MODEL-BASED INFERENCE AND EXPERIMENTAL DESIGN FOR INTERFERENCE USING PARTIAL NETWORK DATA

STEVEN WILKINS REEVES[§], SHANE LUBOLD[°], ARUN G. CHANDRASEKHAR[‡,⋆], AND TYLER H. MCCORMICK[§,¶]

Abstract. The stable unit treatment value assumption states that the outcome of an individual is not affected by the treatment statuses of others, however in many real world applications, treatments can have an effect on many others beyond the immediately treated. For instance, an individual's likelihood of being infected by influenza once a vaccine is available depends on whether their close contacts received the vaccine. Many models of interference involve interference propagated by some network structure. In many empirically relevant situations however, complete network data (required to adjust for these spillover effects) are too costly or logistically infeasible to collect. Partially or indirectly observed network data (e.g., subsamples, aggregated relational data (ARD), egocentric sampling, or respondent-driven sampling) reduce the logistical and financial burden of collecting network data, but the statistical properties of treatment effect adjustments from these design strategies were, until now, largely unknown. In this paper, we present a framework for the estimation and inference of treatment effect adjustments using partial network data through the lens of structural causal models. We also illustrate procedures for better design and analysis of experiments under assumptions of partial network data. We derive single network asymptotic results applicable to a variety of choices for an underlying graph model. We validate our approach using simulated experiments on observed graphs with applications to information diffusion.

1. Introduction

Interference describes the phenomenon where an individual is affected by the treatment[1] received by others; this indirect impact is commonly referred to as a spillover effect. Understanding interference thoroughly is crucial for effectively assessing the impacts of treatment policies on entire populations or specific subgroups. The effects of treatments on others are often propagated through a network structure. Most approaches in estimating causal effects under interference require complete network data, however this is usually infeasible in practice. Instead, one can collect a coarsening (i.e. a subset or aggregate) of the network data. In this paper we introduce a framework for estimation and inference of causal effects under *partial* [2] network data.

Interference and spillovers are common in practice and occurs in multiple scientific domains, including the study of infectious diseases (Hudgens and Halloran, 2008; Tchetgen

Date: This Version: May 15, 2024.

We thank Lori Beaman, Vincent Boucher, Carlos Cinelli, Paul Goldsmith-Pinkham, Kosuke Imai, Fabrizia Mealli, Alex Philip and Alex Volfovsky for helpful comments and discussion. Correspondence: tylermc@uw.edu.

[§]Department of Statistics, University of Washington.

[‡]Department of Economics, Stanford University.

[⋆]J-PAL, NBER.

[°]US Census Bureau.

[¶]Department of Sociology, University of Washington.

[1]As many of our applications pertain to information diffusion, may use the term seed and treatment interchangeably throughout.

[2]We formalize *partial* network data in Section 2.

and VanderWeele, 2012), public policy (Malani et al., 2021; Imai et al., 2021), information diffusion (Banerjee et al., 2013, 2019), technology adoption (Beaman et al., 2021), online platforms (Saveski et al., 2017; Pouget-Abadie et al., 2018, 2019) and online marketplaces (Ha-Thuc et al., 2020; Johari et al., 2022). A large portion of the literature considers how one is influenced by their peers (for instance, in educational attainment) (Manski, 1993; Bramoullé et al., 2009; De Giorgi et al., 2010; Epple and Romano, 2011; Goldsmith-Pinkham and Imbens, 2013).

The violation of the stable unit treatment value assumption (SUTVA) has substantial implications for statistical inference because it makes it difficult to clearly differentiate treatment units from control units without many independent networks. In principle, each potential outcome (the counterfactual outcome under a treatment assignment) can potentially be a function of all other treatments in the population. This leads to numerous new causal effects beyond the usual average treatment effects, however, in order to learn these from data, one must assume some additional structure in the problem as only one treatment assignment will be observed which can influence all units. A useful, somewhat general vocabulary is to write causal effects through exposure maps. Given a fully measured network, a combination of individual heterogeneity, including but not limited to network position, and the full treatment vector can be reduced to exposure classes of treatment (Aronow and Samii, 2017). With a bounded degree assumption, identification, consistency, and asymptotic normality of the treatment effects can be identified (though this condition can be relaxed as was illustrated in Chandrasekhar et al. (2023)). A semiparametric theory has also been developed for fully observed networks with bounded degrees (van der Laan, 2012; Ogburn et al., 2022).

If only a coarsening of the network is observed (i.e. a subset or aggregation of network information), then naturally this leads to a misspecification of the exposure map, and is a challenge in most real datasets involving network data.

We are also interested in how to optimally allocate treatments/seeds, for instance, to maximize the propagation of information diffusion. However, it does not follow immediately from the exposure mapping framework alone how to transfer the knowledge of one experiment to a new population since potential outcomes can be arbitrarily different across individuals with the same exposure. In contrast, many empirical applications propose some mechanism by which interference is propagated; as it is important to understand the mechanism in order to guide policy in the future (e.g. technology adoption (Beaman et al., 2021) or information diffusion (Banerjee et al., 2019)). Pearl (2009) refers to these two paradigms as the second (reasoning about effects) and third (reasoning about counterfactuals) rungs on the ladder of causality (with the first being reasoning about associations). In many cases, inference on counterfactuals is exactly what is required.

A major constraint however, is that we require a framework for estimating causal effects using partial network data, since complete network data is often infeasible to collect. In this paper, we represent commonly used mechanisms of spillovers through a class of structural causal models (SCMs) (Pearl, 2009), and illustrate estimation procedures for causal effects using partial network data, which we define as some coarsening of the original network $G^* = \zeta(G)$. For example, this could include a subgraph or summary statistics such as aggregate counts of the network (i.e. number of connections to nodes with a given trait t, which is exactly aggregated relational data (ARD)).

In this paper we first discuss the related literature on both interference and partial network data collection. We next introduce the class of structural causal models which will serve as the basis for the defining causal effects. Afterwards, we contrast this other common finite sample interpretations of interference. Later we illustrate how, given a parameterization of the outcome model, one can recover the parameters using partial network data through an

iterated expectations argument. In doing so, we must impose a generative model for the network, which we choose to be a class of node-exchangeable graphons. Our estimation procedures require the estimation of the graph model, which we parameterize through common models of graph formation, such as stochastic block-models and latent space models. We illustrate the estimation rates required in order to estimate these network models for valid asymptotics. The core intuition is that in order to develop an estimator, we first assign a graph model which can be estimated at a rate such that the estimation of the model parameter is negligible. In lieu of using the full network data, we can infer outcome model parameters using the distribution of the graphs conditional on the observed data.

We then leverage these asymptotic results in order to establish more efficient experimental design procedures for learning the outcome model, and introduce a Bayesian optimization approach that can be applied for both seeding and experimental design. We lastly illustrate these aspects of using our framework for estimation, seeding and experimental design in a set of semi-synthetic experiments.

We find that leveraging this model structure can lead to more efficient estimation of causal parameters, such as the global average treatment effect, even compared to full data methods like Horvitz-Thompson estimators. Additionally, we find that using experimental design can achieve greater efficiency in estimating parameters than merely collecting complete network data. Finally, we demonstrate how our framework can be used to develop novel seeding strategies using various types of partial network data.

1.1. **Related Work.** Complete network data may be extremely expensive to collect restricted due to privacy considerations (Breza et al., 2020). Researchers often encounter partial network data, whether from survey samples of social interactions, coarse geographic and mobility information, kinship data from censuses, or aggregated transaction flows between banks and firms among others.

The main intuition is that each of the partial network designs mentioned above can be used to estimate a breakdown of each respondent's network in terms of observable characteristics. ARD asks about these characteristics directly. Subsampling, along with characteristic information about each sampled node, can be used to estimate this breakdown within the observed subgraph, similar to egocentric sampling and respondent-driven sampling. Clustering across these observed characteristic groups provides estimates of latent types for each respondent based on the composition of their network, and these latent types can then be used to estimate mixing across latent type groups in the network. Breza et al. (2023) utilize a similar framework, but only in the context of ARD and to estimate statistics about the unobserved network (e.g. centrality) and not studying interference directly.

Perhaps the most direct approach is the subsampling of nodes, where a fraction of nodes are selected from the population and the researchers enumerate the complete graph between those nodes. If random sampling of nodes is infeasible, or if populations are particularly sensitive or subject to stigma, a link-tracing design like snowball sampling or respondent-driven sampling provides a partial, though selective, view of the underlying graph Heckathorn (1997); Goel and Salganik (2009, 2010); Green et al. (2020).

When enumerating edges, even among a subset of nodes, is infeasible, researchers turn to methods that are possible using standard surveys. One possible survey method involves the collection of aggregated relationship data (ARD). Rather than collecting all edges or relationships in a network of interest, a researcher conducting an ARD survey asks respondents "How many people do you know with trait X?" for various traits. ARD was originally proposed to estimate the size of hard-to-reach populations, such as the number of HIV-positive men in the United States (Killworth et al., 1998; Scutelniciuc, 2012; Jing et al., 2014) and has been used in further applications such as models of financial contagion (Acemoglu et al.,

2015). ARD has been shown to be 70 to 80% cheaper to collect than full network data (Breza et al., 2020) and has been shown to be useful in many network inference problems (Breza et al., 2020, 2023). Egocentric sampling is another approach commonly used on standard surveys. In this design, researchers ask respondents to think about specific individuals in their network (rather than aggregates, as in ARD) and then answer questions about those individuals. Inference for parameters under missing network settings was developed in Chandrasekhar and Lewis (2011) as a general framework for correcting inference based on network data with missingness. While asymptotic results were originally developed for multiple independent networks, we extend these results to enable inference in the single network setting.

Under interference, the first task is to define the estimand of interest. For instance, one may want to estimate the global average treatment effect (GATE) which is the effect of treating everyone, integrating over the various peer effects, as compared to treating no one Ugander et al. (2013). In other circumstances one may be interested in predicting the effect of a particular treatment allocations. This for instance includes finding the most influential individuals Kempe et al. (2003); Banerjee et al. (2019). These particular combinations may be restricted, for instance, because a policymaker may face budget constraints (e.g., subsidies are limited to the ultra-poor, agricultural extension focuses on model farmers, Anderson and Feder (2007)). Or it may be because fundamentally the peer effects dynamic is thought to be non-monotone (Banerjee et al., 2018): treating everyone may change interaction dynamics in equilibrium. A more general object is simply comparing the average treatment effect difference between two exposure configurations (Aronow and Samii, 2017).

For the most part, the literature on SUTVA violations and exposure maps has focused on settings where the graph is perfectly observed, with some exceptions for specific average causal effect estimation Sävje et al. (2021); Yu et al. (2022); Cortez et al. (2022). And while this is certainly both a reasonable place to start and one that may work in a number of settings (e.g., social media), in myriad contexts researchers and policymakers do not have access to such granular interaction data. Instead, we will use models for peer influence, such as contagion (Banerjee et al., 2013; Beaman et al., 2021; He and Song, 2023), or hearing models (Banerjee et al., 2019) are used in order to give structure to the peer effects. A similar understanding of these effects through structural causal models is proposed in Auerbach and Tabord-Meehan (2021), however their focus is on nonparametric estimation, while our focus involves estimation, inference and design with partial network data. A distinct but related line of work seeks to detect whether interference is present at all (Athey et al., 2018).

