# Causal inference for social network data

Elizabeth L. Ogburn, Oleg Sofrygin, Iván Díaz, and Mark J. van der Laan

#### Abstract

We extend recent work by van der Laan (2014) on causal inference for causally connected units to more general social network settings. Our asymptotic results allow for dependence of each observation on a growing number of other units as sample size increases. We are not aware of any previous methods for inference about network members in observational settings that allow the number of ties per node to increase as the network grows. While previous methods have generally implicitly focused on one of two possible sources of dependence among social network observations, we allow for both dependence due to contagion, or transmission of information across network ties, and for dependence due to latent similarities among nodes sharing ties. We describe estimation and inference for causal effects that are specifically of interest in social network settings.

# 1 Introduction and Background

Social networks have long been of interest to sociologists (Moreno, 1937) and in the past decade they have gained prominence across a broad range of disciplines, from economics to epidemiology (Knoke and Yang, 2008). A social network is, roughly, a set of individuals who share some kind of social relationships with one another – e.g. friends, neighbors, family members, or coworkers. If one individual's outcome can be affected by his or her social contacts' outcomes, then we say that the outcome exhibits induction or contagion, and the causal effects are called peer effects. If an individual's outcome may be affected by his or her contacts' treatments of exposures, then those treatments or exposures are said to exhibit interference.

Many aspects of social networks are of interest to researchers, from the clustering of individuals into communities to the probability distributions that describe the generation of new relationships between individuals in the network. There is increasing interest in identifying and estimating causal effects in the contexts of social networks, that is causal effects that one individual's behavior, treatment assignment, beliefs, or health outcome could have on his or her social contacts' behaviors, exposures, beliefs, or health statuses. There have been a number of high profile articles that use standard methods like generalized linear models (GLM) and generalized estimating equations (GEE) to attempt to infer causal peer effects from network data (e.g. Christakis and Fowler, 2007, 2008, 2010), and this work

<sup>\*</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

<sup>&</sup>lt;sup>†</sup>Department of Biostatistics, University of California, Berkeley, CA, USA

<sup>&</sup>lt;sup>‡</sup>Division of Biostatistics and Epidemiology, Weill Cornell Medicine, New York, NY, USA

has inspired several research programs that study peer effects using the same statistical methods (Ali and Dwyer, 2010; Cacioppo et al., 2009; Madan et al., 2010; Rosenquist et al., 2010; Wasserman, 2013). However, these methods have come under considerable criticism from the statistical community (Cohen-Cole and Fletcher, 2008; Lyons, 2011; Shalizi and Thomas, 2011). These statistical models are not equipped to deal with dependence across individuals and are rarely appropriate for estimating effects using network data (Ogburn and VanderWeele, 2014). In some settings it may be possible to use them to test for the presence of network dependence, but some properties of such tests are unknown (VanderWeele et al., 2012; Shalizi, 2012; Ogburn and VanderWeele, 2014).

Spatial autoregressive (SAR) models have been applied to the study of peer effects and induction in network settings (e.g. Goetzke, 2008; Lee, 2004; Lin, 2005; O'Malley and Marsden, 2008). Because the endogenous and exogenous variables are measured at the same time, they parameterize an equilibrium state rather than causal relationships. Few data generating processes give rise to true equilibrium states (Besag, 1974; Lauritzen and Richardson, 2002; Thomas, 2013); therefore SAR models may often be misspecified or uninformative about causal relationships.

Very recently, researchers interested in causal inference for interconnected subjects have begun to develop methods designed specifically for the network setting. Many methods for interference—the effect of one individual's treatment or exposure on others' outcomes—are relevant to the analysis of network data (Aronow and Samii, 2013; Athey et al., 2016; Bowers et al., 2013; Eckles et al., 2014; Graham et al., 2010; Halloran and Struchiner, 1995; Halloran and Hudgens, 2011; Hong and Raudenbush, 2006, 2008; Hudgens and Halloran, 2008; Rosenbaum, 2007; Rubin, 1990; Sobel, 2006; Tchetgen Tchetgen and VanderWeele, 2012; VanderWeele, 2010; VanderWeele and Tchetgen Tchetgen, 2011a,b). However, the inferential methods developed in this context generally require observing multiple independent groups of units, which corresponds to observing multiple independent networks, or else they require treatment to be randomized. Ideally, we would like to be able to perform inference even when all observations are sampled from a single social network and in observational settings in addition to randomized experiments.

Methodology has not kept apace with interest in causal inference using data from individuals connected in a social network, and researchers have resorted to using standard statistical methods to analyze this new type of data, or to collecting multiple independent groups of individuals and using the groups, rather than the individuals, as the unit of statistical inference. The former strategy is not statistically appropriate and the he latter strategy, while statistically valid, can be an inefficient use of data. Furthermore, it limits the settings and effects that one can study, because in some settings it may be too expensive or labor intensive to collect data from many independent groups or independent groups may not exist (e.g. a global infectious disease epidemic).

In this paper we extend recent work by van der Laan (2014) on causal inference for causally connected units to more general social network settings. van der Laan (2014) introduced methods for causal inference from a single collection of interconnected units when each unit is known to be independent of all but a small number of other units. Asymptotic results rely on the number of dependent units being fixed as the total number of units goes to infinity. We propose new methods that allow similar causal and statistical inference without requiring the number of dependent units to be fixed as sample size increases. We are not aware of any previous methods for inference about

network members in observational settings that allow the number of ties per node to increase as the network grows. As we discuss in Section 2.3, below, this is a crucial feature of most realistic models for social network generation. While previous methods have generally implicitly focused on one of two possible sources of dependence among social network observations, we allow for both dependence due to contagion or transmission of information across network ties, and dependence due to latent similarities among nodes sharing ties. We describe estimation and inference for causal effects that are specifically of interest in social network settings.

In Section (2) we give some background on causal inference for social network data, discussing briefly the relationship between causal structural equation models and network edges, the types of statistical dependence likely to be found in social network data, and asymptotic growth. In Section (3.1) we present our target of inference and the identifying assumptions that we will use in the methods that follow. We present the efficient influence function for our target parameter under the conditional independence assumptions from van der Laan (2014). When these independence assumptions are relaxed, this will still be an influence function for our target parameter but it may not be efficient. In Section (3.3) we describe estimation procedures—using the (efficient) influence function from (3.1)—which will be efficient under the stronger independence assumptions but still consistent and asymptotically normal under the weaker independence assumptions. In Section (3.4) we prove our main result, which is the asymptotic normality of our estimator under an asymptotic regime in which the number of ties per node grows unboundedly with n. In Section (4) we discuss estimation of rather esoteric causal effects that are specifically of interest in social network settings. Section (5) includes simulation results, and Section (6) concludes.

# 2 Background and setting

#### 2.1 Networks and structural equation models

A network is a collection of units and information about the presence or absence of pairwise ties between them. The presence of a tie between two units indicates that the units share some kind of a relationship; what types of relationships are encoded by network ties depends on the context. For example, in a social network we might define a tie to include familial relatedness, friendship, or shared place of work. In the study of sexually transmitted diseases, such as HIV, the sexual contact network is of great interest; here ties represent sexual partnership within a given time frame. Some types of relationships are mutual, for example familial relatedness and shared place of work. Others, like friendship, can go in only one direction: Tom may consider Sue to be his friend, while Sue does not consider Tom to be her friend. Edges can be binary (present or absent), multiplex (categorical with different levels for different types of relationships), or weighted (with continuous or ordinal weights) according to the strength of the relationship.

Networks are most often studied through the lens of graph theory. Units are represented by *nodes* and ties are represented by *edges*. Two nodes connected by an edge are called neighbors. A node whose characteristics we wish to explain or model is called an *ego*; the ego's neighbors are its *alters*. If ties must be mutual then the graph is *undirected*; if a tie can go in only one direction then the graph is *directed*. For simplicity we will assume all networks are undirected in what follows, but our methods are

equally applicable to directed networks. In an undirected network, the *degree* of a node is the number of edges it has or, equivalently, the number of alters. The graph theoretic approach to networks is immensely powerful and useful. It has been used to study all manner of network structures, network generating models, and diffusion processes on networks (see Newman, 2009 and references therein for examples). However, this approach largely reduces networks to topological features, and it is not well suited for statistical and causal inquiries into mechanisms of diffusion and influence across network nodes.

Underlying inquiries into causal effects across network nodes is a representation of the network as a structural equation model. Consider a network of n subjects, indexed by i, with binary undirected ties  $A_{ij} \equiv I$  {subjects i and j share a tie}. The matrix  $\mathbf{A}$  with entries  $A_{ij}$  is the adjacency matrix for the network. Associated with each subject is a vector of random variables,  $O_i$ , including an outcome  $Y_i$ , covariates  $C_i$ , and an exposure or treatment variable  $X_i$ , all possibly indexed by time t. In numerous applications across the social, political, and health sciences, researchers are interested in ascertaining the presence of and estimating causal interactions across alter-ego pairs. Is there interference, i.e. does the treatment of subject i have a causal effect on the outcome of subject j when i and j share a network tie? Is there peer influence, i.e. does the outcome of subject i at time i have a causal effect on a future outcome of subject i when i and i are adjacent in the network? These inquiries can be formalized with the help of a causal structural equation model, informed by the network.

A structural equation model is a system of equations of the form  $y_i = f_i [pa_i(Y), \varepsilon_i]$ , where  $pa_i(Y)$ , the set of parents of Y, is a collection of variables that are causes of Y for subject i, and  $\varepsilon_i$  is an error term that may include omitted causes of Y. In general  $C_i$  and  $X_i$  will be included in  $pa_i(Y)$ . See Pearl (2000) for further discussion of SEMs. When causal inference is performed on network data, the network ties inform which variables are to be included in  $pa_i(Y)$ . For example, if interference might be present, then the collection of treatment variables for i's alters,  $\{X_j : A_{ij} = 1\}$ , must be included in the set  $pa_i(Y)$ . If contagion might be present then  $\{Y_{j,t-k} : A_{ij} = 1\}$  must be included in the set  $pa_i(Y_t)$ , where t indexes time and t is an outcome-specific lag time such that no causal effect can be transmitted from one person to another in less than t time steps (see Ogburn and VanderWeele, 2013 for discussion of lag times).

It is important that the network be completely and accurately specified; missing ties are akin to missing components of a multidimensional treatment vector because they result in important elements of exposure of interest being left out of the SEM. Whenever an inquiry into causal effects is informed by a social network, measurement error in the network is tantamount to measurement error in the exposure of interest. The network for which data is collected must be calibrated to the causal question of interest. If we are interested in peer effects on academic achievement among elementary school children and think that being in the same classroom is the relationship that determines whether or not two children affect one another's outcomes, then being in the same classroom is the relationship that determines whether or not a network tie exists, and a network that captures interaction during playground sports is not informative or useful. In other words, a tie between nodes i and j represents the possibility of a causal effect of an element of  $O_i$  on an element of  $O_j$  at a later time, and vice versa. These issues have not been made explicit in much of the existing literature on causal inference for network data; equating a network with the underlying SEM can help to make them precise.

## 2.2 Networks and dependence

Perhaps the greatest challenge and barrier to causal and statistical inference using observations from a single, interconnected social network is dependence among observations. The literature on statistics for dependent data is vast and multifaceted, but very little has been written on the dependence that arrises when observations are sampled from a single network. Most of the literature on dependent random variables assumes that the domain from which observations are sampled (e.g. time or geographic space) has an underlying Euclidean topology. The principles behind asymptotic results in the Euclidean dependence literature are simple and intuitive. They rely on a combination of stationarity assumptions, i.e. assumptions that certain features of the data generating process do not depend on an observation's location in the sample domain, and assumptions that bound the nature and the amount of dependence in the data. Most frequently these are mixing assumptions, which describe the decay of the correlation between observations as a function of the distance between them. Sometimes the stronger assumption of m-dependence is made, according to which two observations are independent if they are sampled from locations that are m or more units apart. Intuitively, in order to extract an increasing amount of information from a growing sample of dependent observations, old observations must be predictive of new observations, which is ensured by stationarity assumptions, and the amount of independence in the sample must grow faster than the amount of dependence, which is ensured by mixing conditions or m-dependence.

This literature is not immediately applicable to the network setting. Roughly, this is due to the difference between Euclidean and network topology. While it is possible to embed a network in  $\mathbb{R}^d$ in such a way that preserves distances, to do so is to allow d to increase as n increases. Euclidean dependence results generally require d to be fixed, implying that, as new observations are sampled at the boundary of a Euclidean domain, the average and maximum pairwise distance between observations increases. Networks, on the other hand, often do not have a clear boundary to which we can add observations in such a way that ensures growth in the sample domain. In a large sample with Euclidean dependence, most observations will be distant from most other observations. This is not necessarily the case in networks. The maximum distance between two nodes can be small even in very large networks, and even if the maximum distance between two nodes is large, there may be many nodes that are close to one another. Even under m-dependence with small m, networks exist in which most observations remain less than m units apart as  $n \to \infty$ . Therefore, mixing conditions and m-dependence do not necessarily result in more independence than dependence in a large sample from a network. Research indicates that social networks generally have the small-world property (sometimes referred to as the "six degrees of separation" property), meaning that the average distance between two nodes is small (Watts and Strogatz, 1998). Therefore distances in real-world networks may grow slowly with sample size. Of course some types of networks, e.g. lattices, embed in  $\mathbb{R}^d$  as n grows, but these are generally trivial cases that are not useful for naturally occurring networks like social networks.

Dependence in networks is of two varieties, each with its own implications for inference. In the literature on spatial and temporal dependence, dependence is often implicitly assumed to be the result of latent traits that are more similar for observations that are close in Euclidean distance than for distant observations. This type of dependence is likely to be present in many network contexts as well. In networks, edges present opportunities to transmit traits or information, and this direct transmission

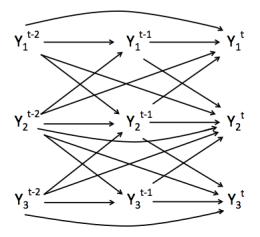
is an important additional source of dependence that depends on the underlying network structure.

Latent variable dependence will be present in data sampled from a network whenever observations from nodes that are close to one another are more likely to share unmeasured traits than are observations from distant nodes. Homophily, or the tendency of people who share similar traits to form network ties, is a paradigmatic example of latent variable dependence. If the outcome under study in a social network has a genetic component, then we would expect latent variable dependence due the fact that family members, who share latent genetic traits, are more likely to be close in social distance than people who are unrelated. If the outcome were affected by geography or physical environment, latent variable dependence could arise because people who live close to one another are more likely to be friends than those who are geographically distant. Of course, these traits can create dependence whether they are latent or observed. But if they are observed then conditioning on them renders observations independent; therefore the methodological challenges are greater when they are latent. Just like in the spatial and temporal dependence context, there is often little reason to think that we could identify, let alone measure, all of these sources of dependence. In order to make any progress towards valid inference in the presence of latent trait dependence, some structure must be assumed, namely that the range of influence of the latent traits is primarily local in the network and that any long-range effects are negligible. In a structural equation model, latent trait dependence would be captured by dependence among the error terms across subjects.

Whenever one subject's treatments, outcomes, or covariates affect other subjects' treatments, outcomes, or covariates, observations across subjects are dependent. However, this kind of dependence, which arrises from causal effects between subjects, has structure lacking in latent trait dependence. Figure 1 depicts contagion in a network of three individuals. This diagram is the directed acyclic graph representation (Pearl, 1995; Ogburn and VanderWeele, 2013) of the following structural equation model: At each time t,  $Y_i^t$  is affected by i's own past outcomes and those of i's social contacts. Individual 2 shares ties with 1 and 3 but individuals 1 and 3 are not connected. This structure implies conditional independences: because any transmission from individual 1 to 3 must pass through 2 we have that  $Y_1^{t-2} \perp Y_3^t \mid Y_2^{t-1}$ ; because information cannot be transmitted instantaneously we have that  $Y_1^{t-2} \perp Y_2^{t-2}$ . If observations are observed at closely spaced time intervals then these conditional independences can be harnessed for inference. There is no reason to think that any such conditional independences would hold with latent variable dependence.

In this paper, we accommodate both dependence due to direct transmission and dependence due to latent traits. We assume that both kinds of dependence are limited to dependence neighborhoods determined by the underlying social network: each subject, or node, i can directly transmit to information, outcomes, or exposures to the nodes with which i shares a network tie, and each node i can share latent traits with the nodes with which i shares a network tie or a mutual connection. That a subject can only transmit to his or her immediate social contacts may be a reasonable assumption (indeed, we can define network ties in such a way as to make this true by definition), but it is likely unrealistic to assume that latent variable dependence only affects nodes at a distance of one or two ties, as we assume throughout. This represents a first step towards valid statistical and causal inference under more realistic assumptions than have been required by previous work, but future work is needed to address more realistic—i.e. longer range—forms of latent variable dependence.

