Estimating treatment effects with partial network data

April 27, 2023

1 Methods

1.1 Data Generating Model

We suppose that there are n individuals denoted by a set $N = \{1, 2, \ldots, n\}$. Let $Y_i \in \mathbb{R}$ denote an outcome of interest, let $A_i \in \{0, 1\}$ denote an individual's binary treatment and $V_i \in \mathcal{V}$ denote their treatment **exposure** as in Aronow and Samii [2017]. The exposure generalizes the treatment from the binary case, to a higher dimensional treatment, based on other's treatment in the population. We take the standard assumption that interference is propagated over some network $\mathcal{G} = (N, E)$ where N is the set of individuals in the population and $E \in N \times N$ are the set of edges. The graph $\mathcal{G}$ can also be represented by the adjacency matrix G_n. In general G_n can be directed, but for now we will simply assume it is undirected.

We suppose that there exists some function f_V such that $V_i = f_V(\boldsymbol{a}, \theta_i)$ where θ_i is the local graph information for individual i. This is commonly the i^{th} row of neighbours, but it need not be restricted to immediate neighbors. Finally, we assume that the graph G_n is generated by an exchangeable random graph with parameters Z_i for each individual. In practice, this could be the stochastic block-model where Z_i is discrete, but nonparametric identification need not require this. Additionally, we suppose that there may be additional covariates X_i which are of interest for heterogeneous effects, which may be related to the outcome, but do not necessarily contribute to the graph. Let $Y_i(\boldsymbol{a})$ be the potential outcome for some outcome of interest under a treatment vector $\boldsymbol{a} \in \{0, 1\}^n$. We will consider parameters of interest which are functions of the potential outcomes in the graph.

1.2 Parameters of Interest

We are primarily considering parameters of the form $\Psi(\boldsymbol{a}, x, z, G_n) = \frac{1}{n}\sum_{i=1}^{n} P(Y_i(\boldsymbol{a})|X_i = x, Z_i = z, G_n)$. These are the conditional average treatment effects of assigning a treatment vector $\boldsymbol{a}$, possibly conditional on x and z and conditional on the graph that was observed. These include the particular cases, marginalizing over either Z_i or X_i. These parameters be $\Psi(\boldsymbol{a}, x) = \frac{1}{n}\sum_{i=1}^{n} P(Y_i(\boldsymbol{a})|X_i = x, Z_i = Z_i, G_n)$ or $\Psi(\boldsymbol{a}, x) = \frac{1}{n}\sum_{i=1}^{n} P(Y_i(\boldsymbol{a})|Z_i = Z_i, G_n)$.

1.3 Nonparametric Identification

In order to estimate the corresponding parameters of interest, we need to first introduce some assumptions.

Definition 1.1 (Exposure Weak Ignorability)**.** *We say that an exposure assignment is* ***weakly ignorable*** *if the following holds:*

$$Y_i(v) \perp\!\!\!\perp \{V_i = v\} | Z_i$$

This simply states that once we condition on the parameters which generate the graph, then the potential outcome for the observes exposure v is independent of the treatment.

Definition 1.2 (Exposure Consistency)**.** *Exposure Consistency holds if*

$$V_i = v \implies Y_i = Y_i(v)$$

where $Y_i(v)$ is the potential outcome of individual i for the exposure v.

The last assumption we need is independence of the graph from the outcome, conditional on the generative model parameters and the exposure.

Definition 1.3 (Conditional Independence of the Graph and Outcome)**.** *We assume that the outcome is conditionally independent of the outcome conditional on the exposure and the graph generative parameters*

$$Y_i \perp\!\!\!\perp G_{ij} | V_i, Z_i \text{ for all } i, j$$

Therefore we can derive the causal effects of interest via the following.

$$\begin{aligned}
P(Y_i(\boldsymbol{a}) | Z_i = z, G_n) &= P(Y_i(v) | Z_i = z, G_n) \text{ By the exposure mapping assumption} && (1)\\
&= P(Y_i(v) | V_i = v, Z_i = z, G_n) \text{ By Exposure Weak Ignorability} && (2)\\
&= P(Y_i | V_i = v, Z_i = z, G_n) \text{ By Consistency} && (3)\\
&= P(Y_i | V_i = v, Z_i = z) \text{ Graph Conditional Independence} && (4)
\end{aligned}$$

Therefore, causal effects of interest can be estimated as long as we model properly $P(Y_i | V_i = v, Z_i = z)$. In particular, we may be concerned with the mean regression $\mathbb{E}[Y_i | V_i = v, Z_i = z] = m(v, z)$

Lastly, if we want to include auxillary covariates X_i, we suppose that each of these properties above are not affected when also conditioning on X_i.