2. Environment

2.1. **Data.** In this section, we introduce the relevant notation. Let $i \in \{1, 2, \ldots, n\} = V$ denote a populations of interacting individuals and let $\mathcal{G} = V \times E$ be the network[3] by which interference is propagated; where V is the set of node vertices and $E \subset V \times V$ is a set of edges (either directed or undirected). We can represent this graph by the adjacency matrix $G \in \{0, 1\}^{n \times n}$. We consider binary treatments denoted by a treatment vector $\boldsymbol{a} \in \{0, 1\}^n$ and let denote the potential outcome $Y_i(\boldsymbol{a}) \in \mathbb{R}$, under a treatment assignment $\boldsymbol{a}$, and Y_i denote the actual observed outcome. Lastly, we assume that we have access to pre-treatment node-level covariates $X_i \in \mathbb{R}^m$.

In the following sections let O and o denote the usual big and little oh notation and O_P and o_P denote the stochastically bounded and convergence to 0 in probability for sequences of random variables.

[3]We can also extend this to weighted graphs, however binary networks are presented for simplicity.

2.2. **A structural causal model.** In order to develop an estimation strategy for causal effects using partially measured networks, we use the framework of structural causal models, a nonparametric extension of structural equation models (Pearl, 2009). Similar approaches have been studied by Ogburn et al. (2022) and Auerbach and Tabord-Meehan (2021) in the case of fully observed networks.

We let $Y_i(\boldsymbol{a})$ denote the potential outcome of Y_i under a treatment allocation $\boldsymbol{a}$. The exposure mapping V_i is represented as a function f_V such that $V_i = f_V(\boldsymbol{a}, \varphi_i(G)) \in \mathbb{R}^{p_V}$ where φ_i is the relevant graph information for individual i relative to their position with respect to treated individuals. We also allow for the potential outcome to be modulated by some additional confounder $S_i = f_S(\boldsymbol{X}, \vartheta_i(G)) \in \mathbb{R}^{p_S}$. We propose the following general structural model for the outcomes Y_i:

$$
\begin{aligned}
\boldsymbol{a} &\sim P_{\boldsymbol{A}} \\
S_i &= f_S(\boldsymbol{X}; \vartheta_i(G)) \\
V_i &= f_V(\boldsymbol{a}; \varphi_i(G)) \\
Y_i &= f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y)
\end{aligned} \tag{1}
$$

The benefits of structural causal models are that they allow for the characterization of all causal effects in a system, as well as the distributions of counterfactuals. However, they require correct specification of the causal process, i.e. correct specification of the exposure map and the relevant confounders. Even if one can propose a model for interference, estimation is not straightforward due to the fact that we only observe partial graph information in G^*. We will also demonstrate that many common models of interference can be expressed as structural causal models, and can be thought of as parameterizations of $f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y) = f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y; \beta_0)$. We will then illustrate how one can recover the model parameters using partial network data.

We differentiate the two types of target parameters. The first are the **outcome model** parameters which parameterize the distribution of (Y, S, V). Specifically, $f_Y(S_i, V_i, \epsilon_Y) = f_Y(S_i, V_i, \epsilon_Y; \beta_0)$ for some parameterization β. We denote these parameters $\beta_0 \in \mathbb{R}^p$. Such parameters may be defined through a moment equation m, $\mathbb{E}[m(Y_i, S_i, V_i, \beta_0)] = 0$ or more explicitly through a regression parameterization. In a model of simple diffusion, this is simply the probability of infecting a neighbouring node $q \in [0, 1]$.

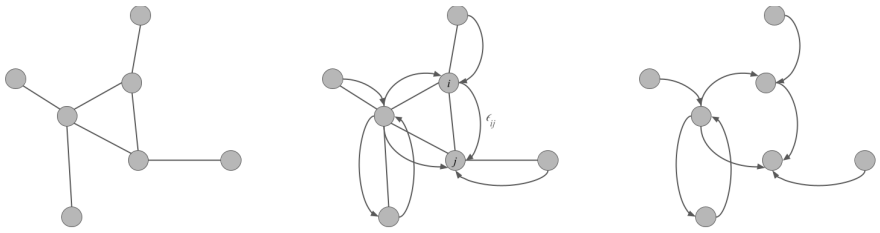
The second set of parameters we consider are the **causal** parameters. We can define these parameters independent of any model. The main causal parameter we will consider is the expected average potential outcome on the complete network G, $\Psi(\boldsymbol{a}|G) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i(\boldsymbol{a})]$. This quantity will implicitly also depend on the graph model G. Similar target parameters can be defined through an average potential outcome conditional on a covariate x: $\Psi(\boldsymbol{a}|x, G) = \frac{1}{n_x}\sum_{i=1}^n \mathbb{E}[Y_i(\boldsymbol{a})]I(X_i = x)$ where $n_x = \sum_{i=1}^n I(X_i = x)$. Leveraging the structural causal model, we can define the these causal effects in terms of the structural causal model. We illustrate conditions for identification of these causal effects in Section 2.4. While we concentrate on defining causal quantities through conditional means, the nonparametric identification argument is broadly applicable and can be extended to other functionals of the conditional distribution of potential outcomes, such as quantiles.

Inference for the causal relationship between Y_i and V_i amounts to learning the relationship between Y_i and S_i, V_i. We consider settings where the assignment of treatments can be manipulated by an experimenter, which we discuss in Section 4. If one leverages this model, either through assumption or estimation, of the generation of the outcomes then one can use a structural causal model to generate expected potential outcomes under different

treatment assignments $f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y)$, which is precisely what is done in the case of seeding. A contrast of these frameworks is included in the appendix in Section 7.1.

Adding such structure to the model of potential outcomes is common in some fields, such as economics, where a researcher often proposes a micro-foundation model of the mechanism by which information or behavior is propagated across a network[4] and fall into our structural causal model framework. For example Banerjee et al. (2013) consider a latent diffusion process for information passing of micro-lending, Banerjee et al. (2019) consider a hearing model of information diffusion, and Beaman et al. (2021) consider a complex contagion approach to behavior adoption of agricultural techniques. Moreover Centola and Macy (2007) suggest that *information* is often propagated by single links, whereas *behavior* often requires multiple neighbors to achieve network spread. Our framework also includes common structure in linear in means type models (Manski, 1993) of which extensions have also been used to identify causal effects such the global average treatment effect (Chin, 2019). Later on we highlight how to leverage experimental design to better estimate model parameters, and how one can leverage partial network data for seeding. Next we highlight how a common mechanism of diffusion can be thought of as a structural causal model.

2.2.1. *Example: Diffusion as a structural causal model.* We highlight simple diffusion can be understood as a structural causal model, as this is often a simple baseline model from which more complex models of interference are built from (e.g. Banerjee et al. (2013)). We will illustrate how given a set of exogenous noise, it can be used to describe all counterfactual outcomes in the population. Suppose that at time 0, a set of seeds $\boldsymbol{a}$ are infected. At each time $t \in \{1, 2, \ldots, T\}$ the infection is passed to each of its neighbors with probability q, after which, the infected node is no longer infectious. Denote whether a node is infected at time t with $Y_{it} = 1$. We consider the outcome $Y_i = 1$ if a node was infected at any point during $t \in \{0, 1, 2, \ldots, T\}$. We highlight a simple example for $T = 2$. Let the exogenous noise variable $\boldsymbol{\epsilon}_Y$ as a set of Bernoulli random variables $\epsilon_{ij} \sim \text{Bernoulli}(q)$ which indicate whether i would pass an infection to node j if it were infected. Let $\boldsymbol{E}_{ij} = \epsilon_{ij}$ and $\boldsymbol{D} = \boldsymbol{E} \odot G$. We illustrate this in Figure 1.

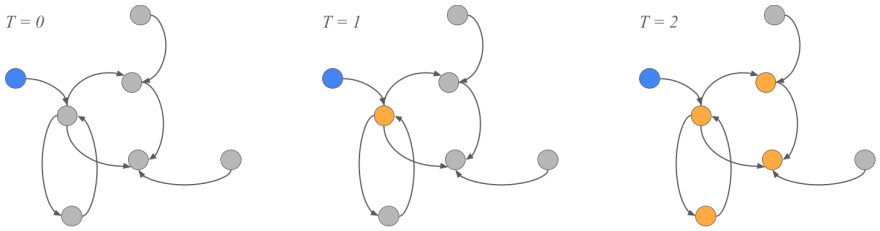


(a) Base Network (b) Noise draw ϵ_{ij} (c) Directed Network $\boldsymbol{D}$.

FIGURE 1. Draw of directed network $\boldsymbol{D}$.

Therefore, given a random sample of the directed graph $\boldsymbol{D}$, one can characterize what would have happened if a node were treated, which is precisely the counterfactual. For instance in Figure 2 we seed the left most which proceeds to propagate in steps 1 and 2. In fact given the draw $\boldsymbol{D}$, this infected node will maximize the spread, when it is not the optimal seed when averaging over samples of the contagion. Additionally, one can construct a relevant exposure map for any fixed number of time steps T.

[4]Many of these models include an element of time. In our setting, we suppose that we only observe the outcomes after some pre-defined fixed time T, i.e. $Y_i(\boldsymbol{a}) = Y_{i,T}(\boldsymbol{a})$.



(a) Contagion process at time period $T = 0$.

(b) Time period $T = 1$.

(c) Time period $T = 2$.

FIGURE 2. Contagion process where a single node is seeded in time $T = 0$ in blue, and infected nodes displayed in orange at times $T = 1$ and $T = 2$.

2.2.2. *Examples of Exposure Maps.* It is instructive to consider several different examples of exposure maps used in practice. We provide these as choices which can fit into our framework.

Local Interaction Effects

Simple examples of local network effects may include the total number of treated neighbors

$$V_i = \sum_j G_{ij} a_j$$

or the treated fraction of one's neighbors

$$V_i = \sum_j \frac{G_{ij}}{d_i} a_j$$

where $d_i = \sum_j G_{ij}$ is degree.

Risk-Sharing Networks (Ambrus et al., 2014).

Equilibrium risk sharing is that the graph consists of C mutually exclusive communities such that any endowment vector within the community is aggregated and shared evenly. Let treatment $\boldsymbol{a}$ be an "endowment" and let $\overline{\boldsymbol{a}}_c = \sum_{j \in c} a_j$ be the sum of the endowment vector for community c, with $|c|$ denoting its size. Then

$$V_i = f_V(\boldsymbol{a}, \varphi_i) = \overline{\boldsymbol{a}_c} \cdot |c|^{-1}.$$

That is, the exposure is just a function of the total endowment of the community and nothing more.