Figure 1: Dependence due to direct transmission



## 2.3 A note on asymptotic growth

There are many complex issues surrounding asymptotic growth of networks (e.g. Diaconis and Janson, 2007; Shalizi and Rinaldo, 2013), and a large literature on graph limits (Lovász, 2012). These issues are largely beyond the scope of this paper, but we believe that our methods are consistent with the large and realistic class of network-generating models. In particular, observed social networks and models proposed for generating social networks tend to have heavy-tailed degree distributions, with most nodes having low degree but a non-trivial proportion of nodes having high degree, with the maximum degree dependent on the size of the network. Some researchers speculate that the heavy right tails of social network degree distributions tend to approximately follow power laws:  $Pr(degree = k) \sim k^{-\alpha}$  for  $2 < \alpha < 3$  (Barabási and Albert, 1999; Lovász, 2012; Newman and Park, 2003), in which case  $Pr(degree > K) = O(K^{1-\alpha})$  for any fixed K. Even if degree distributions depart from power law distributions (Clauset et al., 2009) they are frequently incompatible with the assumption of bounded degrees, which has been used in previous methods for inference about observations sampled from a single social network. Our new methods are not able to accommodate the most highly connected nodes from a power law degree sequence, but they can nevertheless be used to perform inference about the other nodes in a network that has a power law degree distribution (see Section 4.4).

Our theoretical results require an asymptotic regime in which the number of nodes in the network, n, goes to infinity. Formalizing asymptotic growth of network-generating models is beyond the scope of this paper; we take for granted sequence of networks with increasing n such that key features of the network-whatever features inform the causal effects of interest-are preserved. The structural equation model that specifies the distributions of covariates, treatment, and outcome does not change as n increases.

# 3 Methods

In this section we describe estimation of and inference about the causal effect of a treatment or exposure, X, which can depend upon covariates but is otherwise independent of outcomes and network topology. This case includes randomized and non-randomized exposures that may or may not be subject to interference, specifically direct interference (Ogburn and VanderWeele, 2013), where the effect of one individual's exposure on another's outcome is unmediated by the first individual's outcome. (In Section 4 we discuss causal effects that correspond to interference due to contagion, where the effect of one individual's exposure on another's outcome is mediated by the first individual's outcome, and allocational interference, where treatment changes network ties.) The approach we describe below is different from traditional approaches to interference in that they are justified when partial interference does not hold. As far as we are aware, this is the first approach to interference that references an asymptotic regime in which the number of ties for a given individual may grow with sample size. The estimating procedure that we describe in this section was first proposed by van der Laan (2014), but we generalize the results to a broader class of causal effects and to more general and pervasive forms of dependence among observations. The conditions under which the resulting estimators are consistent and asymptotically Normally distributed are different and weaker here than those in van der Laan (2014).

Table 1: Properties of marginal estimands and of estimands conditional on C

Properties that we have demonstrated for the two classes of estimands	Estimand class	
	Marginal	Conditional
nonparametrically identified with or without latent variable (LV) dependence	yes	yes
estimator is CAN with or without LV dependence	yes	yes
efficient estimator is available with LV dependence	no	no
efficient estimator is available without LV dependence	yes	yes
consistent and computationally tractable variance estimation with LV dependence	no	yes
consistent and computationally tractable variance estimation without LV dependence	yes	yes

For the remainder of Section 3, we describe CAN estimators of causal effects under two different sets of assumptions. One set of assumptions allows dependence due to direct transmission but not latent variable dependence, as in van der Laan (2014); under this set of assumptions our estimators inherit the efficiency properties from van der Laan (2014). The other set of assumptions allows dependence due to direct transmission and latent variable dependence; under this set of assumptions our estimators are CAN but may not be efficient. Our main result is the proof of asymptotic normality under an asymptotic regime in which the number of ties for a given individual may grow with sample size in Section 3.4.

In Section 3.5 we describe statistical inference for the estimators introduced in Section 3. We consider two different classes of estimators: estimators that marginalize over baseline covariates and estimators that condition on baseline covariates. In some cases, variance estimation is facilitated by conditioning on covariates. Under the assumptions encoded in the structural equation model in Section 3.1, the conditional estimator is in fact consistent for the marginal estimand. However, conditional estimators have smaller variance and inference about the conditional estimand cannot be interpreted as inference about the marginal estimand. All of our estimands and estimators condition on the observed network as given by the adjacency matrix **A**. Table 1 summarizes the relationships among the two sets of assumptions (with and without latent variable dependence) and the two classes of estimators (marginal over **C** and conditional on **C**) according to their properties and according to the limitations of our proposed methods.

In 4 we describe more elaborate interventions and causal questions that can be addressed by our methods and that may be of interest specifically in the context of social network data. However, we focus throughout on single treatment. Longitudinal interventions are also possible but we leave the details for future work. In addition, we state our results under the assumption that all variables take values on discrete sets. Analogous results are valid for other types of random variables: it is straightforward to extend our notation and results to continuous covariates and outcomes, but continuous treatments are more complicated (see van der Laan, 2014).

# 3.1 Structural equation model

Recall that  $C_i$  denotes a vector of covariates,  $X_i$  is a treatment variable, and  $Y_i$  is the outcome for subjects  $i \in \{1, ..., n\}$ . In addition, let  $K_i = \sum_{j=1}^n A_{ij}$ , that is,  $K_i$  is degree of node i or the number of individuals sharing a tie with individual i. The degree of subject i and the degrees of i's alters may be included in the covariate vector  $C_i$ . We define  $\mathbf{Y} = (Y_1, ..., Y_n)$  and  $\mathbf{C}$  and  $\mathbf{X}$  analogously. We

use a structural equation model to define the causal effects of interest, as in Section 2, but note that analogous definitions may be achieved within the potential outcome framework (van der Laan, 2014).

We assume that the data are generated by sequentially evaluating the following set of equations:

$$C_{i} = f_{C} [\varepsilon_{C_{i}}]$$
  $i = 1, ..., n$   
 $X_{i} = f_{X} [\{C_{j} : A_{ij} = 1\}, \varepsilon_{X_{i}}]$   $i = 1, ..., n$   
 $Y_{i} = f_{Y} [\{X_{j} : A_{ij} = 1\}, \{C_{j} : A_{ij} = 1\}, \varepsilon_{Y_{i}}]$   $i = 1, ..., n,$  (1)

where  $f_C$ ,  $f_X$ , and  $f_Y$  are unknown and unspecified functions and  $\varepsilon_i = (\varepsilon_{C_i}, \varepsilon_{X_i}, \varepsilon_{Y_i})$  is a vector of exogenous, unobserved errors for individual i. This set of equations encodes the assumptions of direct interference in an observational or randomized setting: the treatment X may or may not depend on C, depending on the specification of  $f_X$ , and the outcome Y depends on X and C but not on previous outcomes or on A, except possibly through C. Time ordering is a fundamental component of a structural causal model. For example, we assume that C is first drawn for all units, so that, in addition to  $C_i$ , the other components of the vector C may affect the value of  $X_i$ .

In addition, nonparametric identification of causal effects requires the following assumptions on the error terms from the SEM:

$$(\varepsilon_{X_1}, ..., \varepsilon_{X_n}) \perp (\varepsilon_{Y_1}, ..., \varepsilon_{Y_n}) \mid \mathbf{C},$$
 (A1)

$$\varepsilon_{X_1},...,\varepsilon_{X_n}\mid \mathbf{C}$$
 are identically distributed and  $\varepsilon_{Y_1},...,\varepsilon_{Y_n}\mid \mathbf{C},\mathbf{X}$  are identically distributed, (A2a)

$$\varepsilon_{X_i} \perp \varepsilon_{X_j} \mid \mathbf{C} \text{ and } \varepsilon_{Y_i} \perp \varepsilon_{Y_j} \mid \mathbf{C}, \mathbf{X} \text{ for } i, j \text{ s.t. } A_{ij} = 0 \text{ and } \exists ! k \text{ with } A_{ik} = A_{kj} = 1$$
 (A2b)

$$\varepsilon_{C_i}, i = 1, ..., n$$
, are identically distributed, and (A3a)

$$\varepsilon_{C_i} \perp \varepsilon_{C_i}$$
 for  $i, j$  s.t.  $A_{ij} = 0$  and  $\exists ! k$  with  $A_{ik} = A_{kj} = 1$ . (A3b)

Assumption (A1) is a no unmeasured confounding assumption; it ensures that **C** suffices to control for confounding of the effect of **X** on **Y**. It implies that any latent variable dependence affects **X** and **Y** separately; in general a latent variable that affected **X** and **Y** jointly would constitute a violation of this assumption. Assumptions (A2b) and (A3b) ensure that any unmeasured sources of dependence—i.e. latent trait dependence—only affect pairs of observations up to a distance of two network ties—that is, friends or friends-of-friends. Assumption (A3) can be omitted if attention is restricted to causal effects conditional on **C**.

Although our main result, given in Theorem 1 below, holds for inference in the SEM defined by assumptions (A1)–(A3b), some asymptotic properties are guaranteed only when stronger versions of assumptions (A2b) and (A3b) hold. We therefore introduce alternative assumptions

$$\varepsilon_{X_1}, ..., \varepsilon_{X_n} \mid \mathbf{C} \text{ are i.i.d. and } \varepsilon_{Y_1}, ..., \varepsilon_{Y_n} \mid \mathbf{C}, \mathbf{X} \text{ are i.i.d., and}$$
 (A4)

$$\varepsilon_{C_i}, i = 1, ..., n$$
, are i.i.d. (A5)

Note that, although the variance-covariance structure of the SEM given in (1) is affected by the dependence allowed in (A2b) and (A3b), the mean structure is unaltered by the choice of assumptions

(A2) and (A3) or (A4) and (A5). Therefore, any estimator that is unbiased under (A4) and (A5) will remain unbiased when these are relaxed to (A2) and (A3). In Section 3.2 we discuss nonparametric identification of causal parameters, which is agnostic to the choice of the weaker or stronger independence assumptions. In Section 3.3 we derive estimators under assumptions (A1), (A4), and (A5)—that is, in the presence of dependence due to direct transmission but not latent variable dependence. We use the stronger assumptions because the resulting model is amenable to familiar tools for deriving semiparametric estimators. In Section (3.4) we prove that the estimators derived under assumptions (A1), (A4), and (A5) are CAN under the weaker set of assumptions (A1)—(A3b). In Section 3.5 we discuss inference under each of the two sets of assumptions.

# 3.2 Definition and nonparametric identification of causal effects

In principle it is possible to perform statistical inference in the model defined by assumptions (A1)-(A3b) or by assumptions (A1), (A4), and (A5). However, in practice, with finite data and finite computing resources, we may need to make dimension-reducing assumptions on the forms of  $f_X$  and  $f_Y$ . This is done by considering summary functions  $s_X$  and  $s_Y$  and random variables  $W_i = s_{X,i} (\{C_j : A_{ij} = 1\})$  and  $V_i = s_{Y,i} (\{C_j : A_{ij} = 1\}, \{X_j : A_{ij} = 1\})$  such that the model may be written as

$$C_i = f_C [\varepsilon_{C_i}]$$
  $i = 1, ..., n$   
 $X_i = f_X [W_i, \varepsilon_{X_i}]$   $i = 1, ..., n$   
 $Y_i = f_Y [V_i, \varepsilon_{Y_i}]$   $i = 1, ..., n$ .

For example,  $s_{X,i}\left(\{C_j:A_{ij}=1\}\right)=\left(C_i,\sum_{j:A_{ij}=1}C_j\right)$  implies that the treatment status of unit i only depends on i's own covariate value and on the sum of the covariate values of the units sharing a tie with i. Analogously,  $s_{Y,i}\left(\{C_j:A_{ij}=1\},\{X_j:A_{ij}=1\}\right)=\left(C_i,\sum_{j:A_{ij}=1}C_j,X_i,\sum_{j:A_{ij}=1}X_j\right)$  is an example of a summary function for  $f_Y$ . For convenience we use the notation  $s_{X,i}(\mathbf{C})$  and  $s_{Y,i}(\mathbf{C},\mathbf{X})$  below; however, this notation should not undermine the important fact that  $W_i$  can only depend on the subset and  $\{C_j:A_{ij}=1\}$  and  $\{X_j:A_{ij}=1\}$  of  $\mathbf{C}$  and  $\mathbf{X}$ , as these are the only components of  $\mathbf{C}$  and  $\mathbf{X}$  that are parents of  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, in the causal model defined by the network-as-structural-causal-model. For notational convenience, in what follows we augment the observed data random vector with  $V_i$  and  $W_i$ , recognizing that these are deterministic functionals of  $C_i$  and  $X_i$ , defined by  $s_{Y,i}$  and  $s_{X,i}$ , and are therefore technically redundant:  $O_i=(C_i,W_i,X_i,V_i,Y_i)$ .

Potential or counterfactual outcomes can be defined in terms of hypothetical interventions to the underlying structural causal model. For example, a hypothetical intervention that deterministically sets  $X_i$  to a user-given value  $x_i^*$  for i = 1, ..., n is given by

$$C_{i} = f_{C} \left[ \varepsilon_{C_{i}} \right]$$

$$X_{i} = x_{i}^{*}$$

$$i = 1, \dots, n$$

$$Y_{i}(\mathbf{x}^{*}) = f_{Y} \left[ V_{i}(\mathbf{x}^{*}), \varepsilon_{Y_{i}} \right]$$

$$i = 1, \dots, n,$$

where  $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ . Here  $Y_i(\mathbf{x}^*)$  denotes the potential outcome of individual i in a hypothetical

world in which  $P(\mathbf{X} = \mathbf{x}^*) = 1$ . Analogously,  $V_i(\mathbf{x}^*) = s_{Y,i}(\mathbf{C}, \mathbf{x}^*)$  is a counterfactual random variable in a hypothetical world in which  $P(\mathbf{X} = \mathbf{x}^*) = 1$ . Note that, although  $V_i(\mathbf{x}^*)$  is counterfactual, its value is determined by the observed realization of  $\mathbf{C}$  and by the user-specified value  $\mathbf{x}^*$ , and it is therefore known. In order to streamline notation as we describe increasingly complex interventions, we denote the counterfactual variables  $V_i(\mathbf{x}^*)$  and  $Y_i(\mathbf{x}^*)$  by  $V_i^*$  and  $Y_i^*$ , respectively. The causal parameter of interest throughout is the expected average potential outcome in this same hypothetical world, i.e.  $E\left[\bar{Y}^*\right]$ , where  $\bar{Y}^* = \frac{1}{n}\sum_{i=1}^n Y_i^*$ .

We are now ready define notation that we will use throughout the remainder of the paper for functionals of the distribution of  $\mathbf{O}$ . Let  $p_C(\mathbf{c}) = P\left(\mathbf{C} = \mathbf{c}\right)$ ,  $g(\mathbf{x}|\mathbf{w}) = P\left(\mathbf{X} = \mathbf{x} \mid \mathbf{W} = \mathbf{w}\right)$ ,  $g_i(x|w) = P\left(X_i = x \mid W_i = w\right)$ ,  $p_Y(\mathbf{y}|\mathbf{v}) = P\left(\mathbf{Y} = \mathbf{y} \mid \mathbf{V} = \mathbf{v}\right)$ , and  $p_{Y,i}(y|v) = P\left(Y_i = y \mid V_i = v\right)$ . Define the two marginal distributions  $h_i(v) = P\left(V_i = v\right)$  and  $h_{i,x^*}(v) = P\left(V_i^* = v\right)$ , noting that both  $h_i$  and  $h_{i,x^*}$  are determined by g and g and are therefore observed data quantities. Finally,  $m(v) = \sum_y y \, p_Y(y|v)$  is the conditional expectation of Y given V = v.

In addition to assumptions (A1)-(A3b) or (A1), (A4), and (A5), identification of  $E\left[\bar{Y}^*\right]$  requires the positivity assumption that

$$P(V = v | \mathbf{C} = \mathbf{c}) > 0$$
 for all  $v$  in the range of  $V_i^*$  and for all  $\mathbf{c}$  such that  $P(\mathbf{C} = \mathbf{c}) > 0$ . (A6)

This assumption states that, within levels of  $\mathbf{C}$ , the values of V determined by the hypothetical intervention  $\mathbf{x}^*$  have positive probability under the observed-data-generating distribution. Now the causal parameter  $E\left[\bar{Y}^*\right]$  is identified by

$$\psi = \frac{1}{n} \sum_{i=1}^{n} E[m(V_i^*)] = \frac{1}{n} \sum_{i=1}^{n} \sum_{v} m(v) h_{i,x^*}(v).$$
 (2)

This identification result is equivalent to

$$\psi = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathbf{c}} m(s_{Y,i}(\mathbf{c}, \mathbf{x}^*)) p_C(\mathbf{c}).$$
(3)

From (3), it is clear that the conditional causal parameter  $E\left[\bar{Y}^* \mid \mathbf{C} = \mathbf{c}\right]$  is identified by  $\frac{1}{n} \sum_{i=1}^n m(s_{Y,i}(\mathbf{c}, \mathbf{x}^*))$ .