As an alternative view of the problem, and relating it to other work in the interference literature. We can consider the nonparametric structural equation model such as those in Ogburn et al. [2017], Cristali and Veitch [2022]. Assuming for now that covariates are not included.

$$\begin{aligned}
Z_i &= f_Z[\epsilon_{Z_i}]\\
G_{ij} &= f_G[Z_i, Z_j, \epsilon_{ij}]\\
A_i &= f_A[Z_i, \epsilon_{A_i}]\\
V_i &= f_V(\boldsymbol{A}, \theta_i(G_n))\\
Y_i &= f_Y[V_i, Z_i, X_i, \epsilon_{Y_i}]
\end{aligned}$$

where f_S are the nonparamteric structural equations and ϵ_S are exogenous noise variables. The graph in this case is just generated by a simple graphon model based on node level parameters Z_i.

This assumption relies on assigning treatment, conditional on the network generating parameters Z_i. We can equivalently represent this This structural equation model, can be interpreted using the causal DAG (directed acyclic graph) in Figure 7.

1.4 Types of Experiments and Their Available Data

We consider saturation randomization design experiments. The saturation randomized design partitions the dataset into K clusters of size n_k respectively, then assigns a random τ_k fraction of each of the clusters to the given treatment, for a total number of treated of $n_{a=1} = \sum_{k=1}^{K} \tau_k n_k$. In practice, this design is particularly useful when there are a limited number of treatments one can assign. If τ_k are random as proposed in Hudgens and Halloran [2008], this design generalizes the completely randomized design and cluster randomized design that are common in practice, and is known as *randomized saturation randomized design*. Alternative versions where the saturation level τ_k is fixed at each cluster are known as *deterministic saturation randomized design*, and have been explored in Cai et al. [2022]. Under strict budgeting constraints on the number of treatments available, the deterministic version is a more natural object of study. In all of our settings, we consider the clusters, to be the memberships of the stochastic block-model which generated the graphs.

Next we consider several data collection experiments.

1.4.1 Known Clusters

This first design is for when the clusters in the network are known a-priori. This could be useful in a setting in which there is an identifiable group for which the graph clusters naturally such as religion. Let Z_i denote the identifiable trait. In this data type, we assume we have a vector $D_i \in \mathbb{Z}_{\geq 0}^K$ consisting of a vector of responses to the ARD question of "How many people of X trait do you know?", we will also collect the trait of the individual $Z_i \in \{0,1\}^K$, a binary indicator.

1. For each $i \in \{1, 2, \ldots, n\}$
   (a) Collect ARD vector D_i and other covariates X_i
   (b) Assign treatment a_i according to the group saturation level τ_{Z_i}
2. Wait a specified length of time for the effect of the treatment to set in
3. Collect response Y_i for each $i \in \{1, 2, \ldots, n\}$

1.4.2 Unknown Clusters

If there are unknown clusters, we can first collect baseline data as well as ARD data. Since we collect baseline data. We denote this first population N_0 and this need not have the same individuals as the treatment populations, however this is important for establishing the clustering we observe in the graph, as well as collecting baseline data for stabilizing the variance of some of the estimates.

1. For each $i \in N_0$
   (a) Collect ARD vector D_i, as well as baseline Y_i and other covariates X_i
2. Define a clustering $C(D_i) \to \widehat{Z}_i$ based on the ARD data.
3. For each $i \in \{1, 2, \ldots, n\}$
   (a) Collect ARD vector D_i and other covariates X_i
   (b) Assign treatment a_i according to the group saturation level $\tau_{\widehat{Z}_i}$
4. Wait a specified length of time for the effect of the treatment to set in
5. Collect response Y_i for each $i \in \{1, 2, \ldots, n\}$

1.4.3 Staggered Rollout

(To finish later)

1.5 Choice of Exposure Mapping

As we saw in Section 1.3, identification of causal effects of interest relies on properly specifying the regression models $E[Y_i|Z_i = z, V_i = v] = m(z, v)$, or $E[Y_i|Z_i = z, V_i = v, X_i = x] = m(z, v, x)$. This will rely on a choice of the exposure mapping $V_i = f_V(\boldsymbol{a}, \theta_i(G_n))$. The choice of the exposure mapping will be a modelling decision left to the practitioner. Choosing the correct one may be based on domain knowledge.