Hearing Information

Many phenomena in the world, such as the spread of infectious diseases, information, or social behaviors, can be effectively modeled by contagion processes. These processes describe how certain phenomena propagate through networks (Keeling and Rohani, 2008; Centola and Macy, 2007; Barrat et al., 2008; Pastor-Satorras et al., 2015; Cencetti et al., 2023).

We consider a message-passing model similar to that employed in Banerjee et al. (2019). Here, the treatments $\boldsymbol{a}$ represent a seed piece of information disseminated across a population via a network over time steps t, where t ranges from 1 to T, and no further message propagation occurs after T time steps. Subsequently, we define the "hearing matrix" $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{H}_{(0)}, T) = \sum_{t=1}^{T} \left(\boldsymbol{H}_{(0)}\right)^t$, where $\boldsymbol{H}_{ij}$ denotes the expected number of times, after T time steps, that person j hears a piece of information originating from person i, and $H_{(0),ij}$ represents the probability of transmitting a message from person i to person j.

Let $N_{j,t}$ denote the number of times the person j hears a piece of information and let $N_j = \sum_{t=1}^{T} N_{j,t}$ denote the total number of messages received over the time period. We

suppose that a response Y_i is related to the number of times they received an exposure through a link function Λ.

Given a seed vector $\boldsymbol{a}$, we can define the expectation of N_j as:

$$\mathbb{E}[N_i|\boldsymbol{a}] = \sum_{t=1}^{T} \boldsymbol{e}_i^T \boldsymbol{W}^t \boldsymbol{a}.$$

Lastly, we relate this to the outcome

$$\mathbb{E}[Y_i|N_i] = \Lambda(\beta_0 + \beta_1 M_i)$$

For example, in Banerjee et al. (2019) $Y_i \in \{0, 1\}$ denotes whether an individual called in to a free lottery.

In general, N_i is not known from the graph itself, but $\mathbb{E}[N_i|\boldsymbol{a}]$ can be calculated from a seed vector and the graph G. When Λ is linear, this expectation can be passed directly through; otherwise, one can include an over-dispersion random effect term $\varepsilon_i = M_i - \mathbb{E}[N_i|\boldsymbol{a}]$. A simple diffusion example is to suppose that there is a common probability δ which indicates the probability of passing a message along any edge of the network $\boldsymbol{W} = \delta G$ for some $\delta \in [0, 1]$ where the information passes directly across the nodes of a network.

The simplest version of the model does not account for additional heterogeneity in the susceptibility to influence. For example, individuals with higher degrees $d_i = \sum_{j=1}^{n} G_{ij}$ may be less susceptible to influence by any one of their neighbors, suggesting the need for a degree effect.

$$\mathbb{E}[Y_i|N_i] = \Lambda(\beta_0 + \beta_1 N_i + \beta_2 d_i).$$

In our setting, the model class we consider is fairly general. We can include features based on covariates at node level and connections to others in a graph mapped through S_i and similarly for exposure mapping V_i. For example, we could suppose that at each step of the network, there is a different effect on the outcome as parameterized by the following linear model.

$$\mathbb{E}[Y_i|S_i, V_i] = \Lambda(\beta_0 + \sum_{t=1}^{T} \beta_t \boldsymbol{e}_i^T G^t \boldsymbol{a}) = h(S_i, V_i; \beta)$$

In general, V_i can can also interact with the node-level covariates X_i, or graph level information such as degree d_i. In the remainder of this paper, we assume that the exposure mapping is known.

2.3. **The Missing Data Problem.** In our setting, we do not have access to the full graph G, but rather, have access to some partial set of information $G^* = \zeta(G)$. Some examples include the induced subgraph or egocentric sampling (Freeman, 1982; Almquist, 2012) , respondent driven sampling (Heckathorn, 1997), aggregated relational data (Killworth et al., 1998), respondent driven sampling (Heckathorn, 1997; Goel and Salganik, 2009, 2010; Green et al., 2020) and more. The well known "national longitudinal study of adolescent to adult health" (add health) dataset asked students to list up to 5 friends in their social network; a form of missing data (Harris et al., 2019). We highlight several examples of partial network data below. Later on in Section 3.6 We discuss how these coarsened data can be used to estimate models of network formation.

Example 2.1 (Induced subgraph)**.** *We sample $m \leq n$ of nodes in the graph randomly, with at least one node from each of the K communities. Let $G^* = G_{I_m, I_m}$ be the sub-graph induced by these m nodes where $I_m \subset \{1, 2, \ldots, n\}$ are the set of nodes that are sub-sampled from the whole population.*

Example 2.2 (Respondent Driven Sampling). *Let $i \in I_m \subset \{1, 2, \ldots, n\}$ denote the indices of a sample of individuals obtained through respondent driven sampling. An initial number of individuals are recruited as seeds, and subsequent individuals are recruited via referrals from the others in a population. Under this process we receive a subgraph of connected individuals G_{I_m, I_m} as well as the list of connections to additional nodes $I_{n\setminus m} = \{1, 2, \ldots, n\} \setminus I_m$ $G^* = G_{I_m, I_m}, G_{I_m, I_{n\setminus m}}$.*

Example 2.3 (Aggregated Relational Data). *Aggregated relational data consists of aggregated sums of connections to nodes of a given trait. Typically this is collected from a survey consisting of questions such as "How many many people do you know with [X] trait?". For a set of T traits, this consists of*

$$X^*_{it} = \sum_{i=1}^{n} G_{ij} I(t_j = t)$$

In many examples, partial network data is used to infer properties of the missing part of the graph, using a generative model, which we discuss in greater detail in Section 3.6.

In order to infer about the distribution of the missing part of the graph, we propose that $G \sim \theta_0$ where we assume that $\theta_0 \in \Theta$ denotes the parameters of a random graph model. In this case, for each i, there is an a latent ξ_i parameter such that

$$P(G_{ij} = 1 | \xi_i, \xi_j) = \widetilde{g}(\xi_i, \xi_j)$$

for some function symmetric, measurable $\widetilde{g}$ known as a graphon Lovász and Szegedy (2006). Many common graph models, such as latent space models Hoff et al. (2002), are included in this category. Graphons are appealing in this context because, following Airoldi et al. (2013); Gao et al. (2015), they can be approximated arbitrarily well using latent types assigned to each node. Said another way, graphons introduce complex dependence in the network-generating mechanism through clustering induced by latent types associated with each node.

2.4. **Nonparametric Identification of Causal Effects.** In order to identify the causal parameters from the data, we must ensure that the potential outcomes satisfy the following properties.

Definition 2.1 (Exposure Weak Ignorability). *We say that an exposure assignment is **weakly ignorable** if the following holds:*

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

Once we condition on S_i, then the potential outcomes are independent of the actual exposure received, i.e. all potential confounding is through the graph statistic S_i. For instance, in simple contagion models nodes are considered equivalent and this holds trivially without the need for conditioning. In general, one should condition on any quantity which drives the heterogeneity of the potential outcomes. In Section 5.1 we consider an example used by Ugander and Yin (2023) where conditioning on degree is sufficient for any randomization.

Definition 2.2 (Exposure Consistency). *Exposure consistency holds if*

$$V_i = v \implies Y_i = Y_i(v) = Y_i(\boldsymbol{a})$$

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

Exposure consistency can be thought of having a correct exposure mapping. Meaning that the potential outcomes are realized whenever a particular exposure is observed.

Definition 2.3 (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*

$$Y_i(\boldsymbol{a}) \perp\!\!\!\perp G | V_i, S_i$$

The last assumption states that once we have adjusted for V_i and S_i, then the potential outcomes are independent of the network G.

Under these assumptions, the causal effects can be identified through conditional distributions of the observed data.

$$\begin{aligned} P(Y_i(\boldsymbol{a}) = y | S_i = s, G) &= P(Y_i(v) = y | S_i = s, G) \text{ By the exposure mapping} \\ &= P(Y_i(v) = y | V_i = v, S_i = s, G) \text{ By Weak ignorability} \\ &= P(Y_i = y | V_i = v, S_i = s, G) \text{ By Consistency} \\ &= P(Y_i = y | V_i = v, S_i = s) \text{ Graph Conditional Independence} \end{aligned}$$

$$\implies \Psi(\boldsymbol{a}|G) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i | V_i = f_V(\boldsymbol{a}, \varphi_i), S_i = f_S(\boldsymbol{X}, \vartheta_i)]$$

For brevity, we denote the true conditional mean $\mathbb{E}[Y_i | V_i = v, S_i = s] = h_0(s, v)$. Therefore, the identification of the causal model can be achieved through correct modeling of the potential outcome. We therefore consider a model class $h(s, v; \beta)$ that will be used to model the outcome $h_0(s, v)$.

One may also be interested in the expected average potential outcome on a single network. However, we may not observe that network directly. Given a network model θ_0, observed graph data G^*, and a conditional model $h(s, v; \beta)$ we can also define the expected average treatment effect,

$$\Psi(\boldsymbol{a}|\beta, G^*, \theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[h(f_V(\boldsymbol{a}, \boldsymbol{X}; \varphi_i), f_S(\boldsymbol{X}; \vartheta_i); \beta) | \boldsymbol{a}, \boldsymbol{X}, G^*, \theta] \tag{2}$$

where under the correct model conditional model and graph model $\mathbb{E}[\Psi(\boldsymbol{a}|G)|\boldsymbol{a}, \boldsymbol{X}, \theta_0] = \Psi(\boldsymbol{a}; \beta_0, G^*, \theta_0)$ where we average over draws of the graph model, conditional on the observations. Later in Section 3.5, we illustrate when this population average effect under any draw of the network $\Psi(\boldsymbol{a}|G)$ will be close to the average over the model class $\Psi(\boldsymbol{a}; \beta_0, G^*, \theta_0)$.

2.4.1. *Why not IPW estimators?* In many nonparametric approaches to estimating causal quantities under interference, inverse probability weighted (IPW) estimates can be developed given a randomization scheme $P_{\boldsymbol{A}}$ (Aronow and Samii, 2017). This is useful as it can be used to develop estimators for causal effects any exchangeability assumptions on the potential outcomes. However when V_i is not observed directly, we must leverage additional structure in order to estimate any causal effects.

Our objective is to understand the model's structure and often apply it to tasks such as seeding. Thus, we rely on a correct model specification. The challenge with developing an IPW estimator arises when exposure is not observed. In such cases, it becomes impossible to determine which potential outcome was observed, violating the causal consistency assumption. Specifically, we don't know which potential outcome Y_i represents (i.e., which exposure $v, Y_i = Y_i(v)$).

3. Inference

We next discuss our general procedure for the estimation of parameters using partial network data. We illustrate the algorithm for performing this estimation in Section 3.2 followed by some asymptotic results. In order to develop the asymptotic results, we require

two theoretical tools. Firstly, this is an appropriately fast rate of estimation of the network model parameters θ_0. The second is an appropriate central limit theorem for examples where the outcomes are believed to be correlated.