# 3.3 Estimation

Estimation of and inference about  $E\left[\bar{Y}^*\right]$  requires a statistical model  $\mathcal{M}$  for the distribution of the observed data  $P(\mathbf{O})$ . That is,  $\mathcal{M}$  is a collection of distributions over  $\mathbf{O}$  of which one element is the true data-generating distribution. Our target of inference is a pathwise differentiable mapping  $\Psi: \mathcal{M} \to \mathbb{R}$  such that  $\psi$  is  $\Psi(P)$ , the mapping evaluated at the true data-generating distribution. Under assumptions (A1), (A4), and (A5) the probability distribution of the observed data may be factorized as

$$P(\mathbf{O} = \mathbf{o}) = P(\mathbf{C} = \mathbf{c}) g(\mathbf{x}|\mathbf{w}) p_Y(\mathbf{y}|\mathbf{v}), \tag{4}$$

suggesting that  $\mathcal{M}$  requires three components: a model for  $p_C$ , a model for g, and a model for P(Y|V). Furthermore, the identification results in (2) and (3) indicate that  $\psi$  depends on P(Y|V) only through m. The empirical distribution  $\hat{p}_C$  can be used throughout to nonparametrically estimate  $p_C$ , but, when  $\mathbb{C}$  is high-dimensional, g and m cannot be non-parametrically estimated at rates of convergence that are fast enough to satisfy the regularity conditions of Theorem 1 (see Appendix). Therefore, in order to define the parameter mapping we require a statistical model  $\mathcal{M} = \mathcal{M}_g \times \mathcal{M}_m$ , where  $\mathcal{M}_g$  is a collection of conditional distributions for X given W such that the true conditional distribution is a member, and  $\mathcal{M}_m$  is a collection of expectations of Y relative to conditional distributions of Y given V such that the true conditional expectation of Y given Y is a member. Estimation of Y is based on the efficient influence function for the parameter mapping  $\Psi: \mathcal{M} \to \mathbb{R}$ .

Under assumptions (A1), (A4), and (A5), the efficient influence function, D, evaluated at a fixed value  $\mathbf{o}$  of  $\mathbf{O}$ , is given by

$$D(\mathbf{o}) = \sum_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} E\left[m\left(V_{i}^{*}\right) \mid C_{j} = c_{j}\right] - \psi + \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{h}_{x^{*}}(v_{i})}{\bar{h}(v_{i})} \left\{y_{i} - m\left(v_{i}\right)\right\},$$
 (5)

where  $\bar{h}(v_i) = \frac{1}{n} \sum_{j=1}^n h_j(v_i)$ ,  $\bar{h}_{x^*}(v_i) = \frac{1}{n} \sum_{j=1}^n h_{j,x^*}(v_i)$ ,  $v_i = s_{Y,i}(\mathbf{c}, \mathbf{x})$ , and  $V_i^* = s_{Y,i}(\mathbf{C}, \mathbf{x}^*)$ . The influence function has expected value equal to 0 at the true  $\psi$ ; this fact can be used to generate unbiased estimating equations for  $\psi$ . Estimating equations based on the efficient influence function are doubly robust: the right hand side of Equation (5) has expected value equal to 0 if  $m(\cdot)$  is replaced with an arbitrary functional of V or if  $g(\cdot)$  is replaced with an arbitrary functional of W, as long as one of the two remains correctly specified. (Recall that  $g(\cdot)$ , along with  $p_C$ , determines  $\bar{h}_{x^*}(v_i)$  and  $\bar{h}(v_i)$ .) This implies that an estimating equation based on Equation (5) will be unbiased for  $\psi$  if either model  $\mathcal{M}_m$  for  $m(\cdot)$  or model  $\mathcal{M}_g$  for  $g(\cdot)$  is correctly specified, i.e. contains the truth, even if one is not. This influence function is efficient in that, when  $m(\cdot)$  is correctly specified, it has the smallest variance among all influence functions in model  $\mathcal{M}_g$ .

The efficient influence function in a model that does not make any distributional assumptions about **C**, that is under assumptions (A1) and (A4) only, is given in equation (6) below.

$$D'(\mathbf{o}) = \frac{1}{n} \sum_{i=1}^{n} \left( E[m(V_i^*) \mid \mathbf{C} = \mathbf{c}] - \psi + \frac{\bar{h}_{x^*}(v_i)}{\bar{h}(v_i)} \{y_i - m(v_i)\} \right).$$
 (6)

We use this influence function in the estimating procedure below. This is also the influence function used to derive estimators conditional on  $\mathbf{C}$ , in which case the first two terms cancel out; we will denote the conditional influence function wit  $D_c(\mathbf{o})$ .

In social network settings  $\mathbf{C}$  is likely to be high-dimensional – recall that  $C_i$  includes all relevant covariates for individual i and for i's social contacts, in addition to a list of network ties. Even though we assume that  $m(\cdot)$  and  $g(\cdot)$  depend on  $\mathbf{C}$  only through the dimension-reducing functions  $s_Y(\cdot)$  and  $s_X(\cdot)$ , if  $m(\cdot)$  and  $g(\cdot)$  are estimated nonparametrically care should be taken to ensure rates of convergence satisfying the regularity conditions of Theorem 1. When  $m(\cdot)$  and  $g(\cdot)$  are estimated with working models, the doubly robust property is advantageous, affording two independent opportunities for valid inference about  $\psi$ . If models for both  $m(\cdot)$  and  $g(\cdot)$  are correctly specified, then the solution to estimating equations based on the efficient influence function are asymptotically efficient in the semiparametric model defined by the identifying assumptions and by the correctly specified model for  $g(\cdot)$ . See Van der Laan and Robins (2003); Tsiatis (2007) for a review of semiparametric inference in

i.i.d. settings.

A traditional estimating equation approach to inference about  $\psi$  has two disadvantages in finite samples. First, it does not result in substitution estimator, leading to the possibility that finite-sample estimates will fall outside of the parameter space when the parameter space for  $\psi$  is bounded. Second, it is sensitive to extreme values of the term  $\bar{h}_{i,x^*}(v)/\bar{h}(v_i)$  in the righthand side of Equation (5). For these reasons, we propose a targeted maximum loss-based estimator (TMLE) of  $\psi$ ; TMLEs are substitution estimators and are not as sensitive to the near violations of the positivity assumption that can occur in finite samples and result in extreme values of  $\bar{h}_{x^*}(v_i)/\bar{h}(v_i)$ .

Targeted maximum likelihood estimation is a general template for estimation of smooth parameters in semi- and nonparametric models. The estimation algorithm is constructed to solve the efficient influence function estimating equation, thereby yielding, under regularity conditions, asymptotically linear estimators with the same semiparametric efficiency property as the estimating equation approach described above. In our setting, a TMLE is constructed using three elements: (i) a valid loss function L for the outcome regression model m, (ii) initial working estimators  $\hat{m}$  of m and  $\hat{g}$  of g, and (iii) a parametric submodel  $m_{\epsilon}$  of  $\mathcal{M}$ , the score of which corresponds to a particular component of the score based on the efficient influence function  $D(\mathbf{o})$  and such that  $m_0 = m(\cdot)$ . The TMLE is then defined by an iterative procedure that, at each step, estimates  $\epsilon$  by minimizing the empirical risk of the loss function L at  $m_{\epsilon}$ . An updated estimate is then computed as  $\hat{m}_{\hat{\epsilon}}$ , and the process is repeated until convergence. The TMLE is the estimator obtained in the final step of the iteration. The result of the previous iterative procedure is that, at the final step, the efficient influence function estimating equation is solved. For more details about targeted maximum likelihood estimation, see Van der Laan and Rose (2011). Below we describe a TMLE for  $\psi$  based on (6) that requires only one iteration for convergence. We use influence function  $D'(\mathbf{o})$  to derive the TMLE, instead of  $D(\mathbf{o})$ , because it is computationally more tractable and because the choice of influence function does not matter for the conditional parameter that we are interested in when latent variable dependence is present.

In order to use targeted maximum likelihood estimation to estimate  $\psi$ , the parameter identifying  $E\left[\bar{Y}^*\right]$ , initial estimators  $\hat{m}$  and  $\hat{g}$  of m and g may be found through standard maximum likelihood or loss-based estimation methods. For example, if  $Y_i$  is binary, or if it is continuous and bounded and rescaled to fall between 0 and 1, then the negative log likelihood function

$$\sum_{i=1}^{n} \log \left\{ m(V_i)^{Y_i} (1 - m(V_i))^{1 - Y_i} \right\}$$

is a valid loss function for m. Assuming a parametric model  $m_{\theta}(\cdot)$  for  $m(\cdot)$ , a maximum likelihood estimator is given by  $m_{\hat{\theta}}$ , where

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log \{m_{\theta}(V_i)^{Y_i} (1 - m_{\theta}(V_i))^{1 - Y_i}\}$$

may be estimated by running standard binary regression software of the outcomes  $Y_i$  on  $V_i$ . An initial estimator  $\hat{g}$  may be found analogously by optimizing the log likelihood function  $\sum_{i=1}^{n} \log g(X_i|W_i)$ . Alternatively, these loss functions may be used to implement other loss-based estimation techniques such as super learning (van der Laan Mark et al., 2007), which selects the weighted average of a

collection ("library") of estimators that minimizes the cross-validated loss-based risk; the library may include any machine learning algorithms and logistic regression estimators. The empirical distribution  $\hat{p}_C$  is used to estimate  $p_C$ .

Estimation of  $\bar{h}$  and  $\bar{h}_x$  can be carried out by substituting  $\hat{g}$  and  $\hat{p}_C$  for g and  $p_C$  in the expressions

$$\bar{h}(v) = \frac{1}{n} \sum_{j} \sum_{\mathbf{c}, \mathbf{x}} I(s_{Y,j}(\mathbf{c}, \mathbf{x}) = v) g(\mathbf{x} | \mathbf{w}) p_C(\mathbf{c}) \text{ and}$$

$$\bar{h}_{x^*}(v) = \frac{1}{n} \sum_{j} \sum_{\mathbf{c}} I(s_{Y,j}(\mathbf{c}, \mathbf{x}^*) = v) p_C(\mathbf{c}),$$

where I(E) is the indicator function of the event E. We denote by  $\hat{\bar{h}}$  and  $\hat{\bar{h}}_{x^*}$  the corresponding estimates of  $\bar{h}$  and  $\bar{h}_{x^*}$ .

An alternative and computationally more feasible approach is to estimate  $\bar{h}(v)$  by directly optimizing the log likelihood function  $\sum_{i=1}^{n} \log \bar{h}(V_i|W_i)$ , as if the pooled sample  $(V_i, W_i)$  were i.i.d. It can be shown that this results in a valid loss function for  $\bar{h}$ , even for dependent observations  $(V_i, W_i)$ , for  $i = 1, \ldots, n$ . Similarly, one can construct a direct estimator of  $\bar{h}_{x^*}$ , by first creating a sample  $(V_i^*, W_i)$  and then directly optimizing the log likelihood function  $\sum_{i=1}^{n} \log \bar{h}_{x^*}(V_i^*|W_i)$ , as if the pooled sample  $(V_i^*, W_i)$  were i.i.d.

Now the targeted minimum loss based estimator of  $\psi$  is computed as follows:

1. Define the auxiliary weights  $H_i$  as the ratio of estimated densities of  $V^*$  and V evaluated at the observed value  $V_i$ . Compute the auxiliary weights as

$$H_i = \frac{\hat{\bar{h}}_{x^*}(V_i)}{\hat{\bar{h}}(V_i)}.$$

- 2. Compute initial predicted outcome values  $\hat{Y}_i = \hat{m}(V_i)$  and predicted potential outcome values  $\hat{Y}_i^* = \hat{m}(V_i^*)$  evaluated at the counterfactual value  $V_i^* = s_{Y,i}(\mathbf{C}, \mathbf{x}^*)$ .
- 3. Construct a TMLE model update  $\hat{m}_{\hat{\epsilon}}$  of  $\hat{m}$  by running a weighted intercept-only logistic regression model with weights  $H_i$  defined in step (1),  $Y_i$  as the outcome and including  $\hat{Y}_i$  as an offset. That is, define  $\hat{\epsilon}$  as the estimate of the intercept parameter  $\epsilon$  from the following weighted logistic regression model

$$\operatorname{logit} \hat{m}_{\epsilon}(v) = \operatorname{logit} \hat{m}(v) + \epsilon,$$

where 
$$logit(x) = log(\frac{x}{1-x})$$
.

4. Compute updated predicted potential outcomes  $\tilde{Y}_i^*$  as the fitted values of the regression from step 3, evaluated at  $v^*$  rather than v (that is, at  $\hat{Y}_i^*$  instead of  $\hat{Y}_i$ ):

$$\tilde{Y_i}^* = \text{expit}\{\text{logit}\hat{Y}_i^* + \hat{\epsilon}\},\$$

where  $\operatorname{expit}(x) = \frac{1}{1+e^{-x}}$ , i.e., the inverse of the logit function.

5. Compute the TMLE  $\hat{\psi}$  as

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y_i}^*.$$

The TMLE inherits the double robustness property of the estimating equation estimator we described above: it will be consistent for  $\psi$  if either the working model  $\hat{g}$  for g or the working model  $\hat{m}$  for m is correctly specified. This resulting estimator remains CAN for  $\psi$  under assumptions (A2) and (A3) instead of (A4) and (A5), and the same procedure can be used to estimate the parameter conditional on  $\mathbb{C}$ .

## 3.4 Asymptotic normality

van der Laan (2014) proved the asymptotic normality of the estimators described above under assumptions that the degree  $K_i$  is uniformly bounded above as  $n \to \infty$  by a constant K. In order to accommodate more realistic models of asymptotic growth in the network context, we consider a new asymptotic regime in which  $K_i$  may grow as  $n \to \infty$ .

**Theorem 1:** Let  $K_{max,n} = max_i\{K_i\}$  for a fixed network with n nodes. Suppose that  $K_{max,n}^2/n \to 0$  as  $n \to \infty$ . If at least one of the two models for  $m(\cdot)$  and  $g(\cdot)$  is correctly specified and under independence assumptions (A1) through (A3b), positivity assumption (A6), and regularity conditions (see Appendix),

$$\sqrt{C_n} \left( \hat{\psi} - \psi \right) \stackrel{d}{\longrightarrow} N(0, \sigma^2),$$

 $n/K_{max,n}^2 \le C_n \le n$ . The asymptotic variance of  $\hat{\psi}$ ,  $\sigma^2$ , is given by the variance of the influence curve of the estimator.

The proof of Theorem 1 is in the Appendix. In Section 4.4, below, we discuss settings in which the conditions for this theorem fail to hold, and ways to recover valid inference for conditional estimands in some of these settings. Broadly, the proof has two parts: first, to show that the second order terms in the expansion of  $\hat{\psi} - \psi$  are stochastically less than  $1/\sqrt{C_n}$ , and second, to show that the first order terms converge to a normal distribution when scaled by a factor of order  $\sqrt{C_n}$ . The proof that the second order terms are stochastically less than  $1/\sqrt{C_n}$  is an extension of the empirical process theory of Van Der Vaart and Wellner (1996) and follows the same format as the proof in van der Laan (2014). For the proof that the first order terms converge to a normal distribution, we rely on Stein's method of central limit theorem proof (Stein, 1972). Stein's method allows us to derive a bound on the distance between our first order term (properly scaled) and a standard normal distribution; this bound depends on the degree distribution  $K_1, ..., K_n$ . We show that this bound converges to 0 as  $n \to \infty$  under regularity conditions and our running assumption that  $K_{max,n}^2 = o(n)$ .

When all nodes have the same number of ties, i.e.  $K_i = K_{max,n}$  for all i, then the rate of convergence will be given  $\sqrt{C_n} = \sqrt{n/K_{max,n}^2}$ . When  $K_{max,n}$  is bounded above as  $n \to \infty$ , then the rate of convergence will be  $\sqrt{n}$ . This is the setting covered in van der Laan (2014). When  $K_{max,n} \to \infty$  but some nodes have fewer than  $K_{max,n}$  ties, the exact rate of convergence is between  $\sqrt{n/K_{max,n}^2}$  and  $\sqrt{n}$  (inclusive) but is difficult or impossible to determine analytically, as it may depend intricately on the topology of the network. The inferential procedures that we describe below do not require

knowledge of the rate of convergence.