We next discuss some options for exposure maps. Let $G_i^{(k)}$ denote the graph of connections to individuals of group k. Some simple choices for the regression function may include:

$$\text{Fraction of Treated Neighbors } V_i = (a_i, \frac{\boldsymbol{a}^T G_i}{\boldsymbol{a}^T \mathbf{1}}) =: (a_i, q_i) \tag{5}$$

$$\text{Fraction of Treated Neighbors by Group } V_i = (a_i, \frac{\boldsymbol{a}^T G_i^{(1)}}{\boldsymbol{a}^T \mathbf{1}}, \frac{\boldsymbol{a}^T G_i^{(2)}}{\boldsymbol{a}^T \mathbf{1}}, \ldots, \frac{\boldsymbol{a}^T G_i^{(K)}}{\boldsymbol{a}^T \mathbf{1}}) =: (a_i, q_i^{(1)}, \ldots, q_i^{(K)}) \tag{6}$$

For each of these, since we do not observe q_i directly, we must approximate it using the model. Let $\widetilde{V}_i = \mathbb{E}[V_i|\boldsymbol{a}, D_i, Z_i; \Pi_0]$ where Π_0 are the model parameters, and D_i is the data available

about individual i (i.e. ARD). In the typical fashion of other ARD papers, we will replace the V_i term with $\widetilde{V}_i$ as in Breza et al. [2020, 2019].

There is a subtle distinction in this setting between what is substituted in our setting vs in previous work. In Breza et al. [2019] the replaced term is $\mathbb{E}[V_i|\boldsymbol{a}, Z_i; \Pi_0]$, it is the mean of the graph statistic over all graphs generated with parameters Π_0. Alternatively, $\mathbb{E}[V_i|D_i, \boldsymbol{a}, Z_i; \Pi_0]$ is the average graph statistic, generated with parameters Π_0 conditional on the fact that D_i is the trait level vector that was generated. With a simple simulation, we will show that under randomized saturation design, neither framework will be sufficient for estimating the model parameters.

Example 1.1. *Suppose that we have a stochastic block model with adjacency matrix $P \in [0,1]^{K\times K}$, and also there are known block memberships $Z_i \in \{1, 2, \ldots, K\}$. Suppose also that there is a trait matrix $Q \in [0,1]^{T\times K}$ such that $Q_{tk} = P(T_i = t|Z_i = k)$. Suppose that $D_i \in \mathbb{R}^T_{\geq 0}$ is the vector of ARD responses given each trait. Suppose also for simplicity, that $\sum_{t=1}^T D_{it} = d_i$, the true degree. Furthermore, let $Q^* \in [0,1]^{K\times T}$ where $Q^*_{kt} = P(Z_i = k|T_i = t)$ denote the conditional distribution of an individual's cluster given their observed trait.*

Then let $V_i = (\frac{\boldsymbol{a}^T G_i}{d_i})$ denote the fraction of neighbors connected to individual i. And let τ_k denote the fraction of individuals in group k that are treated.

$$\begin{aligned}\mathbb{E}[V_i|\boldsymbol{a}, \boldsymbol{Z}; \pi_0] &= \mathbb{E}\left[\frac{\boldsymbol{a}^T G_i}{d_i}\middle|\boldsymbol{a}, \boldsymbol{Z}; \pi_0\right] \\ &= f(Z_i)\end{aligned}$$

by exchangeability, this is going to just be a function of Z_i the true membership of node i. However, in the conditional definition, then conditional on the observed data D_i

$$\begin{aligned}\mathbb{E}[V_i|X_i, \boldsymbol{a}, \boldsymbol{Z}; \pi_0] &= \mathbb{E}\left[\frac{\boldsymbol{a}^T G_i}{d_i}\middle|X_i, \boldsymbol{a}, \boldsymbol{Z}; \pi_0\right] \\ &= \frac{\sum_{k=1}^K \tau_k Q^*_{k\cdot} X_i}{d_i}\end{aligned}$$

Hence if we do not condition on X_i, then $\mathbb{E}[V_i|\boldsymbol{a}, \boldsymbol{Z}; \pi_0]$ is the same for every node with the same latent community Z_i.