3.1. **Outcome Model Parameters and Estimators.** Next we consider estimating the outcome model parameters β_0. We present two methods for estimating such parameters, instrumentation in a linear model, and Z estimators.

3.1.1. *Estimation in Linear Models.* We first illustrate a methods of identification of the conditional model under a linear model assumption.

$$Y_i = \beta_0^T f(S_i, V_i) + \varepsilon_i$$

where $\mathbb{E}[\varepsilon_i] = 0$ and there can be general correlation $\text{Var}[\boldsymbol{\varepsilon}] = \Sigma$. In general, we do not have access to network data, however one can recover the model parameters through conditional expectation

$$\mathbb{E}[Y|S(G), V(G), G, \boldsymbol{a}, \boldsymbol{X}] = \beta_0^T f(S(G), V(\boldsymbol{a}, G))$$
$$\mathbb{E}[\mathbb{E}[Y|S(G), V(G), G, \boldsymbol{a}, \boldsymbol{X}]|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0] = \beta_0^T \mathbb{E}[f(S(G), V(\boldsymbol{a}, G))|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0]$$

where we create a new set of features $F_i = \mathbb{E}[f(S_i(G), V_i(\boldsymbol{a}, G))|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0]$ using the marginalization over samples of the network model. Here identification comes from the variation of these averaged features F_i over the population. More clearly, letting $\boldsymbol{F} \in \mathbb{R}^{n \times p}$ denote the design matrix of this model, identification comes from the linear independence of the columns of $\boldsymbol{F}$. We discuss the conditions required for the correlation matrix for estimation in practice in Section 3.

3.1.2. *Z estimators.* In other cases, parameters may be defined through a moment equation, and can be used to construct a Z-estimator. These may include parameters in a moment equation such as generalized linear models (GLMs) $\mathbb{E}[Y|S, V] = \Lambda^{-1}(\beta_0^T f(S, V))$. These parameters can be identified using an estimating equation approach where given a moment function $\widetilde{m}(Y_i, S_i, V_i; \beta)$ such that $\mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta)|\boldsymbol{a}, \boldsymbol{X}, G] = 0$ if and only if $\beta = \beta_0$. Through the use of iterated expectations, we can define a new estimating equation, by marginalizing over the draws of the graph model

$$m_i(Y_i; \beta, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta) = \mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta)|Y_i, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta] \tag{3}$$

then applying iterated expectations

$$\mathbb{E}[m_i(Y_i, S_i, V_i; \beta_0, \theta_0)|G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0] = \mathbb{E}[\mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta_0)|\boldsymbol{a}, \boldsymbol{X}, G]|G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0] = 0.$$

Identification in this case comes from the variation of the exposure and the confounders, such that

$$\mathbb{E}[m_n(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{V}; \beta, \theta_0)|G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} m_i(Y_i, S_i, V_i; \beta, \theta_0)\middle| G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0\right] = 0$$

if and only if $\beta = \beta_0$. Exact conditions will vary depending on the specific parameter estimated, however for GLMs a similar identification strategy as the linear model can be used.

3.2. **Inference with partially measured data.** We introduce a general procedure for estimating the outcome model parameters. We also illustrate inference for estimation of a causal target parameter on a particular graph G. Later in Section 4 we illustrate how to more efficiently design experiments in order to estimate the relevant parameters. We present a pseudo-code approach to the estimation procedure as follows. Let $\widetilde{Z}_i = (Y_i, S_i, V_i)$ denote the full (including unobserved) data, and let $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{a}, \boldsymbol{X}, G^*)$ denote the observed data.

(1) Define an model for the relationship of $\boldsymbol{Y}$ given the exposures $\boldsymbol{V}$ and confounders $\boldsymbol{S}$ (for instance, a regression model $\mathbb{E}[Y|V,S] = h(v,s;\beta_0), \beta \in \mathcal{B} \subset \mathbb{R}^p$ with parameters which can be estimation via the estimating function $\widetilde{m}_n(\widetilde{\boldsymbol{Z}};\beta)$. Let $\widetilde{m}_n(\widetilde{\boldsymbol{Z}};\beta) = \frac{1}{n}\sum_{i=1}^n m(\widetilde{Z}_i;\beta)$ denote the empirical estimating function.
(2) Estimate a model of the network, using the node-level covariates $\widehat{\theta} := \widehat{\theta}(G^*)$.
(3) Estimate $\widehat{\beta}$ by solving the estimating equation $m_n(\boldsymbol{Y};\widehat{\beta}, G^*, \widehat{\theta}) = 0$, where $m_n(\boldsymbol{Y};\widehat{\beta}, G^*, \widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n m_i(Y_i;\beta, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta)$ where m_i is defined in equation (3).
(4) (optional) Plug in $\widehat{\beta}$ to $\Psi(\boldsymbol{a}|\widehat{\beta}, G^*, \widehat{\theta})$.

Step 1 asks the practitioner to propose a response model given the treatment, i.e. the causal model in Section 2.2. Step 2 is a standard step in which we estimate the generative model given the partial network data and the node covariates observed. Step 3 estimates the parameter by marginalizing the estimating function over the graph model. Lastly, Step 4 is optional if the target parameter is a plug-in estimator of the causal parameter using the regression model. We provide an example of this in Section 5.1.

This procedure is standard for estimating parameters partially measured networks (Chandrasekhar et al., 2013; Breza et al., 2023). In the following section, we provide additional details for asymptotics results that the previous literature lacked.

As we have already discussed Step 1 in the as a motivation for our problem, we illustrate the rest of the steps in further detail in the following subsections. First, we discuss the estimation of the stochastic blockmodel using a variety of data types (Step 2). Included in this section is a rate of estimation with respect to the graphon model. Secondly we discuss inference on the parameter $\widehat{\beta}$ which is estimated using the marginalized estimating equation (Step 3). Lastly we discuss inference for the plug-in estimate of causal parameters (Step 4).

3.3. **A central limit theorem for dependent data.** Next we consider the regularity conditions for asymptotic normality of Z-estimators. In order to do so, we utilize the central limit theorem derived in Chandrasekhar et al. (2023). The conditions therein are implied by common assumptions of dependence such as M-dependence which includes α-,ϕ- or ρ-mixing (Bradley, 2005). We begin first with a definition of affinity sets then proceed with the theorem.

Definition 3.1 (Affinity sets)**.** *Denote a triangular array of mean* 0 *random vectors* $W_{1:n}^{(n)}$ *with dimension* p*. Let* $\mathcal{A}_{(i,d)}^{(n)}$ *denote an affinity set which contains all of the variables in the triangular array which are highly correlated with* $W_{i,d}^n$*, the* d^{th} *dimension of the* i^{th} *random variable.*

The affinity sets can be used to construct a matrix which contains the bulk of the covariance across observations and dimensions. Let $||\cdot||_F$ denote the Frobenius norm. If the matrix $\Gamma_{n,dd'} = \sum_{i=1}^n \sum_{(j,d')} \text{cov}(W_{i,d}^{(n)}, W_{j,d'}^{(n)})$ satisfies the following regularity conditions, then a central limit theorem holds $\Gamma_n^{-1/2} S_n \to_d N(0, I_p)$ where $S_n = \sum_{i=1}^n W_i^{(n)}$. The regularity conditions can be understood as control of the covariance within affinity sets (4), control of the covariance across affinity sets (5) and control of the covariance outside of the

affinity sets (6). We collectively refer to these as the affinity set conditions.

$$\sum_{(i,d):(j,d'),(k,d'')} \mathbb{E}[W_{i,d}|W_{j,d'}W_{k,d''}] = o(||\Gamma_n||_F^{3/2}), \tag{4}$$

$$\sum_{(i,d),(j,d');(k,d''),(l,\widehat{d})} \text{cov}(W_{i,d}W_{k,d''}, W_{j,d'}W_{l,\widehat{d}}) = o(||\Gamma_n||_F^2), \tag{5}$$

$$\sum_{(i,d)} \mathbb{E}[|\boldsymbol{W}_{-i,d}\mathbb{E}[W_{i,d}\boldsymbol{W}_{-i,d}]|] = o(||\Gamma_n||_F). \tag{6}$$

Theorem 3.2 (Theorem 1 from Chandrasekhar et al. (2023))**.** *Denote a mean* 0 *triangular array of random vectors* $W_{1:n}^{(n)}$*. If a collection of affinity sets* $\mathcal{A}_{(i,d)}^{(n)}$ *satisfy the conditions of equations* (4)*,* (5) *and* (6)*. Then*

$$\Gamma_n^{-1/2} S_n \to_d N(0, I_p)$$

The authors illustrate some examples under which these conditions are sufficient for the affinity set conditions. Leveraging this central limit theorem, we can now proceed with our main theorem.

3.4. **Asymptotic Results.** We develop asymptotic results for two classes of estimators. In order to ensure the negligibly of the estimation rate of the graph model.

Before introducing the theorem, we introduce some additional notation. Define $g_i(\boldsymbol{Z};\beta) = m_i(\boldsymbol{Y};\boldsymbol{a},\boldsymbol{X},\beta,G^*,\theta_0)$ to be the moment function evaluated using the true generative model and correspondingly

$$g_n(\boldsymbol{Z};\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{Z};\beta).$$

We can also define the (normalized) random vector of the estimating function evaluated at the correct model parameters $\mathcal{E}_i = \frac{1}{n} g_i(\boldsymbol{Z};\beta_0)$. And lastly let $D_n(\boldsymbol{Z};\beta_0) = \nabla_\beta g_n(\boldsymbol{Z};\beta_0) \in \mathbb{R}^{p\times p}$ denote the gradient of the estimating equation $g_n(\boldsymbol{Z};\beta)$.

We next focus on step 3 of our procedure discussing the asymptotic distribution of this new estimating equation averaged over the graph draws

$$m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) = 0.$$

In order for the development of valid inference, we will need to estimate the nuisance graph model at a reasonably fast rate, such that we can ignore graph estimation component in the regression. We next present the theorem and further discuss the assumptions.