#### 3.5 Inference

A 95% confidence interval for  $\psi$  is given by  $\hat{\psi}_n \pm 1.96\sigma/\sqrt{C_n}$ . In practice neither  $\sigma$  nor  $C_n$  are likely to be known, but available variance estimation methods estimate the variance of  $\hat{\psi}_n$  directly, incorporating the rate of convergence without requiring it to be known a priori.

In i.i.d settings, the variance of doubly robust estimators can be bootstrapped or, under the assumption that both nuisance models are correctly specified, estimated with the plug-in estimator of the square of the estimator's influence function. When the nuisance parameters are estimated with parametric models, standard sandwich variance estimators are appropriate whether one or both of the nuisance models are correctly specified. In the present, non-i.i.d. setting, variance estimation is considerably more challenging. In principle, the variance of  $\hat{\psi}$  can be estimated using the empirical average of the square of the influence function, substituting  $\hat{\psi}$  for  $\psi$  and the fitted values from the working models  $\hat{g}$  and  $\hat{m}$  for g and g. Although this variance may be anticonservative if one, but not both, of the working models  $\hat{g}$  and  $\hat{m}$  is correctly specified, using flexible or non-parametric specifications for these models increases opportunities to estimate both consistently. However, unlike in i.i.d. settings, the expectation of the square of the empirical version of the influence function given in Equation (5) does not reduce to the sum of squared influence terms for each observation. Instead, it includes double sums for all pairs of observations that are not marginally independent of one another. These terms capture covariances between dependent observations; these extra covariance terms reflect a larger variance and a slower rate of convergence due to dependence across observations.

When dependence is due to direct transmission, that is, under assumptions (A1), (A4), and (A5), two alternative variance estimation procedures are available. One option is to estimate the variance of the influence function  $D'(\mathbf{o})$  given by Equation (6). Our TMLE is based on  $D'(\mathbf{o})$ , but because this is the efficient IF in a model that makes fewer assumptions than (A1), (A4), and (A5), it has larger variance than  $D(\mathbf{o})$  and provides a valid (asymptotically conservative) variance estimate even when estimation is based on  $D(\mathbf{o})$ . For consistent and computationally feasible estimators for the variance of  $D'(\mathbf{o})$  see Sofrygin and van der Laan (2015).

An alternative approach to estimate the variance of  $\hat{\psi}$  under assumptions (A1), (A4), and (A5) is to employ the following version of a parametric bootstrap. This approach might offer improvements in finite-sample performance over the previously described approach. Briefly, our proposed procedure proceeds by iteratively sampling n observations from the existing fit of the likelihood, fitting the univariate least favorable submodel parameter  $\epsilon$ , and computing the bootstrapped TMLE. After enough iterations one can obtain the Monte Carlo variance estimator of  $\hat{\psi}$  by evaluating the empirical variance of the bootstrap TMLEs. In more detail, for each of M bootstrap iterations, indexed as  $b=1,\ldots,M$ , first n covariates  $\mathbf{C}^b=(C_1^b,\ldots,C_n^b)$  are sampled with replacement (assuming  $\mathbf{C}$  are i.i.d.), then the existing model fit  $\hat{g}$  is applied to sampling of n exposures  $\mathbf{X}^b=(X_1^b,\ldots,X_n^b)$ , followed by a sample of n outcomes  $\mathbf{Y}^b=(Y_1^b,\ldots,Y_n^b)$  based on the existing outcome model fit  $\hat{m}$ . Note that we are also assuming that the corresponding bootstrap random summaries  $W_i^b$  and  $V_i^b$ , for  $i=1,\ldots,n$ , were constructed by applying the summary functions  $s_X$  and  $s_Y$  to  $\mathbf{C}^b$  and  $(\mathbf{C}^b,\mathbf{X}^b)$ , respectively. This bootstrap sample is then applied to obtain the predicted values from the existing auxiliary covariate

fit  $(\hat{h}_{x^*}/\hat{h})(V_i^b)$ , for  $i=1,\ldots,n$ , followed by a bootstrap-based fitting of  $\epsilon$ , and finally, evaluation of bootstrap TMLE. Note that the TMLE model update is the only model fitting step needed at each iteration of the bootstrap, which significantly lowers the computational burden of this procedure. The variance estimate is then obtained by taking the empirical variance of bootstrap TMLE samples  $\hat{\psi}^b$ . In the present setting, due to dependence across observations, one must be judicious with applications of the bootstrap. For example, the parametric bootstrap procedure described above requires conditional independence of  $X_i$  given  $W_i$  and  $Y_i$  given  $V_i$ , along with the consistent modeling of the corresponding factors of the likelihood. It may seem natural to sample  $V_i$  directly from its corresponding auxiliary model fit, but this is likely to result in an anti-conservative variance estimates, since the conditional independence of  $V_i$  is unlikely to hold by virtue of its construction as a summary measure of the network.

When latent variable dependence is present, that is under assumptions (A1) through (A3), consistent and computationally feasible variance estimation procedures are not currently available for either  $D'(\mathbf{o})$  or  $D(\mathbf{o})$ , because existing methods require bootstrapping some of the observed data. Without latent variable dependence we can take advantage of marginal and conditional independences to employ i.i.d. or parametric bootstrap methods, but latent variable dependence requires new methods for dependent data bootstrap. For this reason, we instead estimate the conditional parameter with influence function  $D_C(\mathbf{o})$ . A simple plug-in estimator is available for the variance of this influence function (see the Appendix and van der Laan, 2014).

# 4 Extensions

In this section we extend the estimation procedure to two causal effects of great interest in the context of social networks: to social contagion, or peer effects; to stochastic interventions; and to interventions on the network topology itself, i.e. interventions on  $\mathbf{A} = [A_{ij} : i, j \in \{1, ..., n\}]$  where, as above,  $A_{ij} \equiv I$  {subjects i and j share a tie}. First we introduce dynamic and stochastic interventions.

### 4.1 Dynamic and stochastic interventions

Interventions that assign treatment as a user-specified function  $d_X(\cdot)$  of  $\mathbb{C}$  correspond to substituting  $d_{X,i}(\mathbb{C})$  for  $x_i^*$  in the intervention model, definitions, and estimating procedure above. This is an example of a "dynamic" intervention: treatment is deterministically specified conditional on covariates but is but allowed to depend ("dynamically") on covariates.

We can also identify the effects of interventions that replace  $f_X$  with a new, user-specified function. This is an example of a stochastic intervention: the intervention changes the distribution of X but does not eliminate the stochasticity introduced by  $\varepsilon_X$ ; it is represented by an intervention SEM that replaces the equation for  $X_i$  with  $X_i^* = r_X [W_i^*, \varepsilon_{X_i}]$  for a user specified function  $r_X$ . For discussions of stochastic interventions and their identifying assumptions in i.i.d. settings, see Muñoz and van der Laan (2012); Haneuse and Rotnitzky (2013); Young et al. (2014). In the social network setting, stochastic interventions that change the dependence of  $X_i$  on C and of and  $Y_i$  on C and C are of particular interest. For example, consider data generated by an SEM in which  $f_X$  depends on C only through  $W_i = \frac{1}{|A_i|} \sum_{j:A_{ij}=1} C_j$ , i.e. the mean of C among the set of alters of i. We might be interested in

the mean counterfactual outcome under a stochastic intervention that forces  $f_X$  to depend instead on  $W_i^* = \max_{j:A_{ij}=1} \{C_j\}$ , i.e. the maximum value C among the alters of i. This particular stochastic intervention modifies  $f_X$  only through W; it is represented by an intervention SEM that replaces the equation for  $X_i$  with  $X_i^* = f_X[W_i^*, \varepsilon_{X_i}]$ . This is not a deterministic intervention, rather it assigns X according to a stochastic distribution: for each x in the support of X,  $X_i$  is set by the intervention to x with probability  $P[X = x|W = \max_{j:A_{ij}=1} \{C_j\}]$ .

We formally define the class of stochastic interventions that alter the dependence of  $X_i$  on  $\mathbf{C}$  and of and  $Y_i$  on  $(\mathbf{C}, \mathbf{X})$ , discuss identifying assumptions and estimation procedures, and then describe some such interventions of particular interest. Let  $s_{X,i}^*(\cdot)$  and  $s_{Y,i}^*(\cdot,\cdot)$  be user-specified functionals. They are denoted by an asterisk because they index hypothetical interventions rather than realized data-generating mechanisms. Let  $W_i^* = s_{X,i}^*(\mathbf{C})$  and  $V_i^* = s_{Y,i}^*(\mathbf{C}, \mathbf{X}^*)$ . We are concerned with the class of stochastic interventions given by

$$C_{i} = f_{C} \left[ \varepsilon_{C_{i}} \right]$$

$$i = 1, \dots, n$$

$$X_{i}^{*} = f_{X} \left[ W_{i}^{*}, \varepsilon_{X_{i}} \right]$$

$$i = 1, \dots, n$$

$$Y_{i}^{*} = f_{Y} \left[ V_{i}^{*}, \varepsilon_{Y_{i}} \right]$$

$$i = 1, \dots, n.$$

$$(7)$$

This can be interpreted as an intervention where, for each  $x^*$  in the support of X and for i=1,...,n,  $X_i$  is set to  $x^*$  with probability  $P\left[X=x^*|W=s_{X,i}^*(\mathbf{C})\right]$  and  $V_i$  is set to  $s_{Y,i}^*(\mathbf{C},\mathbf{x}^*)$  deterministically for each possible realization  $\mathbf{x}^*$ . Because Y depends on  $\mathbf{X}$  only through V, this is equivalent to an intervention that sets  $V_i$  to v with probability  $P\left[\mathbf{X} \in \left\{\mathbf{x}^*: s_{Y,i}^*(\mathbf{C},\mathbf{x}^*)=v\right\} \mid \mathbf{W}=\mathbf{s}_X^*(\mathbf{C})\right]$ , where  $\mathbf{s}_X^*(\mathbf{C})=\left(s_{X,1}^*(\mathbf{C}),...,s_{X,n}^*(\mathbf{C})\right)$ .

This intervention is identified under the same assumptions as the deterministic interventions described above, with the exception of a positivity assumption that is a slight modification of (A6). Define  $\mathcal{X}^* = \{\mathbf{x}^* : P\left[\mathbf{X} = \mathbf{x}^* | \mathbf{W} = \mathbf{s}_X^*(\mathbf{C})\right] > 0\}$  to be the set of treatment vectors  $\mathbf{x}^*$  that have positive probability under the stochastic intervention defined by (7). We assume that

$$min_{v \in \mathcal{V}^*} P(V = v | \mathbf{C} = \mathbf{c}) > 0 \text{ for } \mathcal{V}^* = \left\{ s_{Y,i}^*(\mathbf{C}, \mathbf{x}^*) : \mathbf{x}^* \in \mathcal{X}^* \right\} \text{ and for all } \mathbf{c} \text{ in the support of } \mathbf{C}. \tag{8}$$

Note that, in order for this positivity assumption to hold, the supports of  $s_X^*(\cdot)$  and  $s_Y^*(\cdot,\cdot)$  must be of the same dimensions as the supports of  $s_X(\cdot)$  and  $s_Y(\cdot,\cdot)$ , respectively.

The causal parameter of interest is the expected average potential outcome under this hypothetical intervention,  $E\left[\bar{Y}^*\right]$ . Define  $h_i^*(v) = P\left[V_i^* = v\right] = P\left[s_{Y,i}^*(\mathbf{C}, \mathbf{X}^*) = v\right]$ . Then  $E\left[\bar{Y}^*\right]$  is identified by

$$\psi = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathbf{c}, \mathbf{x}} E\left[Y_i | s_{Y,i}^*(\mathbf{c}, \mathbf{x})\right] P\left[\mathbf{X} = \mathbf{x} | \mathbf{W} = \mathbf{s}_X^*(\mathbf{c})\right] p_C(\mathbf{c})$$
$$= \frac{1}{n} \sum_{i=1}^{n} E\left[m(V_i^*)\right] = \frac{1}{n} \sum_{i=1}^{n} \sum_{v} m(v) h_i^*(v).$$

An influence function for  $\psi$ , evaluated at a fixed value of the observed data,  $\mathbf{o}$ , is given by

$$D(\mathbf{o}) = \sum_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} E[m(V_i^*) \mid C_j = c_j] - \psi + \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{h}^*(v_i)}{\bar{h}(v_i)} \{y_i - m(v_i)\},$$

where  $\bar{h}^*(v_i) = \frac{1}{n} \sum_{j=1}^n h_j^*(v_i)$ . (This is the efficient influence function under assumptions (A4) and (A5).) Estimation of  $\bar{h}^*$  is carried out by substituting  $\hat{g}$  and  $\hat{p}_C$  for g and  $p_C$  in the expression

$$\bar{h}^*(v) = \frac{1}{n} \sum_{i} \sum_{\mathbf{c}, \mathbf{x}} I\left(s_{Y, i}^*(\mathbf{c}, \mathbf{x}) = v\right) g\left(\mathbf{x} | \mathbf{s}_X^*(\mathbf{c})\right) p_C(\mathbf{c}).$$

Since  $\hat{p}_C$  is an empirical distribution that puts mass one on the observed value  $\mathbf{c}$ , the estimator  $\hat{h}^*$  reduces to

$$\hat{\bar{h}}^*(v) = \frac{1}{n} \sum_{i=1}^n \sum_{\mathbf{x}} I\left(s_{Y,i}^*(\mathbf{x}, \mathbf{C}) = v\right) \hat{g}(\mathbf{x}|\mathbf{s}_X^*(\mathbf{C})).$$

We denote by  $\hat{\bar{h}}$  and  $\hat{\bar{h}}^*$  the corresponding estimates of  $\bar{h}$  and  $\bar{h}^*$ .

Now the targeted minimum loss based estimator of  $\psi$  is computed according to the steps outlined in Section 3, but with  $V^*$  and  $Y^*$  defined as immediately above.

A special case of this class of stochastic interventions intervenes only on  $s_X$ , like the example discussed above in which the intervention forces  $f_X$  to depend on  $W_i^* = \max_{j:A_{ij}=1} \{C_j\}$  but does not alter the functional form of  $s_Y$ .  $E[\bar{Y}^*]$  under this type of intervention is identified by

$$\psi = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathbf{c}, \mathbf{x}} E[Y_i | \mathbf{C} = \mathbf{c}, \mathbf{X} = \mathbf{x}] P[\mathbf{X} = \mathbf{x} | \mathbf{W} = s_X^*(\mathbf{c})] p_C(\mathbf{c})$$
$$= \frac{1}{n} \sum_{i=1}^{n} E[m(V_i^*)] = \frac{1}{n} \sum_{i=1}^{n} \sum_{v} m(v) h_i^*(v).$$

With  $V_i^*$  defined as  $s_{Y,i}(\mathbf{C}, \mathbf{X}^*)$ , estimation of this class of intervention proceeds as immediately above. The fact that  $\mathbf{X}^*$  is random does not affect the estimation algorithm.

#### 4.2 Peer effects

Define  $Y_i^0$  to be the outcome variable measured at a time previous to the primary outcome measurement  $Y_i$ . Peer effects are the class of causal effects of  $Y_j^0$  on  $Y_i$  for  $A_{ij} = 1$ : the effects of individuals' outcomes on the subsequent outcome of their alters. We can operationalize peer effects as the effects of dynamic interventions where the counterfactual exposure for subject i is given by a user-specified function  $d_X(\cdot)$  of  $\{Y_j^0: A_{ij} = 1\}$ . In order to maintain the identifying assumptions A2b and A3b, the time elapsed between  $Y^0$  and Y must permit transmission only between nodes and their immediate alters. Otherwise, if between  $Y^0$  and Y the outcome could have spread contagiously more broadly, there will be more dependence present than our methods can account for, and also possible confounding of the effect of  $Y_i^0$  on  $Y_j$  for  $A_{ij} = 1$  due to mutual friends.

## 4.3 Interventions on the network topology

An intervention on the network, i.e. an intervention that adds, removes, or relocates ties in the network, is a special case of a joint intervention on  $s_X(\cdot)$  and  $s_Y(\cdot)$ . To see this, note that the network topology, codified by the adjacency matrix  $\mathbf{A}$ , enters the data-generating structural equation model (1) only through  $s_X(\cdot)$  and  $s_Y(\cdot)$ ; therefore we can represent any modification to  $\mathbf{A}$  via the corresponding modification to  $s_X(\cdot)$  and  $s_Y(\cdot)$ . Consider an intervention that replaces the observed adjacency matrix  $\mathbf{A}$  with a user-specified adjacency matrix  $\mathbf{A}^*$ . This is an example of the stochastic interventions described on page 19, with  $s_{X,i}^*(\mathbf{C})$  replaced by  $s_{X,i}^{\mathbf{A}^*}(\mathbf{C}) \equiv s_{X,i}\left(\left\{C_j: A_{ij}^* = 1\right\}\right)$  and  $s_{Y,i}^*(\mathbf{C}, \mathbf{X}^*)$  by  $s_{Y,i}^{\mathbf{A}^*}(\mathbf{C}, \mathbf{X}^*) \equiv s_{Y,i}\left(\left\{X_j^*: A_{ij}^* = 1\right\}, \left\{C_j: A_{ij}^* = 1\right\}\right)$ . The intervention SEM differs from the data-generating SEM only in that  $X_i$  depends on the covariate values for the individuals with whom i shares ties in the intervention adjacency matrix  $\mathbf{A}^*$  and  $Y_i$  depends on the counterfactual treatments and observed covariate values for those same individuals.