In the first case, if we only have access to the outcomes after treatment, then we can never recover even simple parameteric models. This is because the exposure will always be colinear with the membership Z_i. Hence this is insufficient.

The second case however, it seems like there will be variation within a cluster. However, we will show that even under simple examples, this will be insufficient for consistent estimation.

1.6 Lack of Consistency in the linear outcome stochastic block model

In order to estimate the mean model $m(z, v)$ correctly, we would like to first consider an additional assumption. We will assume that the exposure is conditionally independent of the ARD data D_i.

$$V_i \perp\!\!\!\perp D_i | \boldsymbol{a}, \boldsymbol{Z}$$

Consider the linear model in the exposure mapping. For simplicity in notation, let $Z_i' \in \{0,1\}^K$ be the equivalent one-hot encoding of Z_i

$$\mathbb{E}[Y_i | Z_i, V_i] = \beta_Z^T Z_i' + Z_i'^T \beta_{Z,V} V_i$$

where β_Z are the intercepts of each of these models and $\beta_{Z,V}$ is a set of coefficients that vary as a function of the exposure mapping and the block membership.

We can equivalently write this as

$$Y_i = \beta_Z^T Z_i' + Z_i'^T \beta_{Z,V} V_i + \epsilon_i$$

where $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\epsilon}] = \Sigma$ where Σ is a diagonal matrix which allows for heteroskedasticity. Now let $\boldsymbol{W}_i = \text{vect}(Z_i', V_i Z_i'^T)$ be the vector of covariates in the linear model. Then the OLS estimator has variance $\text{Var}[\widehat{\beta}] = (\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \Sigma \boldsymbol{W} (\boldsymbol{W}^T \boldsymbol{W})^{-1}$.

The problem here is that **it is not immediately clear in a stochastic-block model, when we sample W_i randomly, whether this variance is shrinking or not**. In fact, this is not even so far as a problem with our methods of approximating V_i, but even with the true V_i, this variance does not shrink to zero even in a simple example.

This is true even before we have substituted this ARD. When substituting $\widetilde{V}_i$ in for V_i then we can let $V_i = \widetilde{V}_i + \delta_i$. This is simply an error in variable regression. Where we suppose that $\text{Var}[\delta_i] = \Gamma$ where Γ is a diagonal matrix.

Then the regression with the substituted variables can be written as

$$\begin{aligned} Y_i &= \beta_Z^T Z_i' + Z_i'^T \beta_{Z,V} (\widetilde{V}_i + \delta_i) + \epsilon_i \\ &= \beta_Z^T Z_i' + Z_i'^T \beta_{Z,V} (\widetilde{V}_i) + \varepsilon_i \end{aligned}$$

where $\varepsilon_i = Z_i'^T \beta_{Z,V} \delta_i + \epsilon_i$ and hence the OLS estimator has a corresponding variance.

Examples

We first highlight an example of where this is the case, under an extremely simple scenario.

Let $K = 2$, let $P_{11} = P_{22} = 0.5, P_{12} = 0.2$ and let the clusters be divided evenly in the population. Suppose we have a treatment allocation based on a randomized saturation design where we can assign a treatment to 0.3 of group 1 and 0.7 of group 2. Let $V_i = (a_i, \frac{\boldsymbol{a}^T G_i}{d_i})$ be the exposure mapping referring to the fraction of treated neighbors.

Suppose that the data are generated as:

$$Y_i = 1I(Z_i = 2) + 1a_i + 1\frac{\boldsymbol{a}^T G_i}{d_i} + \epsilon_i$$

where $\epsilon_i \sim N(0, 0.3)$. Suppose that the traits are in perfect alignment with the clusters in that everyone with trait 1 is in cluster 1 and vice versa. The true parameter vector is $\beta = (0, 1, 1, 1)$.