Assumption 3.3 (Regularity Conditions for Z-Estimation)**.** *Suppose the following conditions hold for all* n*.*

Consistency for a Z estimator

A1. $\mathbb{E}[g_n(\boldsymbol{Z};\beta)] = 0$ *for* $\beta = \beta_0$ *and for all* $\epsilon > 0$*,* $inf_{||\beta-\beta_0||>\epsilon}\mathbb{E}[g_n(\boldsymbol{Z};\beta)] > 0$

A2. $\sup_{\beta\in\mathcal{B}} \left| \left(\frac{\partial}{\partial\beta}\right)^l g_n(\boldsymbol{Z};\beta) - \left(\frac{\partial}{\partial\beta}\right)^l \mathbb{E}[g_n(\boldsymbol{Z};\beta)] \right| = o_P(1)$ *for* $l \in \{0,1,2\}$

Graph Model Regularity conditions

B1. $\widehat{\theta}$ *is an* $s(n)$*-consistent estimate of the graph parameters* $||\widehat{\theta} - \theta_0|| = o_P(s(n))$

B2. $\sup_{\beta\in\mathcal{B}} |m_n(\boldsymbol{Z};\beta,\theta) - m_n(\boldsymbol{Z};\beta,\theta')| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$ *where* $b_n(\boldsymbol{Z}) = O_P(1)$ *(that is,* $b_n(\boldsymbol{Z})$ *is stochastically bounded).*

Central Limit Theorem

C1. There exists a set of affinity sets $\mathcal{A}^n_{(i,d)}$ for the random vectors $\mathcal{E}_{1:n}$ such that (4), (5) and (6) are satisfied with a corresponding matrix Γ_n. Where $\sqrt{\lambda_{min}(\Gamma_n)} = r(n)$

Theorem 3.4 (Single Network Z-estimator Asymptotics)**.** *Suppose that assumption 3.3 hold and that $s(n) = o(r(n))$. Then:*

$$\Gamma_n^{-1/2} D(\beta_0)(\widehat{\beta} - \beta) \to_d N(0, I_p) \tag{7}$$

Where $\mathbb{E}[\nabla_\beta g_n(\boldsymbol{Z}; \beta)|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0]\Big|_{\beta=\beta_0} = D(\beta_0)$

The first set of assumptions is standard for the consistency of Z-estimators. In general, these can be derived through uniform law of large number conditions as in Andrews (1987); Newey and McFadden (1994). The second set includes a set of conditions under which the estimation of the graph model is negligible, meaning that the estimating functions should be smooth functions of the graph parameters. In Section 8.1 we discuss how $\widehat{\beta}$ is computed in practice. The last set of assumptions, C1 are sufficient for a central limit theorem as in Chandrasekhar et al. (2023). Under independent, or weakly dependent noise, we would have $r(n) \approx n^{-1/2}$ (up to a normalizing constant). Alternatively, one may consider indices where $\mathcal{E}_i$ are highly correlated, across groups. For example there may be k_n groups under which a high correlation exists and $k_n \in (1, n)$ of such blocks where $k_n \to \infty$. See section 4.4 in Chandrasekhar et al. (2023) for a diffusion example in a stochastic blockmodel). In the case where the correlation within blocks is approaching 1, and across blocks is approaching 0, we would observe $r(n) \approx k_n^{-1/2}$. If the number of blocks grows such that $\frac{\sqrt{\frac{\log(n)}{n}}}{k_n} \to 0$ then these asymptotic results would hold while using the latent space model. Of course, even if this does not hold, one may still recover a consistent estimator of the outcome model parameters, but the asymptotic distribution would instead be driven by the estimation of the graph model θ which would require a separate asymptotic approach, and one that may depend on the specific graph model used. We discuss the estimation rates in further detail in Section 3.6.3.

Under the special case of a linear model, $Y_i = \beta^T \widetilde{h}(S_i, V_i) + \epsilon_i$. We can construct an estimator by replacing $\widetilde{h}(S_i, V_i)$ by its conditional mean $\mathbb{E}[h(S_i, V_i)|\boldsymbol{X}, \boldsymbol{a}, \widehat{\theta}]$. The advantage to this method is that the implementation is simpler, whereas the Z-estimator may require further specification of the conditional distribution $P(\boldsymbol{Y}|\boldsymbol{V}, \boldsymbol{S})$, and require an iterative algorithm to fit as we illustrate in Section 8.1.

We next consider the asymptotic results for this estimator.

Theorem 3.5. *Let $\widetilde{H}_i(\theta) = \mathbb{E}[\widetilde{h}(S_i(G), V_i(G))|\boldsymbol{a}, \boldsymbol{X}, G^*; \theta]$. The OLS estimator uses the model averaged coefficients $\widetilde{H}_i(\theta)$ in place of the true unobserved coefficients $\widetilde{h}_i$. Let $\mathsf{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \widetilde{H}_i(\theta)\widetilde{H}_i^T(\theta)$. Given an estimate of the model parameters $\widehat{\theta}$, we define the*

$$\widehat{\beta}_{ols} = \mathsf{H}_n^{-1}(\widehat{\theta}) \frac{1}{n} \sum_{i=1}^n \widetilde{H}_i(\widehat{\theta}) Y_i$$

Let $u_i = (\widetilde{h}(S_i(G), V_i(G)) - \widetilde{H}_i(\theta_0))\beta_0 + \epsilon_i$. Suppose the following conditions hold for all n.
Model Regularity conditions

D1. $\widehat{\theta}$ is a $s(n)$-consistent estimate of the graph parameters $||\widehat{\theta} - \theta_0|| = o_P(s(n))$
D2. $|\mathsf{H}_n(\theta) - \mathsf{H}_n(\theta')| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$ where $b_n(\boldsymbol{Z}) = O_P(1)$ (that is, $b_n(\boldsymbol{Z})$ is stochastically bounded).
D3. $\max_i ||\widetilde{H}_i(\theta) - \widetilde{H}_i(\theta')|| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$
D4. $||H_i(\theta)|| \leq M < \infty$

D5. $\left|\frac{1}{n}\sum_{i=1}^{n}|u_i| - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|u_i|]\right| = o_P(1)$

Lastly, let Γ_n *denote a matrix that satisfies the following central limit theorem for the estimating function*

Central Limit Theorem

E1. For the array of random variables $\mathcal{G}_i = \frac{1}{n}H_i(\theta_0)u_i$*, there exists a set of affinity sets* $\mathcal{A}^n_{(i,d)}$ *such that (4), (5) and (6) are satisfied with a corresponding matrix* Γ_n*, where* $\sqrt{\lambda_{min}(\Gamma_n)} = r(n)$*.*

Then if $r(n) = o(s(n))$

$$\Gamma_n^{-1/2}\mathsf{H}_n(\widehat{\theta})(\widehat{\beta}_{ols} - \beta_0) \to_d N(0, I_p)$$

The same arguments follow in the OLS estimator with respect to the estimation rate of the graph model in comparison to the parameters of interest.

3.5. **Plug-in estimates of the Causal parameter.** For many problems, the parameter of interest is a causal query conditional on the complete graph G as described in Section 2.4. For example, one may care about the expected number of adoptions after seeding an individual in block k v.s. block k'. In this section, we illustrate how to construct an estimate of the causal parameter $\Psi(\boldsymbol{a}|G)$ using our conditional model estimation procedure.

Let $\Psi(\boldsymbol{a}|\theta_0) = \mathbb{E}[\Psi(\boldsymbol{a}|G)|\boldsymbol{a}, \boldsymbol{X}, \theta_0]$ be the average causal effect of policy $\boldsymbol{a}$ over all draws of the graph model θ_0. We will establish conditions under which these two quantities are close to one another.

Recall the true conditional mean function $\mathbb{E}[Y|S_i = s, V_i = v] = h_0(s, v)$. Under a correctly specified conditional model, $h_0(s, v) = h(s, v; \beta_0)$, and $\Psi(\boldsymbol{a}|\theta_0) = \Psi(\boldsymbol{a}|\beta_0, \theta_0)$ where

$$\Psi(\boldsymbol{a}|\beta, \theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(V_i, S_i; \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta]. \tag{8}$$

In order to estimate $\Psi(\boldsymbol{a}|G)$ we plug-in the estimates for the mean model and network model $\Psi(\boldsymbol{a}|\widehat{\beta}, \widehat{\theta})$. We next discuss the asymptotics of the plug-in estimate.

Lemma 3.6 (Inference for a plug-in causal parameter)**.** *Assume the conditions of 3.3. Further, assume:*

$$\sup_{\beta}\left|\mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}; G); \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta] - \mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}; G); \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta']\right| \leq b_i||\theta - \theta'|| \tag{9}$$

where $b_i \leq M < \infty$*. Denote*

$$Q_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}\mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}; G); \beta')|\boldsymbol{a}, \boldsymbol{X}, \theta_0]\Big|_{\beta'=\beta} \in \mathbb{R}^{1\times p}$$

and

$$\widetilde{\omega}_n := Q_n(\beta_0)D_n(\beta_0)\Gamma_n D_n(\beta_0)^T Q_n(\beta_0)^T.$$

If $s(n) = o(\sqrt{\widetilde{\omega}_n})$*. Then*

$$\widetilde{\omega}_n^{-1/2}(\Psi(\widehat{\beta}, \widehat{\theta}) - \Psi(\beta_0, \theta)) \to_d N(0, 1) \tag{10}$$

This lemma is essentially an application of the delta method, with the additional caveat that we estimate θ before the plug-in estimate. As before, this requires a fast estimate of the graph generative model parameter, but we add the slightly different assumption (eq. (9)) that the smoothness in the model class is over the conditional response models $\mathbb{E}[h(S_i, V_i; \beta)|\theta]$, rather than the estimating function $\widetilde{m}(Y, S, V|\beta, \theta)$.

Convergence of the causal parameter to the average over graphs

As we have previously discussed, we can only hope to estimate $\Psi(\boldsymbol{a}|\theta_0)$ as we do not have access to the full graph G. We next introduce a simple conditions under which the parameter $\Psi(\boldsymbol{a}|G)$ is close to its average over draws of the graph $G \sim \theta_0$, $\Psi(\boldsymbol{a}|\theta_0)$.

Assumption 3.7 (v_n-response dependence)**.** *For any graph draw G let $G'^{(ij)}$ denote the graph G with the ij entry swapped from 0 to 1 or vice versa. Let $c_{ij,n}$ denote the bounds of the differences such that*

$$\left|\frac{1}{n}\sum_{i=1}^{n} h_0(S_i(\boldsymbol{X},G),V_i(\boldsymbol{a},G)) - h_0(S_i(\boldsymbol{X},G'^{(ij)}),V_i(\boldsymbol{a},G'^{(ij)}))\right| \leq c_{ij,n} \tag{11}$$

And let $v_n^2 = \sum_{ij:i\neq j} c_{ij,n}^2$

Lemma 3.8. *Under Assumption 3.7*

$$\Psi(\boldsymbol{a}|G) - \Psi(\boldsymbol{a}|\theta_0) = O_P(v_n)$$

The proof is a one-line application of McDiarmid's inequality. Previous related work such as Breza et al. (2023) typically assume that such a quantity is consistent, however here we quantify the rate here. We next highlight an example;

Example 3.1 (Conditional Mean Function Example)**.** *We abbreviate $G = G$ and $G' = G'^{(kl)}$. Let $h_0(S_i(\boldsymbol{X},G),V_i(\boldsymbol{a},G)) = \beta_0 + \beta_1 a_i + \beta_2 X_i + \beta_3 \sum_{l\neq i} \frac{X_l G_{kl}}{n} + \beta_4 \sum_{l\neq k} \frac{a_l G_{il}}{n}$ denote a linear response function dependent on the density of connected neighbors. Suppose that the covariate values are bounded $|X_i| \leq M < \infty$. Then:*

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=1}^{n} h_0(S_i(\boldsymbol{X},G),V_i(\boldsymbol{a},G)) - h_0(S_i(\boldsymbol{X},G'),V_i(\boldsymbol{a},G'))\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^{n}\sum_{l\neq k} \beta_3 \frac{X_l(G_{kl}-G'_{kl})}{n} + \beta_4 \frac{a_l(G_{kl}-G'_{kl})}{n}\right| \\
&\leq \left|\frac{1}{n}\frac{X_j+X_i}{n} + \frac{|a_j|+|a_i|}{n}\right| \\
&\leq (2M+2)/n^2
\end{aligned}$$

Applying 3.8 illustrates that: $\Psi(\boldsymbol{a}|G) - \Psi(\boldsymbol{a}|\theta_0) = O_p(n^{-1})$. Hence in order to estimate the expected average outcome, all we need is a consistent estimate of the model parameters β_0.