Interventions on summary features of the adjacency matrix can also be viewed as stochastic interventions. Instead of replacing  $\mathbf{A}$  with  $\mathbf{A}^*$ , an intervention on features of the network topology replaces  $\mathbf{A}$  with the members of a class  $\mathcal{A}^*$  of  $n \times n$  adjacency matrices that share the intervention features, stochastically according to some probability distribution  $g_{\mathbf{A}^*}$  over  $\mathcal{A}^*$ . For example, we might be interested in interventions that constrain the degree distribution of the network, e.g. fixing the maximum degree to be smaller than some number D or ensuring that no node has degree equal to 0. We might specify  $g_{\mathbf{A}^*}(A) = \frac{1}{|\mathcal{A}^*|} I\left\{A \in \mathcal{A}^*\right\}$ , giving equal weight to each realization in the class  $\mathcal{A}^*$ . Effectively, this kind of intervention sets  $V_i$  to v with probability  $P\left[\mathbf{X} \in \left\{\mathbf{x}^*: s_{Y,i}^{\mathbf{T}^*}(\mathbf{C}, \mathbf{x}^*) = v\right\} \mid \mathbf{W} = \mathbf{s}_X^{\mathbf{A}^*}(\mathbf{C}) \text{ for some } \mathbf{A}^* \in \mathcal{A}^*\right]$ , where  $\mathbf{s}_X^{\mathbf{A}^*}(\mathbf{C}) = \left(s_{X,1}^{\mathbf{A}^*}(\mathbf{C}), ..., s_{X,n}^{\mathbf{A}^*}(\mathbf{C})\right)$ .

As with the stochastic interventions discussed in the previous section, positivity is a crucial assumption for identifying interventions on  $\mathbf{A}$ . The support of  $V^*$  must be the same as the support of V. If replacing  $\mathbf{A}$  with  $\mathbf{A}^*$  (either deterministically or as a random selection from the class  $\mathcal{A}^*$ ) assigns to unit i a value of V that not observed in the real data for a unit in the same  $\mathbf{C}$  stratum as i, then the effect of the intervention that that replaces  $\mathbf{A}$  with  $\mathbf{A}^*$  is not identified for unit i.

We can expand the arguments of  $s_X(\cdot)$  and  $s_Y(\cdot)$  to include network topology explicitly, in addition to  $\mathbb{C}$ , or we can include features of network topology as individual-level covariates. For example, we could include in  $C_i$  the total number of ties for subject i has, a measure of subject i's centrality in the network, or a local clustering coefficient for the neighborhood comprised of subject i's contacts. Whether or not we can define, identify, and estimate interventions involving these features of network topology hinges crucially on the positivity assumption.

Both the degree of subject i and local clustering around subject i are local features of network topology: they depend on  $\mathbf{T}$  only through subject i and subject i's immediate contacts. A local clustering coefficient for node i can be defined as the proportion of potential triangles that include i as one vertex and that are completed, or the number of pairs of neighbors of i who are connected divided by the total number of pairs of neighbors of i (Newman, 2009). This measure of triangle completion captures the extent to which "the friend of my friend is also my friend": triangle completion is high whenever two subjects who share a mutual contact are more likely to themselves share a tie than are two subjects chosen at random from the network. Positivity could hold if, within each level of  $\mathbf{C}$ , subjects were observed to have a wide range of degrees and of triangle completion among their

contacts.

In contrast with degree and local clustering, network centrality is a node-specific attribute that nevertheless depends on the entire network topology. It captures the intuitive notion that some nodes are central and some nodes are fringe in any given network. It can be measured in many different ways, based, for example, on the number of network paths that intersect node i, on the probability that a random walk on the network will intersect node i, or on the mean distance between node i and the other nodes in network (see Chapter 7 of Newman, 2009 for a comprehensive discussion of these and other centrality measures). Centrality is given by a univariate measure for each node in a network, but each node's measure depends crucially on the entire graph. In reality it is not generally possible to intervene on centrality without altering the entire adjacency matrix  $\mathbf{A}$ , which is likely to result in changes to  $s_X(\cdot)$  and  $s_Y(\cdot)$ . Unless it is hypothetically possible to intervene on centrality without changing any other input into the SEM, in this case  $s_X(\cdot)$  and  $s_Y(\cdot)$ , interventions on centrality cannot be identified.

## 4.4 Too many friends, too much influence

The conditions in the theorem on page 16 will be violated for any asymptotic regime in which the degree of any one or more nodes grows at the same or similar rate to the sample size n. This is problematic because social networks frequently have a small number of "hubs"—that is, nodes with very high degree (Newman, 2009), and the occurrence of hubs is a feature of many of the network-generating models that have been proposed for social networks. When a small number of individuals wield influence over a significant portion of the rest of the population, two problems arise for statistical inference about the people connected in the underlying network. First, the number of hubs may grow slowly with n or it could be fixed and not grow at at all. If the hubs are systematically different from the rest of the population (e.g. they have different covariates values or the exposure affects them differentially), then a fixed or slowly growing number of hubs would not allow for consistent inference about this distinct subpopulation. Second, and more importantly, the sweeping influence of hubs creates dependence among all of the influenced nodes that undermines inference. Our methods rely on the independence of  $Y_i$  and  $Y_j$  whenever nodes i and j do not share a tie or a mutual alter. When hubs are present, a significant proportion of nodes will share a connection to one of these hubs, undermining our methods.

We can recover valid inference using our methods if we condition on the hubs, treating them as features of the background network environment rather than as observations. This results in different causal effects or statistical estimands, as all of our inference is conditional on the identity and characteristics of the hubs. Imagine a social network comprised of the residents of a city in which a cultural or political leader is connected to almost all of the other nodes. It may be impossible to disentangle the influence of this leader, which affects every other node, from other processes simultaneously occurring among the other residents of the city. It will certainly be impossible to statistically learn about the hub, as the sample size for the hub subgroup is 1. But it may make sense to consider the hub as a feature of the city rather than a member of the network. We could then learn about other processes occurring among the other residents of the city, conditional on the behavior and characteristics of the leader. For example, we could evaluate the effect of a public health initiative encouraging residents to talk to their friends about the importance of exercise, but we could not evaluate a similar program

targeting the leader's communication about exercise. It is easy to see that, in the latter case, positivity would be violated as there would be no control group of subjects unexposed to the leader's communications. Equally important is the fact that conditioning on the leader is necessary to engender the independence among units required for our theoretical results to hold.

Practically speaking, for real and finite datasets, this implies that the methods we have proposed are inappropriate for networks in which the degree is large, compared to n, for one or more nodes. If many nodes are connected to a significant fraction of other nodes, this problem is intractable. However, if only a small number of nodes are highly connected we can condition on them to recover approximately valid inference using our methods for conditional estimands. There is a theoretical tradeoff between the rate of convergence of our estimators and the order of K relative to n that, in finite samples, becomes a practical tradeoff between generality and variance. Increasing the number of nodes classified as hubs (and conditioned on in subsequent analysis) will increase the rate of convergence by decreasing the order of K for the remaining, non-hub nodes – assuming that the number of hubs remains small compared to n so that the sample size does not decrease significantly when we exclude hubs from the analysis. On the other hand, classifying more nodes as hubs results in analyses that are increasingly specific: conditioning on a single hub may preserve generalizability to other networks (similar cities with similar leaders), but conditioning on many hubs is likely to limit the generalizability of the resulting inference.

# 5 Simulations

We conducted a simulation study that evaluated the finite sample and asymptotic behavior of the TMLE procedure described in Section 3.3. We generated social networks according to the preferential attachment model (Barabási and Albert, 1999), where the node degree (number of friends) distribution followed a power law with  $\alpha=0.5$ , and according to the small world network model (Watts and Strogatz, 1998) with a rewiring probability of 0.1. We generated data with two different types of dependence: first with dependence due to direct transmission only, and second with both latent variable dependence and dependence due to direct transmission.

Our simulations mimicked a hypothetical study designed to increase the level of physical activity in a population comprised of members of a social network. For each community member indexed by  $i=1,\ldots,n$ , the study collected data on i's baseline covariates, denoted  $C_i$ , which included the indicator of being physically active, denoted  $PA_i$  and the network of friends on each subject,  $F_i$ . The exposure or treatment,  $X_i$ , was assigned randomly to 25% of the community. For example, one can imagine a study where treated individuals received various economic incentives to attend a local gym. The outcome  $Y_i$  was a binary indicator of maintaining gym membership for a predetermined follow-up period. We assumed that it was of interest to examine and estimate the average of the mean counterfactual outcomes  $E\left[\bar{Y}^*\right]$  under various hypothetical interventions  $g^*$  on such a community. First, we considered a stochastic intervention  $g_1^*$  which assigned each individual to treatment with a constant probability of 0.35; this differs from the observed allocation of treatment to 25% of the community members. We also considered a scenario in which the aforementioned economic incentive was resource constrained and could only be allocated to up to 10% of community members and estimated the effects of various targeted approaches to allocating the exposure. For

example, we considered an intervention  $g_2^*$  that targeted only the top 10% most connected members of the community, as such a targeted intervention would be expected to have a higher impact on the overall average probability of maintaining gym membership among the community, when compared to purely random assignment of exposure to 10% of the community. Another hypothetical intervention  $g_3^*$  assigned an additional physically active friend to individuals with fewer than 10 friends. Notably,  $g_3^*$  can be thought of as intervening on the structure of the social network itself. Finally, we estimated the combined effect of simultaneously implementing intervention  $g_2^*$  and the network-based intervention  $g_3^*$  on the same community; this is  $g_4^*$ . For simplicity, this simulation study only reports the expected outcome under each of these interventions; causal effects defined as contrasts of these interventions can be easily estimated based on the same methods.

All simulation and estimation was carried out in R language (R Core Team, 2015) with packages simcausal (Sofrygin et al., 2015) and tmlenet (Sofrygin and van der Laan, 2015). The full R code for this simulation study is available in a separate github repository<sup>1</sup>. The simulations were repeated for community sizes of n = 500, n = 1,000 and n = 10,000. The estimation was repeated by sampling 1,000 such datasets, conditional on the same network (sampled only once for each sample size). For the simulations with dependence due to direct transmission, the baseline covariates were independently and identically distributed. The probability of success for each  $Y_i$  was a logit-linear function of i's exposure  $X_i$  (indicator of receiving the economic incentive), the baseline covariates  $C_i$  and the three summary measures of i's friends exposures and baseline covariates. In particular, we also assumed that the probability of maintaining gym membership increased on a logit-linear scale as a function of the following network summaries: the total number of i's friends who were exposed  $(\sum_{j:A_{ij}=1} X_j)$ , the total number of i's friends who were physically active at baseline  $(\sum_{j:A_{ij}=1} PA_j)$  and the product of the two summaries  $(\sum_{j:A_{ij}=1} X_j \times \sum_{j:A_{ij}=1} PA_j)$ . The economic incentive to attend local gym had a small direct effect on each individual who was not physically active at baseline and no direct effect on those who were already physically active. However, physically active individuals were more likely to maintain gym membership over the follow-up period if they had at least one physically active friend at baseline. We repeated these simulations with the addition of latent variable dependence, which we introduced by generating unobserved latent variables for each node which affected the node's own outcome as well as the outcomes of its friends.

We estimated the expected outcome under interventions  $g_1^*$  through  $g_4^*$  using the aforementioned TMLE approach and evaluated its finite sample bias. For the simulations under dependence due to direct transmission, we estimated the marginal parameter  $E\left[\bar{Y}^*\right]$  and compared three different estimators of the asymptotic variance and the coverage of the corresponding confidence intervals. First, we looked at the naive plug-in i.i.d. estimator ("IID Var") for the variance of the influence curve which treated observations as if they were i.i.d. Second, we used the plug-in variance estimator based on the efficient influence curve which adjusted for the correlated observations ("dependent IC Var"), as previously described in Sofrygin and van der Laan (2015). Finally, we used the parametric bootstrap variance estimator ("bootstrap Var"), where each bootstrap iteration sampled n observations from the previous fit of the likelihood (the estimated exposure model  $\hat{g}$  and the mean outcome model  $\hat{m}$ ) and used this bootstrap sample to re-fit the least favorable parametric submodel update  $\epsilon$ . The

<sup>&</sup>lt;sup>1</sup>github.com/osofr/socialnets.sim.study

simulation results showing the mean length and coverage of these three CI types are shown in Figure 2 (preferential attachment network) and Figure 3 (small world network). For the simulations with latent variable dependence, we estimated the conditional parameter  $E[\bar{Y}^*]$  and we compared two plug in variance estimators based on the conditional influence function  $D_C$ : one that assumes conditionally i.i.d outcomes (conditional on **X** and **C**), which would be true if all dependence were due to direct transmission but is violated in the presence of latent variable dependence ("IID Var"), and one that does not make this assumption ("dependent IC Var").

We examined the empirical distribution of the transformed TMLEs, comparing their histogram estimates against the predicted normal limiting distribution, with the results shown in Figure 4, where the histogram plots are displayed by sample size (horizontal axis) and the intervention type (vertical axis). The estimates were first centered at the corresponding true parameter values and then rescaled by their corresponding true standard deviation (SD). We note that our results indicate that the estimators converge to their normal theoretical limiting distribution, even in networks with power law node degree distribution, such as the preferential attachment network model, as well as in the densely connected networks obtained under the small world network model. The results shown in Figure 4 were generated from simulations with dependence due to direct transmission; simulations with latent variable dependence (not shown) evinced similar approximate normality.

One of the lessons of our simulation study is that by leveraging the structure of the network it might be possible to achieve a larger overall intervention effect on a population level (Harling et al., 2016). For example, the results in the left panels of Figures 2 and 3 show that by targeting the exposure assignment to highly connected and physically active individuals, intervention  $g_2^*$  increases the mean probability of sustaining gym membership compared to the similar level of un-targeted coverage of the exposure. We also demonstrated the feasibility of estimating effects of interventions on the observed network structure itself, such as intervention  $g_3^*$ , which can be also combined with economic incentives, as it was mimicked by our hypothetical intervention  $g_2^* + g_3^*$ . These combined interventions could be particularly useful in resource constrained environments, since they may result in larger community level effects at the lower coverage of the exposure assignment.

We evaluated the performance of the different approaches to variance estimation described above, as measured by the coverage of the 95% CIs. Our results from simulations with dependence due to direct transmission show that conducting inference while ignoring the nature of the dependence in such datasets generally results in anticonservative variance estimates and under-coverage of CIs, which can be as low as 50% even for very large sample sizes ("IID Var" in the right panels of Figures 2 and 3). The CIs based on the dependent variance estimates ("dependent IC Var" in the right panel of the same figures) obtain nearly nominal coverage of 95% for large enough sample sizes, but can suffer in smaller sample sizes due to lack of asymptotic normality and near-positivity violations. Notably, the CIs based on the parametric bootstrap variance estimates provide the most robust coverage for smaller sample sizes, while attaining the nominal 95% coverage in large sample sizes for nearly all of the simulation scenarios ("bootstrap Var" in the right panel of the same figures). The apparent robustness of the parametric bootstrap method for inference in small sample sizes, even as low as n = 500, was one of the surprising finding of this simulation study. Future work will explore the assumptions under which this parametric bootstrap works and its sensitivity towards violations of those assumptions. Similarly,



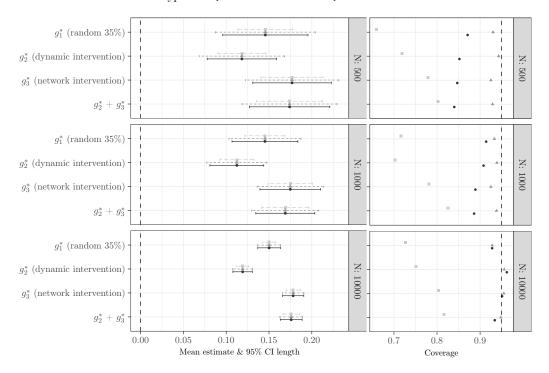


Figure 2: Mean 95% CI length (left panel) and coverage (right panel) for the TMLE in preferential attachment network with dependence due to direct transmission, by sample size, intervention and CI type. Results are shown for the estimates of the average expected outcome under four hypothetical interventions  $(g_1^*, g_2^*, g_3^*)$  and  $g_2^* + g_3^*$ .

in the simulations with latent variable dependence the variance estimates that assume conditionally i.i.d. outcomes, i.e. that dependence may be due to direct transmission but not to latent variables, are anti-conservative.