We then consider $\widetilde{V}_i = (a_i, \frac{\sum_{k=1}^{K} \tau_k Q^* X_i}{d_i})$ and run two regressions, one in which we have access to the true covariates Z_i, V_i and the other where we have access to the model approximated covariates $(Z_i, \widetilde{V}_i)$. We let these respective estimators be $\widehat{\beta}_1$ and $\widehat{\beta}_2$ respectively.

Then we run the simulations for $n \in \{100, 200, 300, 400, 500\}$. Figures 1 shows that even if we have access to the true treatments exposures $\frac{\boldsymbol{a}^T G_i}{d_i}$, due to the insufficient variation of the exposures within a cluster, we cannot recover the parameters under this design.

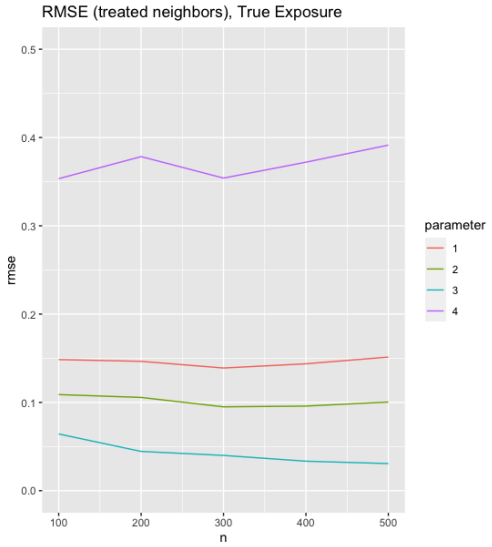

Figure 1: RMSE of $\widehat{\beta}_1$

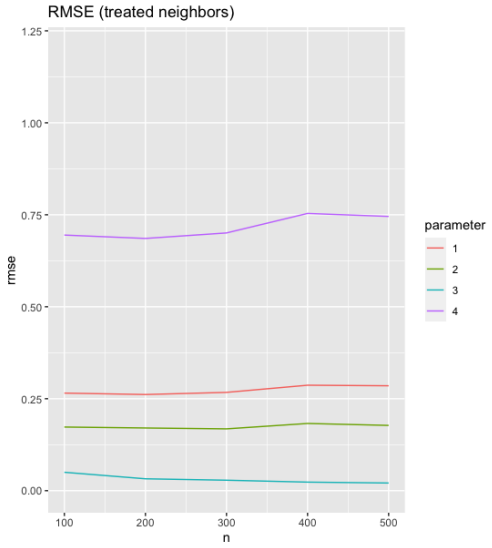

Figure 2: RMSE of $\widehat{\beta}_2$

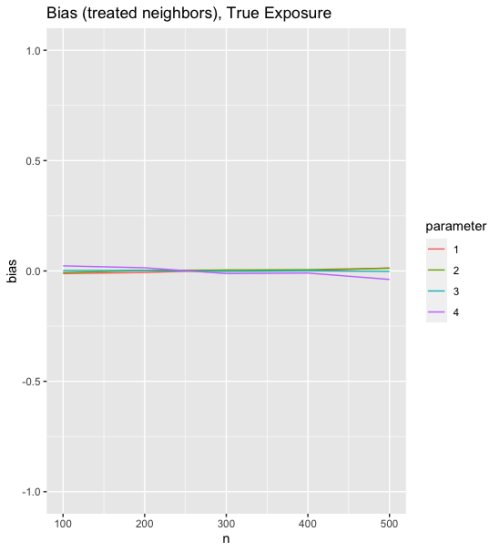

Figure 3: BIAS of $\widehat{\beta}_1$

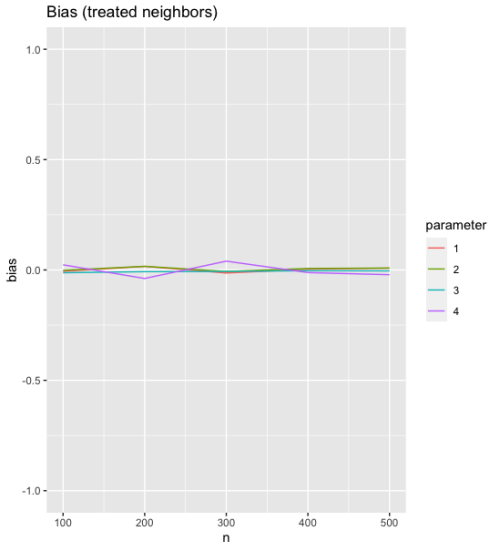

Figure 4: BIAS of $\widehat{\beta}_2$

However, if we consider the baseline values, we can suppose that we have access to outcomes.

