONOW, P. M. and SAMII, C. (2013). Estimating average causal effects under general interference. Technical report.

AVIN, C., SHPITSER, I. and PEARL, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conference on Artificial Intelligence* 357–363. Morgan-Kaufmann, Edinburgh, UK.

BOWERS, J., FREDRICKSON, M. M. and PANAGOPOULOS, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis* **21** 97–124.

CARRELL, S. E., FULLERTON, R. L. and WEST, J. E. (2009). Does your cohort matter? Measuring peer effects in college achievement. *J. Labor Economics* **27** 439–464.

CHRISTAKIS, N. A. and FOWLER, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England J. Medicine* **357** 370–379.

CHRISTAKIS, N. A. and FOWLER, J. H. (2008). The collective dynamics of smoking in a large social network. *New England J. Medicine* **358** 2249–2258.

COHEN-COLE, E. and FLETCHER, J. M. (2008). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *J. Health Econ.* **27** 1382–1387.

DAHLHAUS, R. and EICHLER, M. (2003). Causality and graphical models in time series analysis. In *Highly Structured Stochastic Systems. Oxford Statist. Sci. Ser.* **27** 115–144. Oxford Univ. Press, Oxford. With part A by V. Didelez and part B by H. R. Künsch. MR2082408

DIDELEZ, V., KREINER, S. and KEIDING, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statist. Sci.* **25** 368–387. MR2791673

FREEDMAN, D. A. (2004). Graphical models for causation, and the identification problem. *Eval. Rev.* **28** 267–293.

GRAHAM, B. S., IMBENS, G. W. and RIDDER, G. (2010). Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. Technical report, National Bureau of Economic Research, Cambridge, MA.

GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

GREENLAND, S. and ROBINS, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **15** 413–419.

HALLORAN, M. E. and HUDGENS, M. G. (2012). Causal inference for vaccine effects on infectiousness. *Int. J. Biostat.* **8** Art. 6, front matter + 40. MR2925328

HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6** 142–151.

HECKMAN, J. J., LOCHNER, L. and TABER, C. (1998). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Technical report, National Bureau of Economic Research, Cambridge, MA.

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586.

HONG, G. and RAUDENBUSH, S. W. (2008). Causal inference for time-varying instructional treatments. *J. Educational and Behavioral Statistics* **33** 333–362.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060

MULVANEY-DAY, N. and WOMACK, C. A. (2009). Obesity, identity and community: Leveraging social networks for behavior change in public health. *Public Health Ethics* **2** 250–260.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809

PEARL, J. (1997). Graphical models for probabilistic and causal reasoning. In *The Computer Science and Engineering Handbook* (A. Tucker, ed.) 699–711. CRC Press, Boca Raton, FL.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* 411–420. Morgan-Kaufmann, San Francisco, CA.

PEARL, J. (2003). Statistics and causal inference: A review. *Test* **12** 281–318. MR2044313

PEARL, J. and VERMA, T. S. (1994). A theory of inferred causation. In *Logic, Methodology and Philosophy of Science IX* (*Upsala*, 1991). *Stud. Logic Found. Math.* **134** 789–811. North-Holland, Amsterdam. MR1328002

RICHARDSON, T. S. and ROBINS, J. M. (2013). Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, Univ. Washington, Seattle, WA.

ROBINS, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 70–82. *Oxford Statistical Science Series* **27**. Oxford Univ. Press, Oxford. MR2082403

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537

ROSS, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. Part I. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **92** 204.

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

SACERDOTE, B. (2000). Peer effects with random assignment: Results for Dartmouth roommates. Technical report, National Bureau of Economic Research, Cambridge, MA.

SMITH, K. P. and CHRISTAKIS, N. A. (2008). Social networks and health. *Annual Revue of Sociology* **34** 405–429.

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

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

TIAN, J. and PEARL, J. (2002a). A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence* 567–573. MIT Press, Cambridge, MA.

TIAN, J. and PEARL, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* 519–527. Morgan Kaufmann, San Francisco, CA.

VANDERWEELE, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociol. Methods Res.* **38** 515–544. MR2758165

VANDERWEELE, T. J. and HERNAN, M. A. (2013). Causal inference under multiple versions of treatment. *J. Causal Inference* **1** 1–20.

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011a). Bounding the infectiousness effect in vaccine trials. *Epidemiology* **22** 686–693.

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011b). Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Probab. Lett.* **81** 861–869. MR2793754

VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J. and HALLORAN, M. E. (2012). Components of the indirect effect in vaccine trials: Identification of contagion and infectiousness effects. *Epidemiology* **23** 751–761.

VANDERWEELE, T. J., HONG, G., JONES, S. M. and BROWN, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *J. Amer. Statist. Assoc.* **108** 469–482. MR3174634

VANSTEELANDT, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scand. J. Stat.* **34** 478–498. MR2368794

VERMA, T. S. (1993). Graphical aspects of causal models. Technical Report R-191, Univ. California, Los Angeles, CA.