# 6 Conclusion

We proposed new methods that allow for causal and statistical inference using observations sampled from members of a single interconnected social network when the observations evince dependence due to network ties. In contrast to existing methods, our methods do not require randomization of an exogenous treatment and they have proven performance under asymptotic regimes in which the number of network ties grows (slowly) with sample size. In future work we plan to address a key limitation of the present proposal, namely the assumption that the network is observed fully and without error. We also plan to develop data-adaptive methods for estimating the summary measures  $s_X$  and  $s_Y$ , as it may be unreasonable to expect these to be known a priori. Finally, we plan to develop estimating algorithms for longitudinal settings; the influence function and asymptotic results for these settings are straightforward extensions of the results presented here, but estimation can be challenging.

#### CI.type - dependent IC Var - bootstrap Var - iid Var

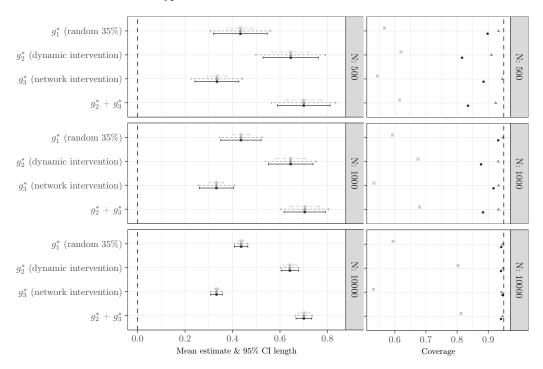


Figure 3: Mean 95% CI length (left panel) and coverage (right panel) for the TMLE in small world network with dependence due to direct transmission, by sample size, intervention and CI type. Results are shown for the estimates of the average expected outcome under four hypothetical interventions  $(g_1^*, g_2^*, g_3^*)$  and  $g_2^* + g_3^*$ .

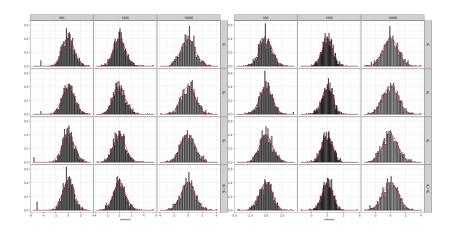


Figure 4: Comparing re-scaled empirical TMLE distributions (black) to their theoretical normal limit (red) with varying sample size (x-axis) and intervention type (y-axis). TMLEs were centered at the truth and then re-scaled by true SD. Results shown for the preferential attachment network (left) and the small world network (right).

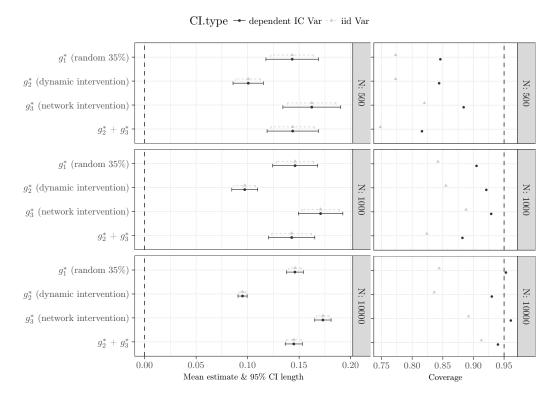


Figure 5: Mean 95% CI length (left panel) and coverage (right panel) for the TMLE in preferential attachment network with latent variable dependence, by sample size, intervention and CI type. Results are shown for the estimates of the average expected outcome under four hypothetical interventions  $(g_1^*, g_2^*, g_3^*)$  and  $g_2^* + g_3^*$ .

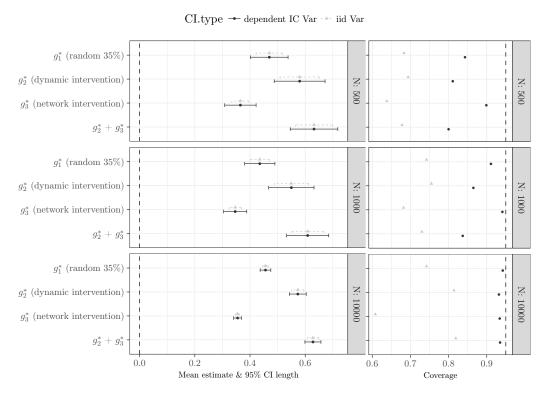


Figure 6: Mean 95% CI length (left panel) and coverage (right panel) for the TMLE in small world network with latent variable dependence, by sample size, intervention and CI type. Results are shown for the estimates of the average expected outcome under four hypothetical interventions  $(g_1^*, g_2^*, g_3^*)$  and  $g_2^* + g_3^*$ .

# Acknowledgements

Elizabeth Ogburn was supported by ONR grant N000141512343. Oleg Sofrygin and Mark van der Laan were supported by NIH grant R01 AI074345-07.

# Appendix: Proof of Theorem 1

# Regularity conditions

For a real-valued function  $\mathbf{c} \mapsto f(\mathbf{c})$ , let the  $L^2(P)$ -norm of  $f(\mathbf{c})$  be denoted by  $||f|| = E[f(\mathbf{C})^2]^{1/2}$ . Define  $\mathcal{M}_m$  and  $\mathcal{M}_{\tilde{h}}$  as the classes of possible functions that can be used for estimating the two nuisance parameters m and  $\tilde{h} \equiv \bar{h}_{x^*}/\bar{h}$ , respectively. Note that a model for g plus the empirical distribution of covariates  $\mathbf{C}$  determines  $\tilde{h}$ . Equivalent assumptions could be stated in terms of g instead of  $\tilde{h}$ , but we focus on  $\tilde{h}$  because that is the functional of g and  $\mathbf{C}$  that we model in our estimating procedure. Assume that the TMLE update  $\hat{m}_{\tilde{e}} \in \mathcal{M}_m$  with probability 1 and assume that  $\hat{h}_{x^*}/\hat{h} \in \mathcal{M}_{\tilde{h}}$  with probability 1. Finally, define the following dissimilarity measure on the cartesian product of  $\mathcal{F} \equiv \mathcal{M}_m \times \mathcal{M}_{\tilde{h}}$ :

$$d\left(\left(h,m\right),\left(\tilde{h},\tilde{m}\right)\right) = \max\left(\sup_{v \in \mathcal{V}} \mid h - \tilde{h} \mid (v), \sup_{v \in \mathcal{V}} \mid m - \tilde{m} \mid (v)\right).$$

The following are the regularity conditions required for Theorem 1, i.e. for asymptotic normality of the TMLE  $\hat{\psi}^*$ .

Uniform consistency: Assume that

$$d\left(\left(\hat{\bar{h}}_{x^*}/\hat{\bar{h}},\hat{m}_{\hat{\epsilon}}\right),\left(\bar{h}_{x^*}/\bar{h},m\right)\right)\to 0$$

in probability as  $n \to \infty$ . Note that this assumption is only needed for proving the asymptotic equicontinuity of our process; it is not needed for proofs of relevant convergence rates for the second order terms.

**Bounded entropy integral:** Assume that there exists some  $\eta > 0$ , so that  $\int_0^{\eta} \sqrt{\log(N(\epsilon, \mathcal{F}, d))} d\epsilon < \infty$ , where  $N(\epsilon, \mathcal{F}, d)$  is the number of balls of size  $\epsilon$  w.r.t. metric d needed to cover  $\mathcal{F}$ .

Universal bound: Assume  $\sup_{f \in \mathcal{F}, \mathbf{O}} | f | (\mathbf{O}) < \infty$ , where the supremum of  $\mathbf{O}$  is over a set that contains  $\mathbf{O}$  with probability one. This assumption will typically be a consequence of the choosing a specific function class  $\mathcal{F}$  that satisfies the above entropy condition.

Positivity: Assume

$$\sup_{v \in \mathcal{V}} \frac{\bar{h}_{x^*}(v)}{\bar{h}(v)} < \infty.$$

Consistency and rates for estimators of nuisance parameters: Assume that  $\|\hat{n} - m\| \|\hat{\bar{h}} - \bar{h}\| = o_P\left(\left(C_n\right)^{-1/2}\right)$ . Note that this rate is achievable if, for example, estimation of  $\bar{h}$  relies on some

pre-specified parametric model, or if both  $\bar{h}$  and m are estimated at rate  $C_n^{-1/4}$ .

Rate of the second order term: Assume that

$$R_{n1} \equiv -\int_{v} \left\{ \left( \frac{\hat{\bar{h}}_{x^*}}{\hat{\bar{h}}} - \frac{\bar{h}_{x^*}}{\bar{h}} \right) (\hat{m}_{\hat{\epsilon}} - m)(v) \bar{h}(v) d\mu(v) \right\} = o_P \left( 1/\sqrt{C_n} \right).$$

Note that this condition is provided here purely for the sake of completeness, since it will satisfied based on the previously assumed rates of convergence for  $\|\hat{m} - m\| \|\hat{\bar{h}} - \bar{h}\|$ . This follows from the fact that the parametric TMLE update step  $\hat{m}_{\hat{\epsilon}}$  of  $\hat{m}$  will have a negligible effect on the rate of convergence of the initial estimator  $\hat{m}$ , that is,  $\hat{m}_{\hat{\epsilon}}$  will converge at "nearly" the same rate as  $\hat{m}$ .

Limited connectivity and limited dependence of Y,X and C: Let  $K_{max,n} = max_i\{K_i\}$  for a fixed network with n nodes. Assume that  $K_{max,n}^2/n$  converges to 0 in probability as  $n \to \infty$ .

Bounded fourth moments:  $E[f_i(\mathbf{O})^4] < \infty$ , where  $f_i(\mathbf{O})$  is the contribution of the *i*th observation to the estimator and is defined below.

A key condition is consistency and rates for estimators of nuisance parameters. This condition will be satisfied, for example, if both nuisance models are parametric and if one of the two is correctly specified, or if both models converge to the truth at rate  $C_n^{1/4}$ . It can in fact be weakened, but for a more general discussion and the corresponding technical conditions we refer to the Appendix of van der Laan (2014). With the exception of the rates of convergence, the more general conditions for asymptotic normality of the TMLE presented in that paper apply to our setting as well.

# 7 Overview of the proof of Theorem 1

We want to show that  $\sqrt{C_n}(\hat{\psi} - \psi)$  converges in law to a Normal limit as n goes to infinity for some rate  $\sqrt{C_n}$  such that  $\sqrt{n/(K_{max}(n))^2} \leq \sqrt{C_n} \leq \sqrt{n}$ , where the rate  $\sqrt{C_n}$  is the order of the variance of the sum of the first-order linear approximation of  $(\hat{\psi} - \psi)$ .

Broadly, the proof has two parts: First, we require that the second order terms in the expansion of  $\hat{\psi} - \psi$  are stochastically less than  $1/\sqrt{C_n}$ , that is that

$$\hat{\psi_n} - \psi = \frac{1}{n} \sum_{i=1}^n \left\{ f_i(\mathbf{O}) - E[f_i(\mathbf{O})] \right\} + o_p \left( 1/\sqrt{C_n} \right),$$

where  $f_i(\mathbf{O})$  is the contribution of the *i*th observation to the estimator. Specifically, for our influence function

$$D(\mathbf{o}) = \sum_{i=1}^{n} \frac{1}{n} \sum_{i=1}^{n} E[m(V_{i}^{*}) \mid C_{j} = c_{j}] - \psi + \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{q}_{x^{*}}(v_{i})}{\bar{q}(v_{i})} \{y_{i} - m(v_{i})\},$$

the contribution of the *i*th observation is

$$f_i(\mathbf{o}) = \sum_{i=j}^n E[m(V_i^*) \mid C_j = c_j] + \frac{\bar{q}_{x^*}(v_i)}{\bar{q}(v_i)} \{y_i - m(v_i)\}.$$

Then proving asymptotic normality of the TMLE amounts to the asymptotic analysis of the sum  $\sum_{i=1}^{n} \{f_i(\mathbf{O}) - E[f_i(\mathbf{O})]\}\$ , and the second part of the proof establishes that the first order terms converge to a normal distribution when scaled by  $\sqrt{C_n}$ , that is that  $\sqrt{C_n} \sum_{i=1}^n \{f_i(\mathbf{O}) - E[f_i(\mathbf{O})]\} \to_d N(0, \sigma^2)$  for some finite  $\sigma^2$ .

The proof that the second order terms are stochastically less than  $1/\sqrt{C_n}$  is an extension of the empirical process theory of Van Der Vaart and Wellner (1996) and follows the same format as the proof in van der Laan (2014). Indeed, the proof offered by van der Laan (2014) holds immediately after replacing the rate or scaling factor  $\sqrt{n}$  with  $\sqrt{C_n}$  throughout. Only one step in the van der Laan (2014) proof relies on the network topology, which is the major difference between the setting in that paper, where the number of network connections is fixed and bounded as n goes to infinity, and the present setting: the proof requires bounding the Orlicz norms of several empirical processes corresponding to components of the influence function for  $\psi$ , and a key step is bounding the expectation of  $E[|X_n(f)|^p]$ , where  $X_n(f)$  is the stochastic process that describes the difference between the empirical (indexed by n) and the true distribution functions of a component of the influence function for  $\psi$ . This step relies on a combinatorial argument about nature of overlapping friend groups in the underlying network, and the argument for the case of growing  $K_i$  is subsumed by the argument for fixed K in van der Laan (2014).

The proof that the first order terms converge to a normal distribution requires a central limit theorem for dependent data with growing and possibly irregularly sized dependency neighborhoods, where a dependency neighborhood for unit i is a collection of observations on which the observations for unit i may be dependent. We prove such a CLT in Lemmas 1 and 2. In next section we use the CLT for growing and irregular dependency neighborhoods, along with an orthogonal decomposition of the first order terms, to prove the remainder of Theorem 1.

# Central limit theorem for first order terms

Proving asymptotic normality of the TMLE amounts to the asymptotic analysis of the sum  $\sum_{i=1}^{n} \{f_i(\mathbf{O}) - E[f_i(\mathbf{O})]\}$ . As a start, decompose  $\sum_{i=1}^{n} \{f_i(\mathbf{O}) - E[f_i(\mathbf{O})]\}$  into a sum of three orthogonal components:

$$f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C}) = f_i(\mathbf{O}) - E[f_i(\mathbf{O}) \mid \mathbf{X}, \mathbf{C}],$$
  
 $f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}) = E[f_i(\mathbf{O}) \mid \mathbf{X}, \mathbf{C}] - E[f_i(\mathbf{O}) \mid \mathbf{C}], \text{ and}$   
 $f_{\mathbf{C},i}(\mathbf{C}) = E[f_i(\mathbf{O}) \mid \mathbf{C}] - E[f_i(\mathbf{O})].$ 

Note that

$$f_i(\mathbf{O}) - E[f_i(\mathbf{O})] = f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C}) + f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}) + f_{\mathbf{C},i}(\mathbf{C})$$

and with slight abuse of notation we will also write  $f_{\mathbf{Y},i}(\mathbf{O})$ ,  $f_{\mathbf{X},i}(\mathbf{O})$  and  $f_{\mathbf{C},i}(\mathbf{O})$ . Let  $f_{\mathbf{Y}}(\mathbf{O}) = \sum_{i=1}^n f_{\mathbf{Y},i}(\mathbf{O})$ ,  $f_{\mathbf{X}}(\mathbf{O}) = \sum_{i=1}^n f_{\mathbf{X},i}(\mathbf{O})$  and  $f_{\mathbf{C}}(\mathbf{O}) = \sum_{i=1}^n f_{\mathbf{C},i}(\mathbf{O})$ . For  $i = 1, \ldots, n$ , let

$$Z_{Y,i} = \frac{f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C})}{\sqrt{Var(\sum_{i=1}^{n} f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C}))}}$$

$$Z_{X,i} = \frac{f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C})}{\sqrt{Var(\sum_{i=1}^{n} f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}))}}$$

$$Z_{C,i} = \frac{f_{\mathbf{C},i}(\mathbf{C})}{\sqrt{Var(\sum_{i=1}^{n} f_{\mathbf{C},i}(\mathbf{C}))}}.$$

and

$$Z'_{Y,i} = \frac{f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C}) | (\mathbf{X}, \mathbf{C})}{\sqrt{Var(\sum_{i=1}^{n} f_{\mathbf{Y},i}(\mathbf{Y}, \mathbf{X}, \mathbf{C}) | (\mathbf{X}, \mathbf{C}))}}}$$
$$Z'_{X,i} = \frac{f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}) | \mathbf{C}}{\sqrt{Var(\sum_{i=1}^{n} f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}) | \mathbf{C})}}$$