3.6. **Network Model Estimation.** We next discuss the estimation of the generative model for the network while using a variety of data types. Using a generative model of the network has is seen in similar approaches such as in moderately dense networks using a dense graph model Breza et al. (2020, 2023). We first highlight several examples of how different datatypes can be used to estimate a stochastic blockmodel and followup with a discussion on other model classes.

3.6.1. *Estimation of the Stochastic Blockmodel Using Sampled Data.* We next illustrate that it is possible to estimate the stochastic blockmodel using a diverse set of partial and sampled network data types. In each case, $\mathbf{P}_{kk'}$ refer to the cross-block probabilities, while $k_i \in \{1,2,\ldots,K\}$ denote the node memberships. We consider *partial network data* to be any subset of the network data which can be used to generate an estimate of the generative model $\widehat{\theta}$.

Example 3.2 (Induced subgraph)**.** *We sample $m \leq n$ of nodes in the graph randomly, with at least one node from each of the K communities. Let G' be the sub-graph induced by these*

m nodes. Let N'_k denote the set of sampled nodes in community k, assumed to be positive for each k. Let

$$\widehat{\mathbf{P}}_{kk'} = \frac{1}{|N'_k||N'_{k'}|} \sum_{i \in N'_k} \sum_{j \in G'_{k'}} G'_{ij} \ .$$

Example 3.3 (Edges missing)**.** *Suppose that edges are missing according to some distribution. Let G' be the observed graph, and suppose that $P(G'_{ij} = 1|X_{ij} = x)$ is the probability of observing the edge G'_{ij}, given dyad-level covariates X and the edge G_{ij}. Suppose that we have a consistent estimator of this conditional response. Then,*

$$\widehat{\mathbf{P}}_{kk'} = \frac{1}{|N'_k||N'_{k'}|} \sum_{i \in N'_k} \sum_{j \in G'_{k'}} \frac{G'_{ij}}{\widehat{P}(G'_{ij} = 1|X_{ij})} \ .$$

Lemma 3.9 (Rates for induced subgraph and Edges Missing)**.** *Consider an estimate for a stochastic blockmodel cross probabilities based on either the induced subgraph or the edges missing example of $m \leq n$. Let $m_k = |N_k| = \rho_k m$ for some $\rho_k \in (0, 1)$. Then with probability at least $1 - \delta$*

$$|\widehat{\mathbf{P}}_{kk'} - \mathbf{P}_{kk'}| \leq \frac{1}{\rho_k \rho_{k'} m} \sqrt{\frac{\log(2/\delta)}{2}} \tag{12}$$

Further, suppose that $\sup_x |\widehat{P}(G_{ij} = 1|X_{ij} = x) - P(G_{ij} = 1|X_{ij} = x)| = o_P(m^{-1})$ with $P(G_{ij} = 1|X_{ij} = x) \geq \lambda > 0$. Then for large enough m, eq. (12) holds for the missing edges example as well.

Our next example is aggregated relational data (ARD). Suppose each individual has discrete trait vector $t_i \in \{0, 1\}^T$. Aggregated relational data consists of outcomes $X^*_{it} \in \mathbb{R}_{\geq 0}$ which denote responses from individual i to the questions of the type: "How many people that have trait t do you know?". (Breza et al., 2023) show that we can consistently estimate the connection probabilities between latent types in a stochastic blockmodel using Aggregated Relational Data using mutually exclusive traits. Latent Space mixture models can also be estimated as shown in McCormick and Zheng (2015). However, we present an improved version of the SBM estimator which allows for an non-mutually exclusive traits. In our next example, we illustrate a method to cluster the data based on the responses, followed by estimating the cross-cluster probabilities.

Example 3.4 (Aggregated Relational Data)**.** *Suppose we collect ARD about the K communities using a survey of T discrete traits where $T \geq K$. Let X^*_{it} denote the ARD survey counts, the number of connections of type t that node i has. Let n_t denote the total number of individuals of trait type t. Let N'_k denote the nodes in our sample in group k, and let n_k denote the number of nodes in the graph in group k.*

Suppose that we learn a clustering of the ARD traits according to the following procedure.

1. *Count the number of individuals with each trait n_t*
2. *Denote the normalized ARD responses $X^\dagger_{it} = X^*_{it}/n_t$.*
3. *Cluster the normalized ARD response vectors $\{X^\dagger_i\}_{i=1}^T$ into K groups using hierarchical agglomerative clustering into a set of clusters $\widehat{k}_i \in \{1, 2, \ldots, K\}$*

After we obtain a clustering, we can estimate the stochastic blockmodel. Let $\widehat{\omega}_{kt} = \widehat{N}_{kt}/N_t$ where N_{kt} are the number of traits in the estimated group k and with trait t, and N_t are the number of individuals with trait t, and $\omega_{kt} = N_{kt}/N_t$, the analogous population quantity.

We next define the probability matrix of observing a connection of group k with a trait t. $\widetilde{\mathbf{P}}_{kt} = \sum_{k'} \mathbf{P}_{kk'} \omega_{k't}$, where $\widetilde{\mathbf{P}}_{kt} = P(G_{ij} = 1|k_j = k, t_i = t)$. This relationship can be

expressed in a linear system $\widetilde{\mathbf{P}} = \Omega\mathbf{P}$ *where* $\Omega \in \mathbb{R}^{T\times K}$ *and* $\Omega_{kt} = \omega_{kt}$. *If* Ω *is of full column rank, then a unique solution will exist. Given an estimate of the latent communities, one can estimate* $\widehat{\Omega}$.

$$\widehat{\mathbf{P}}_{kk'} = \left(\widehat{\Omega}^\intercal\widehat{\Omega}\right)^{-1}\widehat{\Omega}^\intercal\widehat{\widetilde{\mathbf{P}}}$$

where

$$\widehat{\widetilde{\mathbf{P}}}_{kt} = \frac{1}{n_k n_t}\sum_{i\in\widehat{N}_k} X^*_{it}.$$

In general, one can symmmetrize $\widehat{\mathbf{P}}_{kk'}$ after the estimate to ensure the constraints of an undirected stochastic blockmodel are satisfied. Alternatively, once can also minimize the constrained least squares objective

$$\widehat{\boldsymbol{P}} = \underset{0\leq\boldsymbol{P}\leq 1:\boldsymbol{P}=\boldsymbol{P}^\intercal}{\arg\min}\sum_{i=1}^{n}\sum_{t=1}^{T}(\widetilde{X}_{it} - \sum_{k'}\widehat{\Omega}_{k't}P_{k',k_i})^2.$$

This is easily implemented using standard convex solvers such as `CVX` Fu et al. (2020).

Though Breza et al. (2023) develop a procedure for consistently estimating the stochastic blockmodel, we can extend their result and obtain a rate on the estimation of the model parameters in Lemma 3.10. We differentiate between the cross-group probabilities in which the clusters that are estimated $\mathbf{P}^{(\widehat{\boldsymbol{k}})}$ with the cross-group probabilities where they are known with $\mathbf{P}^{(\boldsymbol{k})}$

Lemma 3.10. *Suppose that we use the clustering strategy outlined in Example 3.4 to cluster the observations based on aggregated relational data. Let* $Z_k = (\widetilde{\mathbf{P}}_{k1}, \dots \widetilde{\mathbf{P}}_{kT})$ *and* $\widetilde{\mathbf{P}}_{kt} = P(G_{ij} = 1|k_i = k, t_j = t)$. *Assume also that* $\inf_{k,k'}||Z_k - Z_k||_2 > 0$ *and that* $T \geq K$ *where* T *is the number of discrete traits asked about and* K *is the true number of clusters.*

Let $\widehat{\boldsymbol{k}}$ *denote the estimated cluster memberships and let* $\widehat{\mathbf{P}}^{(\widehat{\mathbf{k}})}$ *be the corresponding estimate of the cross block probabilities. Let* $\Omega_{kt} = N_{kt}/N_t$ *denote the matrix which involves the fraction of the individuals in cluster* k *who also have trait* t, *and* $\widehat{\Omega}$ *the estimated counterpart based on membership clusters. Let* $C_\Omega = \lambda_{max}((\Omega^T\Omega)^{-1})$ *and* $\lambda_{max}(\cdot)$ *denotes the largest eigenvalue of a matrix. We also assume* Ω *is of full column rank.*

Then with probability at least $1 - \delta - \frac{1}{n}$

$$||\widehat{\mathbf{P}}^{(\widehat{\mathbf{k}})} - \mathbf{P}^{(\boldsymbol{k})}||_1 \leq C_\Omega\frac{KT}{n}\sqrt{\frac{\log(2/\delta)\log(KT)}{2}}$$

We contrast our result to the optimal estimation rate of a stochastic blockmodel in Gao et al. (2015) which is $\widetilde{O}_P(n^{-1/2})$, where $\widetilde{O}$ refers to the big O notation with suppressed log-factors. Though it appears initially to be faster than theirs, there is a subtle distinction from their problem and ours. This comes from the fact that our clustering problem is much easier than the one involved in the standard stochastic blockmodel. In our case, the clusters are related to node-level observed traits which adds additional information. As the network grows, the normalized ARD vector converges to its mean for each individual, therefore the clustering problem becomes progressively easier. Conditional on a correct classification, the effective sample size grows as n^2 since there are n possible independent connections for each individual in the graph.

Lastly, we discuss respondent driven sampling. In this setting, community membership can be defined based on a partition of the covariates, thus allowing for an observable trait in the graph, a similar strategy is adopted by Roch and Rohe (2018).

Example 3.5 (Respondent driven sampling)**.** *Let $i \in \{1, 2, \ldots, m\}$ denote the indices of a sample of individuals obtained through respondent driven sampling. An initial number of individuals are recruited as seeds, and subsequent individuals are recruited via referrals from the others in a population. Tran and Vo (2021) develop a consistent estimator for the model parameters of the stochastic blockmodel.*

Let $\widetilde{G}_m$ be the subgraph of G_n sampled from a set of nodes $\{1, 2, \ldots, m\}$. Let M_k denote the number of individuals in the subsample of type k and let $M^{\leftrightarrow}_{kk'}$ denote the number of connected individuals in the subgraph $\widetilde{G}_m$.