$$Y_{i0} = 1I(Z_i = 2) + 1 \times 0 + 1 \times 0 + \epsilon_i$$

Then we can define this OLS estimator $\widehat{\beta}_3$, then we will have a consistent estimator as we see in Figure 5, and its bias in Figure 6. We see that once we add baseline data, this problem is solved and consistency has returned.

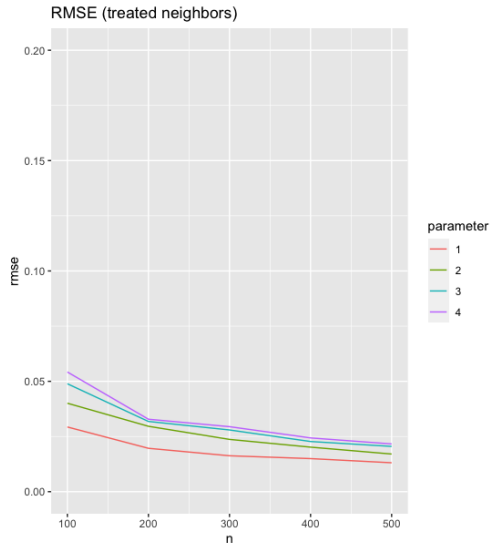

Figure 5: RMSE of $\widehat{\beta}_3$

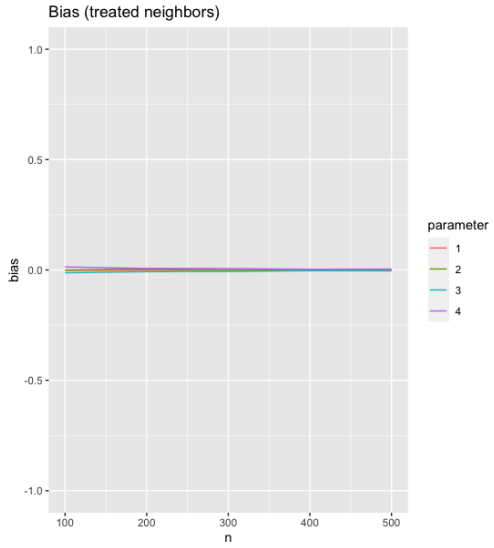

Figure 6: BIAS of $\widehat{\beta}_3$

1.7 Optimal Experimental Design For Linear Models

(TBD)

2 Additional Figures

Here if we want to include a directed acyclic graph of the randomization and treatment, then we have the following.

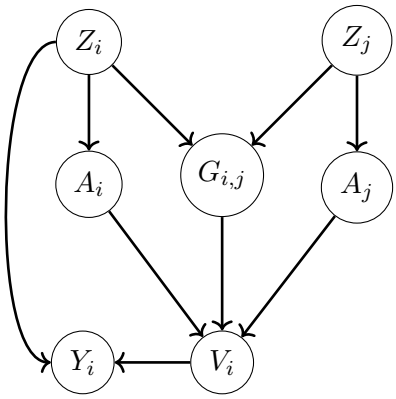

Figure 7: DAG for analogue to Aronow and Samii [2017]

References

P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017. ISSN 1932-6157. doi: 10.1214/16-AOAS1005. URL https://projecteuclid.org/euclid.aoas/1514430272.

E. Breza, A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan. Consistently estimating network statistics using aggregated relational data. *arXiv e-prints*, pages arXiv–1908, 2019.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 2020.

C. Cai, J. Pouget-Abadie, and E. M. Airoldi. Optimizing randomized and deterministic saturation designs under interference. *arXiv preprint arXiv:2203.09682*, 2022.

I. Cristali and V. Veitch. Using embeddings for causal estimation of peer influence in social networks. *arXiv preprint arXiv:2205.08033*, 2022.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

E. L. Ogburn, O. Sofrygin, I. Diaz, and M. J. Van Der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.