We use the prime to denote conditional random variables:  $Z'_{Y,i}$  conditions  $f_{\mathbf{Y},i}(\mathbf{O})$  on  $(\mathbf{X}, \mathbf{C})$  and rescales it by the standard error of  $f_{\mathbf{Y}}(\mathbf{O})|(\mathbf{X}, \mathbf{C})$ . Similarly,  $Z'_{X,i}$  conditions  $f_{\mathbf{X},i}(\mathbf{O})$  on  $\mathbf{C}$  and rescales it by the standard error of  $f_{\mathbf{X}}(\mathbf{O})|\mathbf{C}$ . Let

$$\sigma_{nY}^{2}(\mathbf{x}, \mathbf{c}) = Var\left(\sum_{i=1}^{n} f_{\mathbf{Y}, i}(\mathbf{Y}, \mathbf{x}, \mathbf{c}) | (\mathbf{X} = \mathbf{x}, \mathbf{C} = \mathbf{c})\right)$$
$$\sigma_{nY}^{2} = E_{P_{\mathbf{X}, \mathbf{C}}}\left[\sigma_{nY}^{2}(\mathbf{X}, \mathbf{C})\right],$$

$$\sigma_{nX}^{2}(\mathbf{c}) = Var\left(\sum_{i=1}^{n} f_{\mathbf{X},i}(\mathbf{X}, \mathbf{c}) | \mathbf{C} = \mathbf{c}\right)$$
$$\sigma_{nX}^{2} = E_{P_{\mathbf{C}}}\left[\sigma_{nX}^{2}(\mathbf{C})\right],$$

and

$$\sigma_{nC}^2 = Var\left(\sum_{i=1}^n f_{\mathbf{C},i}(\mathbf{C})\right).$$

Note that by the law of total variance  $\sigma_{nX}^2 = Var(\sum_{i=1}^n f_{\mathbf{X},i}(\mathbf{X},\mathbf{C}))$  and  $\sigma_{nY}^2 = Var(\sum_{i=1}^n f_{\mathbf{Y},i}(\mathbf{Y},\mathbf{X},\mathbf{C}))$ . Let  $Z'_{nY}$  denote  $\sum_{i=1}^n Z'_{Y,i}$ ,  $Z'_{nX}$  denote  $\sum_{i=1}^n Z'_{X,i}$ ,  $Z_{nY}$  denote  $\sum_{i=1}^n Z_{Y,i}$ ,  $Z_{nX}$  denote  $\sum_{i=1}^n Z_{X,i}$ , and  $Z_{nC}$  denote  $\sum_{i=1}^n Z_{C,i}$ . We will establish convergence in distribution of each of the three terms separately. Because  $Z'_{nY}$  and  $Z'_{nX}$  converge to distributions that do not depend on their conditioning events, conditional convergence in distribution implies convergence of  $Z_{nY}$  and  $Z_{nX}$  to the same limiting distributions. Since  $f_Y(\mathbf{O}), f_X(\mathbf{O})$ , and  $f_C(\mathbf{O})$  are orthogonal by construction, the variance of the limiting distribution of their sum is the sum of their marginal variances. If the three processes converge at the same rate the limiting variance will be the sum of the variances of the three processes.

However, the three terms may converge at different rates, in which case the limiting distribution of  $\hat{\psi} - \psi$  will be given by the limiting distribution of the term(s) with the slowest rate of convergence.

In order to show that  $Z'_{nX}$ ,  $Z'_{nY}$ , and  $Z_{nC}$  all converge in distribution to a N(0,1) random variable, we can use three separate applications of the central limit theorem given in Lemma 1, which is based on Stein's method.

Stein's method (Stein, 1972) quantifies the error in approximating a sample average with a normal distribution. (For an introduction to Stein's method see Ross, 2011.) Stein's method has been used to prove CLTs for dependent data with dependence structure given by dependency neighborhoods (Chen and Shao, 2004): the dependency neighborhood for observation i is a set of indices  $D_i$  such that observation i is independent of observation j, for any  $j \notin D_i$ . Conditionally on C,  $f_{X,i}$  and  $f_{X,j}$  are independent for any nodes i and j such that  $A_{ij} = 0$  and there is no k with  $A_{ik} = A_{jk} = 1$ , that is for any nodes that do not share a tie or have any mutual network contacts. The same is true for  $f_{Y,i}$ and  $f_{Y,j}$  conditional on **X** and **C** and for  $f_{C,i}$  and  $f_{C,j}$ . Thus the three collections of random variables  $Z'_{X,1},...,Z'_{X,n},\ Z'_{Y,1},...,Z'_{Y,n},$  and  $Z_{C,1},...,Z_{C,n}$  each has a dependency neighborhood structure with  $D_i = i \cup \{j : A_{ij} = 1\} \cup \{k : A_{jk} = 1 \text{ for } j : A_{ij} = 1\}, \text{ that is the "friends" and "friends"}$ of node i. Define the indicators R(i,j) for any  $(i,j) \in \{1,\ldots,n\}^2$  to be an indicator of dependence between  $Z_{X,i}$  and  $Z_{X,j}$ , R(i,j)=1 iff  $j\in D_i$  or, equivalently, if  $i\in D_j$ . For any  $i\in\{1,\ldots,n\}$  the set  $\{Z'_{X,j}:(R(i,j)=1,j\in\{1,\ldots,n\})\}$  forms the dependency neighborhood of  $Z'_{X,i}$  and the collection  $\{Z'_{X,j}:(R(i,j)=0,j\in\{1,\ldots,n\})\}$  is independent of  $Z'_{X,i}$ . The same logic applies to defining the dependency neighborhoods for  $Z'_{Y,1},...,Z'_{Y,n}$  conditional on **X** and **C**, and for  $Z_{C,1},...,Z_{C,n}$  based on (unconditional) independence of each  $f_{C,i}(\mathbf{O})$  and  $f_{C,j}(\mathbf{O})$ , as determined by the network structure and the distributional assumptions made for the baseline covariates C.

Applied to  $Z'_{nX}$ , Stein's method provides the following upper bound

$$d(Z'_{nX}, Z) \leq \sum_{i=1}^{n} \sum_{j,k \in D_{i}} E |Z'_{X,i} Z'_{X,j} Z'_{X,k}| + \sqrt{\frac{2}{\pi}} \sqrt{Var \left(\sum_{i=1}^{n} \sum_{j \in D_{i}} Z'_{X,i} Z'_{X,j}\right)},$$

where  $Z \sim N(0,1)$  and  $d(\cdot,\cdot)$  is the Wasserstein distance metric (Vallender, 1974).

In order to show that  $Z'_{nX}$  converges in distribution to Z, we must show that the righthand side of the inequality converges to zero as n goes to infinity. We will first show that this convergence holds when  $K_i = |F_i| = K_{max}(n)$  for all i, that is when all nodes have the same number of ties. We will then show that removing any tie from the network preserves an upper bound on the righthand side of the inequality. This completes our proof that for any network such that  $K_i \leq K_{max}(n)$  for all i and  $\frac{K_{max}^2(n)}{n}$  converges to zero as n goes to infinity,  $Z'_{nX}$  converges in distribution to a standard normal distribution. The same argument applied to  $Z_{nC}$  proves that it has a Normal limiting distributions as well.

1. [Applying Stein's Method to the dependent sum] Consider a network of nodes given by adjacency matrix A. Let  $U_1, ..., U_n$  be bounded mean-zero random variables with finite fourth moments and with

dependency neighborhoods  $D_i = i \cup \{j : A_{ij} = 1\} \cup \{k : A_{jk} = 1 \text{ for } j : A_{ij} = 1\}$ , and let  $K_i$  be the degree of node i. If  $K_i = K_{max}(n)$  for all i and  $K_{max}(n)^2/n \to 0$ , then  $\frac{\sum U_i}{\sqrt{var(\sum U_i)}} \stackrel{D}{\to} N(0,1)$ .

Proof. [Proof of Lemma 1] Let  $U_i' = \frac{U_i}{\sqrt{var(\sum U_i)}}$ . Application of Stein's method often involves defining the so-called "Stein coupling" (W,W',G) (Fang, 2011; Fang et al., 2015). Consider the following sum of dependent variables  $W = \sum_{i=1}^n U_i'$ . Define a discrete random variable I distributed uniformly over  $\{1,\ldots,n\}$  and define another random variable  $W' = (W - \sum_{j=1}^n R(I,j)U_j')$ . Finally, define  $G = -nU_I'$  and note that (W,W',G) forms a Stein coupling \citep{fang2011thesis,fang2015rates}. We also let  $D = (W' - W) = -\sum_{j=1}^N R(I,j)U_j'$ . This Stein coupling allows us then to derive the upper bound

$$d(W,Z) \le \sum_{i=1}^{n} \sum_{j,k \in D_i} E \left| U'_i U'_j U'_k \right| + \sqrt{\frac{2}{\pi}} \sqrt{Var \left( \sum_{i=1}^{n} \sum_{j \in D_i} U'_i U'_j \right)}, \tag{9}$$

as shown in Ross (2011). We will now show that, for any network structure,

$$\sum_{i=1}^{n} \sum_{j,k \in D_{i}} E |U'_{i}U'_{j}U'_{k}| + \sqrt{\frac{2}{\pi}} \sqrt{Var \left(\sum_{i=1}^{n} \sum_{j \in D_{i}} U'_{i}U'_{j}\right)}$$

$$= O\left(\frac{\sum_{i,j,k} R(i,j)R(i,k)}{\left[\sum_{i,j} R(i,j)\right]^{3/2}}\right). \tag{10}$$

The righthand side of the above equation is equal to  $\sqrt{\frac{(K_{max}(n))^2}{n}}$  under the assumption of  $K_{max}(n)$  ties for each node  $i = \{1, \dots, n\}$ . By assumption, we also have that  $\frac{K_{max}(n)}{\sqrt{n}}$  converges to zero as n goes to infinity, and therefore if we can show equation (10) we have proved that  $\frac{\sum U_i}{\sqrt{var(\sum U_i)}} \stackrel{D}{\to} N(0, 1)$ . Consider the term

$$\sum_{i=1}^{n} \sum_{j,k \in D_i} E |U'_i U'_j U'_k| = \frac{1}{var(\sum U_i)^{3/2}} \sum_{i=1}^{n} E \left\{ \left| U_i \left( \sum_{j \in D_i} U_k \right)^2 \right| \right\}.$$

By the assumption of bounded 4th moments,  $var(\sum U_i)^{3/2} = O\left(\left[\sum_{i,j} R(i,j)\right]^{3/2}\right)$ , that is,  $var(\sum U_i)$  stabilizes to a constant when scaled by  $\sum_{i,j} R(i,j)$ . Using the fact that each  $|U_i|$  is bounded we get

$$\sum_{i=1}^{N} E \left\{ \left| U_i \left( \sum_{j \in D_i} U_j \right)^2 \right| \right\}$$

$$\leq M \sum_{i=1}^{n} \left\{ \sum_{j,k} R(i,j)R(i,k) \right\}$$

$$= M \sum_{i,j,k} R(i,j)R(i,k),$$

for some positive constant  $M < \infty$ . Combining the above expressions, we get

$$\sum_{i=1}^{n} \sum_{j,k \in D_i} E |U'_i U'_j U'_k| = O \left( \frac{\sum_{i,j,k} R(i,j) R(i,k)}{\left[\sum_{i,j} R(i,j)\right]^{3/2}} \right).$$

Now consider the second term:

$$\sqrt{Var\left(\sum_{i=1}^{n}\sum_{j\in D_{i}}U'_{i}U'_{j}\right)} = \frac{\sqrt{Var\left(\sum_{i=1}^{n}\sum_{j\in D_{i}}U_{i}U_{j}\right)}}{var(\sum U_{i})^{2}}.$$

There are  $\sum_{i,j} R(i,j)$  terms in  $\sum_{i=1}^{n} \sum_{j \in D_i} U_i U_j$ , and the number of terms  $U_k U_l$  with which  $U_i U_j$  has non-zero covariance is  $|D_i \cup D_j| \leq \sum_k R(i,k) + \sum_k R(i,k)$ , so  $Var\left(\sum_{i=1}^{n} \sum_{j \in D_i} U_i U_j\right) \leq M \sum_{i,j} R(i,j) \sum_k R(i,k)$  for some finite M. Therefore  $Var\left(\sum_{i=1}^{n} \sum_{j \in D_i} U_i U_j\right) = O\left(\sum_{i,j,k} R(i,j) R(i,k)\right)$ .  $Var(\sum U_i)^2 = O\left(\left[\sum_{i,j} R(i,j)\right]^2\right)$ , so the second term is of smaller order than the first term. Therefore we have only to consider the first term and we have completed the proof.

**2.** [Bound goes to zero when  $K_i \leq K_{max}(n)$  for all i] Convergence to zero of the righthand side of Equation (9) is preserved under the removal of ties and holds as long as  $K_i \leq K_{max}(n)$  for all i and  $\frac{K_{max}^2(n)}{n}$  converges to zero as n goes to infinity.

*Proof.* [Proof of Lemma 2] Consider a sequence of networks with n going to infinity such that the righthand side of Equation (9) converges to 0, i.e.

$$\sum_{i=1}^{n} \sum_{j,k \in D_i} E\left|U_i'U_j'U_k'\right| + \sqrt{\frac{2}{\pi}} \sqrt{Var\left(\sum_{i=1}^{n} \sum_{j \in D_i} U_i'U_j'\right)} \rightarrow 0.$$

Because the second term is of the same or smaller order than the first, we only have to consider the first term. For this sequence of networks, define  $A_n = \sum_{i=1}^n \sum_{j,k \in D_i} E |U_i'U_j'U_k'|$ . Removing a single tie from the underlying network has the effect of rendering independent some pairs that were previously dependent; We now consider the effect of rendering a single dependent pair independent but otherwise leaving the distributions of the random variables the same. Suppose the pair rendered independent is (l, m). Define a new sequence of networks with n going to infinity to be identical to the previous sequence but with pair (l, m) independent, and let  $A'_n$  be the first term in the righthand side of Equation (9) for this new sequence. Then

$$A'_{n} = A_{n} - 2 \sum_{k \in D_{l} \cup D_{m}} E |U'_{l}U'_{m}U'_{k}|$$

which is bounded above by  $A_n$ .

This completes the proof that  $Z'_{nX}$ ,  $Z'_{nY}$ , and  $Z_{nC}$  have Normal limiting distributions.

3. [Conditional CLT implies marginal CLT] $Z'_{nX}$  converges to Normal distribution after marginalizing over  $\mathbf{C}$  (but conditioning on the network as captured by the adjacency matrix  $\mathbf{A}$ ) and  $Z'_{nY}$  converges to Normal distribution after marginalizing over  $(\mathbf{X}, \mathbf{C})$ . That is,  $Z_{nX}$  and  $Z_{nY}$  both converge to Normal distributions.

*Proof.* [Proof of Lemma 3] For illustration consider  $Z'_{nX} = \sum_{i=1}^{n} Z'_{2,i}$ , where

$$Z_{X,i}' = \left(f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C}) \,| \mathbf{C}\right) / \sqrt{\sigma_{nX}^2(\mathbf{C})}$$

and note that the proof of the convergence of  $Z_{nY}$  is nearly identical. The conditional CLT results from Lemma 1 show that

$$P[Z'_{nX} \le x | \mathbf{C} = \mathbf{c}] = P\left[\left(\sum_{i=1}^{N} \frac{f_{\mathbf{X},i}(\mathbf{X}, \mathbf{c})}{\sqrt{\sigma_{nX}^{2}(\mathbf{c})}} \le x\right) | \mathbf{C} = \mathbf{c}\right]$$

converges to  $\Phi(x)$  for each x and almost every  $\mathbf{c}$ , where  $\Phi$  is the cumulative distribution function of the standard Normal random variable and  $\mathbf{C}$  is a given sequence  $(C_i: i=1,\ldots,n)$ . Let  $P_{\mathbf{C}}$  denote the distribution of  $\mathbf{C}$ . Then

$$P(Z_{nX} \le x) \equiv P\left[\left(\sum_{i=1}^{N} \frac{f_{\mathbf{X},i}(\mathbf{X}, \mathbf{C})}{\sqrt{\sigma_{nX}^{2}}} \le x\right)\right]$$
$$= \int_{\mathbf{c}} P(Z'_{nX} \le x | \mathbf{C} = \mathbf{c}) dP_{\mathbf{C}}(\mathbf{c}).$$

For a given x, the dominated convergence theorem is now applied with  $f_n(\mathbf{c}) = P(Z'_{nX} \leq x | \mathbf{C} = \mathbf{c})$  and the limit given by  $f(\mathbf{c}) = \Phi(x) = m$ , where m is some constant that doesn't depend on  $\mathbf{c}$ . From the previous conditional CLT result it follows that  $f_n(\mathbf{c})$  converges to  $f(\mathbf{c})$  pointwise for each  $\mathbf{c}$ . The next step is to find an integrable function g, such that  $f_n < g$  and  $\int g(\mathbf{c})dP_{\mathbf{C}}(\mathbf{c}) < \infty$ . The proof is then completed by choosing g = 1.