The cross-type probabilities can be estimated as follows:

$$\widehat{\boldsymbol{P}}_{kk'} = \begin{cases} \frac{M^{\leftrightarrow}_{kk'}}{M_k M_{k'}} & \text{When } k \neq k' \\ \frac{M^{\leftrightarrow}_{kk}}{M_k(M_k - 1)} & \text{otherwise} \end{cases}$$

Tran and Vo (2021) illustrate the consistency of these parameters[5]*, in particular $|\widehat{\boldsymbol{P}}_{kk'} - \boldsymbol{P}_{kk'}| = O_P(m^{-1})$*

3.6.2. *Misspecification of the Graph Model.* Here we give further justification for the choice of using a stochastic blockmodel as it serves as a good approximator of a general graphon class. Suppose θ_0 is not in the stochastic blockmodel class, but instead in a smooth graphon class we can bound the bias in the estimation of the relevant model parameters.

Suppose that the edges are generated under a true graphon model. A graphon is a function $\widetilde{g} : [0, 1]^2 \to [0, 1]$ which assigns the pairwise conditional on a sample of $[0, 1]$ random variables.

$$\eta_{ij} = \widetilde{g}(\xi_i, \xi_j)$$
$$\boldsymbol{\xi} \sim_{iid} P_{\boldsymbol{\xi}} \in [0, 1].$$

Let $\mathcal{H}_\alpha(M)$ denote a smooth graphon class defined via the α-M-Hölder class as follows. Let $\mathcal{D} = [0, 1]^2 \cap x \leq y$ denote the domain of (x, y). We define the norm $||\widetilde{g}||_{\mathcal{H}_\alpha}$ as:

$$||\widetilde{g}||_{\mathcal{H}_\alpha} = \max_{j+k \leq \lfloor \alpha \rfloor} \sup_{x,y \in \mathcal{D}} |\nabla_{jk} \widetilde{g}(x, y)|$$
$$+ \max_{j+k = \lfloor \alpha \rfloor} \sup_{(x,y) \neq (x'y') \in \mathcal{D}} \frac{\nabla_{jk} \widetilde{g}(x, y) - \nabla_{jk} \widetilde{g}(x', y')}{(|x - x'| + |y - y'|)^{\alpha - \lfloor \alpha \rfloor}}$$

The Hölder class corresponding to this norm is defined as

$$\mathcal{H}_\alpha(M) = \{||\widetilde{g}||_{\mathcal{H}_\alpha} \leq M : \widetilde{g}(x, y) = \widetilde{g}(y, x); 0 \leq \widetilde{g}(x, y) \leq 1\}.$$

Prior work has focused on the approximability of a stochastic blockmodel to any element of a smooth graphon class. In particular there will always be some assignment of block memberships such that we can bound the 2-norm probability deviation from the true model.

Lemma 3.11. *Suppose that θ_* corresponds to a true graphon model and θ_0 a corresponding approximating stochastic blockmodel satisfying the conditions of 7.1. Denote the population estimating function, as a function of the model parameters*

$$L_n(\beta, \theta) = \mathbb{E}[\widetilde{m}_n(\widetilde{\boldsymbol{Z}}; \beta) | \boldsymbol{a}, \boldsymbol{X}, \theta]$$

where $L_n(\beta_0, \eta_0) = 0$ defines the population parameter β_0 under the misspecified model θ_0, and let $L_n(\beta_*, \theta_*) = 0$ define the population solution β_* to the correctly specified graph model θ_*. Let η_0 and η_* be the pairwise edge probabilities corresponding to the models θ_0, θ_* respectively. Finally assume that:

F1. $\mathcal{B}$ is compact

[5] Theorem 4.2 in their paper

F2. $\sup_{\beta\in\mathcal{B}} |L_n(\beta,\eta) - L_n(\beta,\eta_*)| \leq L||\eta-\eta_*||_2/n$

F3. $\min_j \frac{\partial}{\partial\beta_j} L_n(\beta,\eta_*)\Big|_{\beta=\beta_*} = \lambda > 0$

Then the approximation error under the graph misspecification is bounded by the following rate:

$$||\beta_0 - \beta_*|| = O(\lambda^{-1}K^{-\alpha\wedge 1}) \tag{13}$$

In practice, for datatypes like ARD, since we do not directly select clusters to misrepresent the graph model, we cannot guarantee achieving this bound. However, this is a worst-case bound, and in fact, may be overly conservative to the bias that we observe in practice. This therefore suggests a possibility of future work that involves the sensitivity analysis of both the response function and the latent graph model, which is beyond the scope of this paper.

3.6.3. *Estimation of Additional Network Models.* Though we emphasise the estimation of the stochastic blockmodel, there are several other methods available for estimation of the network formation model. These include the beta model of Chatterjee and Diaconis (2011), in which the graph generation model consists of two model parameters ν_i, ν_i possibly altered through some additional dyadic covariates X^*_{ij}

$$P(G_{ij} = 1|\theta_0) = \widetilde{f}(\nu_i + \nu_j + \beta^T X^*_{ij})$$

where $\widetilde{f}$ is a link function. Alternatively one can consider the latent space model of Hoff et al. (2002) which include latent positions on some unobserved manifold $\mathcal{M}^p$.

$$P(G_{ij} = 1|\theta_0) = \widetilde{f}(\nu_i + \nu_j + d(Z_i, Z_j))$$

In each of these cases Breza et al. (2023) illustrate consistent estimation rates in the $||\widehat{\theta} - \theta_0||_\infty = \mathcal{O}_P\left(\sqrt{\frac{\log(n)}{n}}\right)$ with the use of aggregated relational data. Since this represents the coarsest datatype we expect similar rates to hold for subgraph sampling and respondent driven sampling. Though this rate is too slow for the to ignore the effect of the estimation of the graph model, in examples where one expect a high level of correlation among the outcomes it can be practical to use these methods.

4. Experimental Design

Thus far our focus has been on inference for the parameters of our model, given a particular treatment assignment $\boldsymbol{a}$. We next discuss methods of experimental design which can be used to reduce the variance of our estimators.

We consider saturation randomization experiments. The saturation randomized design partitions the dataset into $\widetilde{K}$ clusters of size n_k respectively, then assigns a fraction τ_k fraction of each of the clusters to the given treatment, for a total number of treated of $n_t = \sum_{k=1}^{\widetilde{K}} \tau_k n_k$[6]. In practice, due to budgeting constraints etc. one can let $\boldsymbol{\tau} \in \mathcal{T} \subset [0,1]^{\widetilde{K}}$ denote a set of constrained saturation levels. For instance, this could reflect resource constraints (such as a limited number of vouchers to give away in a vaccine trial).

[6]If one fits a stochastic blockmodel to the network model, these saturation clusters need not be the same as those in the experiment

4.1. **Bayesian Optimization of Asymptotic Regression Estimators.** Our goal is to optimize the asymptotic variance of a function the model parameter $\widehat{\beta}$ in Section 3. We will develop a Bayesian Optimization procedure to solve for the optimal cluster-level treatment probability vector $\boldsymbol{\tau} \in [0,1]^{\widetilde{K}}$. We highlight this by optimizing the variance of the estimates of linear contrasts of the parameters $\phi^T\beta$. If one uses the stochastic block model for the network model structure, one can define these treatment block to align with the model blocks, however it need not be the case.

Denote the variance of the target contrast parameter conditional on the treatment assignment $\boldsymbol{a}$: $\upsilon^{\phi}(\boldsymbol{a};\theta) = \mathrm{Var}(\phi^T\widehat{\beta}|\boldsymbol{a};\theta)$. Ideally, the goal is to find a treatment assignment a^* that minimizes the variance of the contrast:

$$\arg\min_{\boldsymbol{a}\in\{0,1\}^n} \upsilon^{\phi}(\boldsymbol{a};\theta).$$

Without further structure, this requires exponential time complexity (NP-hard), as it requires a search over 2^n possible discrete choices of treatment assignment. Under the restriction of a saturated randomized design, one can solve a simpler problem, namely, we wish to minimize the average variance over treatment assignments given a saturation level $\boldsymbol{\tau}$ in a much lower dimensional continuous parameter space. Let $P_{\boldsymbol{\tau}}$ denote the distribution of treatment assignments $\boldsymbol{a}$ under a saturation level $\boldsymbol{\tau}$.

$$\mathcal{V}(\boldsymbol{\tau};\theta) = \mathbb{E}_{\boldsymbol{a}\sim P_{\boldsymbol{\tau}}}[\upsilon^{\phi}(\boldsymbol{a};\theta_0)]$$

Our goal will then be to minimize $\upsilon(\boldsymbol{\tau};\theta)$ instead. We evaluate $\mathcal{V}$ using the procedure, for the OLS estimator outlined in Algorithm 1. In the appendix we include a general procedure for Z-estimators.

Algorithm 1 Saturation Randomized Design Variance.

Inputs: Variance structure $\mathrm{Var}[\boldsymbol{u}] = \Sigma$, Model estimate $\widehat{\theta}$.
Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L} \sim \widehat{\theta}|G^*$
Sample R treatments $\{\boldsymbol{a}_r\}_{r=1}^{R}$ according to the block saturation levels $\boldsymbol{\tau}$.
for $r \leftarrow 1$ **to** R **do**
    Compute the averaged features $\widehat{H}_{ir}(\boldsymbol{a}) = \frac{1}{L}\sum_{l=1}^{L} \widetilde{h}(S_i(\widehat{G}^{(l)})V_i(\boldsymbol{a}_r;\widehat{G}^{(l)}))$
    Compute the Hessian $\widehat{\mathsf{H}}_n(\boldsymbol{a}_r) = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}_{ir}(\boldsymbol{a})\widehat{H}_{ir}^T(\boldsymbol{a})$.
    Compute the design matrix $\widehat{H}_r^T(\boldsymbol{a}) \in \mathbb{R}^{n\times p}$ where each row is $\widehat{H}_{ir}(\boldsymbol{a})$.
    Compute the variance for a single draw of the treatment vector $\boldsymbol{a}_r$:

$$\upsilon^{\phi}(\boldsymbol{a}_r;\widehat{\theta}) = \phi^T\widehat{\mathsf{H}}_n^{-1}(\boldsymbol{a}_r)\widehat{H}_r^T(\boldsymbol{a})\Sigma\widehat{H}_r(\boldsymbol{a})\widehat{\mathsf{H}}_n^{-1}(\boldsymbol{a}_r)\phi$$

end for
Average over each of the draws $\mathcal{V}(\boldsymbol{\tau};\widehat{\theta}) = \sum_{r=1}^{R} \upsilon^{\phi}(\boldsymbol{a}_r;\widehat{\theta})$

In order to use Algorithm 1 in practice, one needs to make assumptions on the covariance matrix Σ. One could in principle also make assumptions on the correlation within the dense blocks of the network. The simple evaluation of the average variance $\mathcal{V}(\boldsymbol{\tau};\widehat{\theta})$ is computationally costly, and thus is a natural candidate for Bayesian Optimization.