We have now shown that  $Z_{nY}$ ,  $Z_{nX}$ , and  $Z_{nC}$  are asymptotically Normally distributed. We now show that the sum of the three processes converges in distribution to a Normal random variable. Consider three cases: (1) the three processes have the same rate of marginal convergence in distribution, (2) one of the three processes converges faster than the other two, and (3) two of the processes converge faster than the third. In all three cases the rate of convergence for the sum will be the slowest of the three marginal rates. In case (3), the limiting distribution of the sum is determined entirely by the one process that converges with a slower rate than the other two: the other two processes will converge to constants (specifically to their expected values of 0) when standardized by the slower rate; Slutsky's theorem concludes the proof. We focus on case (1) below; case (2) follows immediately by applying the proof below to the two processes that converge at the same slower rate and applying Slutsky's to the third, faster converging process.

For convenience, in order to show that the sum of the three dependent processes also converges to Normal, define

$$C_n^* := \sigma_{nY}^2 + \sigma_{nX}^2 + \sigma_{nC}^2.$$

Note that  $C_n^*$  is related to  $C_n$  as follows:  $C_n = O(n^2/C_n^*)$ .

**4.** [CLT for the sum of the three orthogonal processes] If all three processes have the same marginal rate of convergence, then

$$\frac{1}{\sqrt{C_n^*}} \left( f_{\mathbf{Y}}(\mathbf{Y}, \mathbf{X}, \mathbf{C}) + f_{\mathbf{X}}(\mathbf{X}, \mathbf{C}) + f_{\mathbf{C}}(\mathbf{C}) \right) \to N(0, 1).$$

*Proof.* [Proof of Lemma 4] Without the loss of generality, we prove that  $Z_{nX} + Z_{nC} \to N(0,2)$  and note that the general result for  $(Z_{nY} + Z_{nX} + Z_{nC})$  follows by applying a similar set of arguments.

Consider the following random vector  $(Z_{nX}, Z_{nC})$  taking values in  $\mathbb{R}^2$ . Let  $F_n(x_1, x_2) \equiv P(Z_{nX} \leq x_1, Z_{nC} \leq x_2)$ , where  $(x_1, x_2) \in \mathbb{R}^2$ . Let  $\Phi^2(x_1, x_2) \equiv P(Z_X \leq x_1)P(Z_C \leq x_2)$ , for  $Z_X \sim N(0, 1)$  and  $Z_C \sim N(0, 1)$ , that is,  $\Phi^2(x_1, x_2)$  defines the CDF of the bivariate standard normal distribution, for  $(x_1, x_2) \in \mathbb{R}^2$ . The goal is to show that  $F_n(x_1, x_2) \to \Phi^2(x_1, x_2)$ , for any  $(x_1, x_2) \in \mathbb{R}^2$ . The convergence in distribution for  $Z_{nX} + Z_{nC}$  will follow by applying the Cramer and Wold Theorem (1936).

Note that

$$P(Z_{nX} \le x_1, Z_{nC} \le x_2)$$
  
= $P(Z_{nX} \le x_1 | Z_{nC} \le x_2) P(Z_{nC} \le x_2).$ 

First, from the previous application of Stein's method, we have that

$$P(Z_{nC} \le x_2) \to \Phi(x_2),$$

where  $\Phi(x_2) \equiv P(Z_C \leq x_2)$ ,  $Z_C \sim N(0,1)$  and  $x_2 \in \mathbb{R}^2$ . Also note that

$$P(Z_{nX} \le x_1 | Z_{nC} \le x_2)$$

$$= \sum_{\mathbf{c} \in \mathcal{C}} P(Z_{nX} \le x_1 | \mathbf{C} = \mathbf{c}) P(\mathbf{C} = \mathbf{c} | Z_{nC} \le x_2),$$

where C denotes the support of  $\mathbf{C}$ ,  $Z_{nX} = \frac{1}{\sqrt{C_n^*}} f_{\mathbf{X}}(\mathbf{X}, \mathbf{C})$ ,  $Z_{nC} = \frac{1}{\sqrt{C_n^*}} f_{\mathbf{C}}(\mathbf{C})$  and

$$P(\mathbf{C} = \mathbf{c} | Z_{nC} \le x_2) = \frac{P(\mathbf{C} = \mathbf{c})I((1/\sqrt{C_n^*}) f_{\mathbf{C}}(\mathbf{c}) \le x_2)}{P((1/\sqrt{C_n^*}) f_{\mathbf{C}}(\mathbf{c}) \le x_2)}.$$

By another application of Stein's method, it was shown that

$$P(Z_{nX} \leq x_1 | \mathbf{C} = \mathbf{c}) \rightarrow \Phi(x_2),$$

for any realization of  $\mathbf{c} \in \mathcal{C}$ . That is, we've shown that the limiting distribution of  $Z_{nX}$  conditional on  $\mathbf{C} = \mathbf{c}$ , does not itself depend on the conditioning event  $\mathbf{C} = \mathbf{c}$ . Applying Lemma 3, we finally conclude that  $F_n(x_1, x_2) \to \Phi^2(x_1, x_2)$ , for any  $(x_1, x_2) \in \mathbb{R}^2$  and the result follows.

# Variance estimation

The estimate of the variance of the TMLE  $\hat{\psi}$  can be obtained from the sum, scaled by  $1/n^2$ , of the three plug-in estimators of

$$\begin{split} \sigma_{nY}^2 &=& \sum_{i,j} E(f_{\mathbf{Y},i}(\mathbf{O}) f_{\mathbf{Y},j}(\mathbf{O})) \\ \sigma_{nX}^2 &=& \sum_{i,j} E(f_{\mathbf{X},i}(\mathbf{O}) f_{\mathbf{X},j}(\mathbf{O})) \\ \sigma_{nC}^2 &=& \sum_{i,j} E(f_{\mathbf{C},i}(\mathbf{O}) f_{\mathbf{C},j}(\mathbf{O})). \end{split}$$

Alternatively, one can estimate the variance from a single plug-in estimator

$$\frac{1}{n^2} \sum_{i,j} E(f_i(\mathbf{O}) f_j(\mathbf{O})).$$

Note that contribution to these variances of any pair i,j not in each others dependency neighborhoods will be 0. Therefore, it is acceptable to sum only over pairs i,j sharing a tie or a mutual contact in the underlying network. Finally, note that we do not need to know the true rate of convergence  $\sqrt{C_n}$  to obtain a valid estimate of the C.I. for  $\psi$ ; this rate is captured by the number of non-zero terms in the variance sums.

# References

Ali, M. M. and Dwyer, D. S. "Social network effects in alcohol consumption among adolescents." *Addictive behaviors*, 35(4):337–342 (2010).

Aronow, P. M. and Samii, C. "Estimating average causal effects under general interference." Technical report, Yale University (2013).

Athey, S., Eckles, D., and Imbens, G. W. "Exact P-values for Network Interference\*." *Journal of the American Statistical Association*, (just-accepted) (2016).

Barabási, A.-L. and Albert, R. "Emergence of scaling in random networks." *science*, 286(5439):509–512 (1999).

Besag, J. "On spatial-temporal models and Markov fields." In Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, 47–55. Springer (1974).

Bowers, J., M, F. M., and C, P. "Reasoning about interference between units: A general framework." *Political Analysis*, 21:97–124 (2013).

Cacioppo, J. T., Fowler, J. H., and Christakis, N. A. "Alone in the crowd: the structure and spread of loneliness in a large social network." *Journal of personality and social psychology*, 97(6):977 (2009).

- Chen, L. H. and Shao, Q.-M. "Normal approximation under local dependence." *The Annals of Probability*, 32(3):1985–2028 (2004).
- Christakis, N. and Fowler, J. "The spread of obesity in a large social network over 32 years." New England Journal of Medicine, 357(4):370–379 (2007).
- —. "The collective dynamics of smoking in a large social network." New England journal of medicine, 358(21):2249–2258 (2008).
- —. "Social network sensors for early detection of contagious outbreaks." *PloS one*, 5(9):e12948 (2010).
- Clauset, A., Shalizi, C. R., and Newman, M. E. "Power-law distributions in empirical data." *SIAM review*, 51(4):661–703 (2009).
- Cohen-Cole, E. and Fletcher, J. "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic." *Journal of Health Economics*, 27(5):1382–1387 (2008).
- Diaconis, P. and Janson, S. "Graph limits and exchangeable random graphs." arXiv preprint arXiv:0712.2749 (2007).
- Eckles, D., Karrer, B., and Ugander, J. "Design and analysis of experiments in networks: Reducing bias from interference." arXiv preprint arXiv:1404.7530 (2014).
- Fang, X. "Multivariate, combinatorial and discretized normal approximations by Stein's method." Ph.D. thesis (2011).
- Fang, X., Röllin, A., et al. "Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics." *Bernoulli*, 21(4):2157–2189 (2015).
- Goetzke, F. "Network effects in public transit use: evidence from a spatially autoregressive mode choice model for New York." *Urban Studies*, 45(2):407–417 (2008).
- Graham, B., Imbens, G., and Ridder, G. "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach." Technical report, National Bureau of Economic Research (2010).
- Halloran, M. and Hudgens, M. "Causal Inference for Vaccine Effects on Infectiousness." The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series, 20 (2011).
- Halloran, M. and Struchiner, C. "Causal inference in infectious diseases." *Epidemiology*, 142–151 (1995).
- Haneuse, S. and Rotnitzky, A. "Estimation of the effect of interventions that modify the received treatment." *Statistics in medicine*, 32(30):5260–5277 (2013).
- Harling, G., Wang, R., Onnela, J.-P., and DeGruttola, V. "Leveraging Contact Network Structure in the Design of Cluster Randomized Trials." Harvard University Biostatistics Working Paper Series, (Working Paper 199) (2016).

- Hong, G. and Raudenbush, S. "Evaluating Kindergarten Retention Policy." *Journal of the American Statistical Association*, 101(475):901–910 (2006).
- —. "Causal inference for time-varying instructional treatments." *Journal of Educational and Behavioral Statistics*, 33(3):333–362 (2008).
- Hudgens, M. and Halloran, M. "Toward causal inference with interference." *Journal of the American Statistical Association*, 103(482):832–842 (2008).
- Knoke, D. and Yang, S. Social network analysis, volume 154. Sage (2008).
- Lauritzen, S. L. and Richardson, T. S. "Chain graph models and their causal interpretations." *Journal of the Royal Statistical Society: Series B*, 64(3):321–348 (2002).
- Lee, L.-F. "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models." *Econometrica*, 72(6):1899–1925 (2004).
- Lin, X. "Peer effects and student academic achievement: an application of spatial autoregressive model with group unobservables." *Unpublished manuscript*, *Ohio State University* (2005).
- Lovász, L. Large networks and graph limits, volume 60. American Mathematical Soc. (2012).
- Lyons, R. "The spread of evidence-poor medicine via flawed social-network analysis." *Statistics*, *Politics*, and *Policy*, 2(1) (2011).
- Madan, A., Moturu, S. T., Lazer, D., and Pentland, A. S. "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks." In *Wireless Health 2010*, 104–110. ACM (2010).
- Moreno, J. L. "Sociometry in relation to other social sciences." Sociometry, 1(1/2):206–219 (1937).
- Muñoz, I. D. and van der Laan, M. "Population intervention causal effects based on stochastic interventions." *Biometrics*, 68(2):541–549 (2012).
- Newman, M. Networks: an introduction. Oxford: Oxford University Press (2009).
- Newman, M. E. and Park, J. "Why social networks are different from other types of networks." *Physical Review E*, 68(3):036122 (2003).
- Ogburn, E. and VanderWeele, T. J. "Causal diagrams for interference." Technical report, Harvard University (2013).
- Ogburn, E. L. and VanderWeele, T. J. "Vaccines, Contagion, and Social Networks." arXiv preprint arXiv:1403.1241 (2014).
- O'Malley, J. A. and Marsden, P. V. "The analysis of social networks." *Health services and outcomes research methodology*, 8(4):222–269 (2008).
- Pearl, J. "Causal diagrams for empirical research." Biometrika, 82(4):669–688 (1995).
- —. Causality: models, reasoning and inference. Cambridge Univ Press (2000).

- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015).
- Rosenbaum, P. "Interference between units in randomized experiments." *Journal of the American Statistical Association*, 102(477):191–200 (2007).
- Rosenquist, J. N., Murabito, J., Fowler, J. H., and Christakis, N. A. "The spread of alcohol consumption behavior in a large social network." *Annals of Internal Medicine*, 152(7):426–433 (2010).
- Ross, N. F. "Fundamentals of Stein's method." Probability Surveys, 8:210-293 (2011).
- Rubin, D. "Comment: Neyman (1923) and causal inference in experiments and observational studies." Statistical Science, 5(4):472–480 (1990).
- Shalizi, C. and Thomas, A. "Homophily and contagion are generically confounded in observational social network studies." *Sociological Methods & Research*, 40(2):211–239 (2011).
- Shalizi, C. R. "Comment on "Why and When 'Flawed' Social Network Analyses Still Yield Valid Tests of no Contagion"." *Statistics, Politics, and Policy*, 5(1) (2012).
- Shalizi, C. R. and Rinaldo, A. "Consistency under Sampling of Exponential Random Graph Models."

  Annals of Statistics, 41(2):508–535 (2013).
- Sobel, M. "What Do Randomized Studies of Housing Mobility Demonstrate?" *Journal of the American Statistical Association*, 101(476):1398–1407 (2006).
- Sofrygin, O. and van der Laan, M. J. "Semi-Parametric Estimation and Inference for the Mean Outcome of the Single Time-Point Intervention in a Causally Connected Population." *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 344) (2015).
- Sofrygin, O. and van der Laan, M. J. tmlenet: Targeted Maximum Likelihood Estimation for Network Data (2015). R package version 0.1.0.
- Sofrygin, O., van der Laan, M. J., and Neugebauer, R. simcausal: Simulating Longitudinal Data with Causal Inference Applications (2015). R package version 0.5.0.

  URL http://CRAN.R-project.org/package=simcausal
- Stein, C. "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables." In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, 583–602 (1972).
- Tchetgen Tchetgen, E. J. and VanderWeele, T. "On causal inference in the presence of interference." Statistical Methods in Medical Research, 21(1):55–75 (2012).
- Thomas, A. C. "The social contagion hypothesis: Comment on 'Social contagion theory: Examining dynamic social networks and human behavior'." *Statistical in Medicine*, 32(4):581–590 (2013).
- Tsiatis, A. Semiparametric theory and missing data. Springer Science & Business Media (2007).
- Vallender, S. "Calculation of the Wasserstein distance between probability distributions on the line." Theory of Probability & Eamp; Its Applications, 18(4):784–786 (1974).

- van der Laan, M. J. "Causal Inference for a Population of Causally Connected Units." *Journal of Causal Inference*, 0(0):2193–3677 (2014).
- Van der Laan, M. J. and Robins, J. M. Unified methods for censored longitudinal data and causality. Springer Science & Springer & S
- Van der Laan, M. J. and Rose, S. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media (2011).
- van der Laan Mark, J., Polley Eric, C., et al. "Super learner." Statistical Applications in Genetics and Molecular Biology, 6(1):1–23 (2007).
- Van Der Vaart, A. W. and Wellner, J. A. "Weak Convergence." In Weak Convergence and Empirical Processes, 16–28. Springer (1996).
- VanderWeele, T. "Direct and indirect effects for neighborhood-based clustered and longitudinal data." Sociological Methods & Research, 38(4):515–544 (2010).
- VanderWeele, T. and Tchetgen Tchetgen, E. "Bounding the Infectiousness Effect in Vaccine Trials." Epidemiology, 22(5):686 (2011a).
- —. "Effect partitioning under interference in two-stage randomized vaccine trials." Statistics & probability letters, 81(7):861–869 (2011b).
- VanderWeele, T. J., Ogburn, E. L., and Tchetgen, E. J. T. "Why and when 'flawed' social network analyses still yield valid tests of no contagion." *Statistics, Politics, and Policy*, 3(1) (2012).
- Wasserman, S. "Comment on "Social contagion theory: Examining dynamic social networks and human behavior" by Nicholas Christakis and James Fowler." *Statistics in Medicine*, 32(4):578–580 (2013).
- Watts, D. J. and Strogatz, S. H. "Collective dynamics of 'small-world' networks." *Nature*, 393(6684):440–442 (1998).
- Young, J. G., Hernán, M. A., and Robins, J. M. "Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data." Epidemiologic Methods, 3(1):1–19 (2014).