We next give a brief overview of Bayesian optimization procedures, then proceed to an algorithm on how this can be implemented in practice.

Bayesian Optimization. Let $\mathcal{V}$ denote our objective function. We place a Gaussian process prior on a function $\mathcal{V}(\boldsymbol{\tau})$ on a set of pilot points $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{n_0}$. This prior is defined as

$$\mathcal{V}(\boldsymbol{\tau}_{1:n_0}) \sim N(\mu_0(\boldsymbol{\tau}_{1:n_0}, \Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})))$$

Where $\text{Cov}[\mathcal{V}(\boldsymbol{\tau}_i), \mathcal{V}(\boldsymbol{\tau}_j)] = \Sigma_0(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j)$ where Σ_0 is some kernel function. Typical choices include Màtern kernels or the Gaussian kernel. In the gaussian case we define $\Sigma_0(x, x') = \alpha_0 \exp\left(-||x - x'||^2\right)$ and $||x - x'||_2^2 = \sum_{r=1}^{p} \alpha_r (x_r - x'_r)^2$. Under this prior, we can define the posterior

$$\begin{aligned}
\mathcal{V}(\boldsymbol{\tau})|\mathcal{V}(\boldsymbol{\tau}_{1:n_0}) &\sim N(\mu_n(\boldsymbol{\tau}), \sigma_n^2(\boldsymbol{\tau})) \\
\mu_n(\boldsymbol{\tau}) &= \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}_{1:n_0})\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})^{-1}(\mathcal{V}(\boldsymbol{\tau}) - \mu_0(\boldsymbol{\tau}_{1:n_0})) + \mu_0(\boldsymbol{\tau}) \\
\sigma_n^2(\boldsymbol{\tau}) &= \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}) - \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}_{1:n_0})\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})^{-1}\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau})
\end{aligned}$$

The Bayesian optimization procedure can be evaluated as follows.

(1) Place a Gaussian process prior on $\mathcal{V}(\boldsymbol{\tau})$
(2) Use a pilot n_0 samples space-filling design to observe $\mathcal{V}$ at the design points.
(3) While $i \leq N$ for some maximum number of draws N
   (a) Update the posterior on $\mathcal{V}(\boldsymbol{\tau})$
   (b) Let $\boldsymbol{\tau}_i$ be the maximizer of the acquisition function $A(\boldsymbol{\tau})$, (e.g. mean or upper confidence bound)
   (c) Observe $\mathcal{V}(\boldsymbol{\tau}_i)$
   (d) n = n+1
(4) Return the point $\boldsymbol{\tau}$ with the smallest $\mathcal{V}(\boldsymbol{\tau})$

Standard methods involve the Upper Confidence Bound (UCB) method which finds the largest value of the posterior for a specified κ, of $\mu_n(\boldsymbol{\tau}) + \kappa\sigma_n(\boldsymbol{\tau})$ to be the acquisition function $A(\boldsymbol{\tau})$. We utilize the `R` package `rBayesianOptimization` which utilizes the `GPfit` package (R Core Team, 2021; Yan, 2021; MacDonald et al., 2015). Additionally, see Frazier (2018) for an more in depth survey of Bayesian Optimization procedures.

The quality of optimization over N iterations to the true minimizer will depend on the smoothness of $\mathcal{V}(\boldsymbol{\tau})$. Since it is possible that under some configurations, the variance can diverge to ∞ (for example, this will tend to happen as $\boldsymbol{\tau} \to 0$). A quick fix is to instead maximize the objective $\exp(-\mathcal{V}(\boldsymbol{\tau}))$. The smoothness of $\exp(-\mathcal{V}(\boldsymbol{\tau}))$ will determine the closeness of the maximizer after N iterations. We can suppose that $\exp(-\mathcal{V}(\boldsymbol{\tau})) \in \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space with corresponding kernel $\Sigma_0(x, x') \leq B < \infty$. The smoothness of this function will determine the rate of this approximation which is described in greater detail in Srinivas et al. (2009). For example, suppose that $\{\boldsymbol{\tau}\}_{m=1}^N$ is a sequence of draws from the UCB algorithm, let $\boldsymbol{\tau}^*$ denote the minimizer of $\mathcal{V}(\boldsymbol{\tau})$. Then if Σ_0 is the square exponential kernel $\exp(-\mathcal{V}(\boldsymbol{\tau})) \leq \frac{1}{N}\sum_{m=1}^{N} \exp(-\mathcal{V}(\boldsymbol{\tau}_m)) + O_P(\frac{B\sqrt{\log(N)^{K+1}} + \log(N)^{K+1}}{\sqrt{N}})$. Similar results can be derived for Matern and linear kernels in Srinivas et al. (2009). We next extend this procedure to account for uncertainty in the graph model.

4.1.1. *Variance Minimization With Model Uncertainty.* As an extension of our variance minimizing procedure, we can incorporate the uncertainty in our estimates of the model parameters. For instance, consider the following parametric bootstrap approach for estimating the model parameters of the stochastic blockmodel when using ARD.

Denote $\widehat{\theta} = (\{\widehat{Z}_i\}_{i=1}^n, \widehat{\boldsymbol{P}})$ the initial estimate of the model as computed from lemma 3.10. We can construct a sampling distribution of $\widehat{\theta}^{(b)}$ using the following procedure. Let X_{it}^* denote the ARD responses of the number of connections individual i has to someone of trait t and let $T_i \in \{0, 1\}^T$ denote the trait memberships of the corresponding individuals.

(1) Estimate $\widehat{\theta}$ from $\boldsymbol{X}^*$ using example 3.4
(2) For $b \in \{1, 2, \ldots, B\}$
   (a) Sample $G^{(b)} \sim \widehat{\theta}$

(b) Construct the ARD vector based on the resampled responses $X_{it}^{*(b)}$ using counts according to connections of $G^{(b)}$ to the nodes with corresponding traits $\{T_i\}_{i=1}^n$

(c) Estimate $\widehat{\theta}$ from $\boldsymbol{X}^{*(b)}$ using example 3.4

This approach can work for any procedures which can allow for a sampling distribution of the model parameters $\{\widehat{\theta}^{(b)}\}_{b=1}^B$. For example Baraff et al. (2016) considers a nonparametric bootstrap for respondent driven sampling.

In all such cases where $\widehat{\theta}$ is modeled with uncertainty, we apply algorithm 1 to each of the b draws. Since the model is equivalent under cluster permutations, we choose the permutation for each $\{\widehat{Z}_i^{(b)}\}_{i=1}^n$ which minimizes the classification error with respect to $\{\widehat{Z}_i\}_{i=1}^n$. This is implemented using standard software, for example in the `label.switching R` package Papastamoulis (2015). Thus far we have discussed assigning seeds or treatments from the perspective of designing more efficient experiments, however, in many applications, one may wish to select nodes which will maximize some outcome over the network, such as diffusion processes.

4.2. **A discussion on Seeding.** In addition to the regression-based diffusion models, we can also use our approach to define the optimal seeding strategies across different clusters. In general, finding the optimal seeds for diffusion is an NP-hard problem (Kempe et al., 2003). In settings where only the block information is available however, the stochastic blockmodel allows for a simpler configuration due to the node-exchangeability within a block. In our setting, since the exact network structure is unknown, we in we can obtain the optimal blocks to assign seeds, however when $K \ll n$ this additional structure greatly reduces the computational burden and we only need to identify the number of seeds to assign to each of the K clusters.

Beyond just diffusion, we can use this framework to simulate from the outcome model $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$. In some cases, this may be an assumed model for the process (i.e. as Beaman et al. (2021) does using complex contagion). In other cases, $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$ may be parameterized using some outcome model which is estimated, (i.e. we simulate from $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y; \widehat{\beta})$). We illustrate this in Algorithm 4.2.

Algorithm 2 Optimal Seeding With Partial Network Data

1: **Inputs:** Number of seeds N, Model estimate $\widehat{\theta}$, number of graph draws L.
2: Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^L \sim \widehat{\theta}|G^*$
3: **for** $\tau \in \mathcal{T}$ **do**
4: Sample L treatments $\{\boldsymbol{a}_l\}_{l=1}^L$ according to the block saturation levels $\boldsymbol{\tau}$.
5: Compute the outcomes $Y_i^{(l,\boldsymbol{a}_l)}$ according to the outcome model $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$.
6: Compute the average (and standard error) over draws of the network $\overline{Y}^{(\boldsymbol{\tau})} = \frac{1}{L}\sum_{l=1}^L Y_i^{(l,\boldsymbol{a}_l)}$
7: **end for**
8: Return saturation level $\boldsymbol{\tau}$ with the largest value of $\overline{Y}^{(\boldsymbol{\tau})}$.

The procedure illustrated in Algorithm 4.2 allows for the seeding using an outcome model in a variety of scenarios using partial network data. We present this result using partial network data, however, if one wishes to use many seeds, one can apply a Bayesian Optimization procedure to Algorithm 4.2 as well. We next illustrate these aspects of estimation, experimental design and seeding in several parts in the following section.

5. Data Analysis

In this section, we highlight the main aspects of this framework. First, we illustrate how this framework can be used to estimate causal effects using partial network data, following this we give examples of experimental design with network uncertainty. Lastly, we discuss seeding with and without estimating the outcome model.

In each of our examples, we use real network datasets. In some cases, we simulate outcomes to highlight the relevant aspect of our framework. When individual level covariates are available, we use those to construct ARD. When this is not available, we define traits by clustering the network using the Leiden algorithm Traag et al. (2019) implemented in the `igraph` Csardi and Nepusz (2006) and denote these cluster memberships the traits. A summary on which datasets used this procedure and which used real traits is illustrated in section 5. This allows us to control the number of traits used in the simulations. All analysis is conducted in `R` R Core Team (2021).

Network Dataset	Traits
Banerjee et al. (2013)	Leiden Cluster $K \in [4, 16]$
Banerjee et al. (2019)	Real Traits (Section 10.2)
Beaman et al. (2021)	Leiden Cluster K = 8

Table 1. Summary of synthetic traits vs. real traits in the simulation and real data analysis settings.

5.1. **Causal Effect Estimation.** In this first example, our goal is the estimation of a causal effect, specifically the global average treatment effect. In order to do so, we must estimate the conditional distribution $\mathbb{E}[Y_i|S_i, V_i]$. Of course, the choice of which are the correct confounders to condition on depend on the specific problem at hand. We consider the example from Ugander and Yin (2023) and generate a set of potential outcomes according to the following model

$$Y_i(\mathbf{0}) = \frac{d_i}{\overline{d}}\left(\alpha + bX_i + \sigma\epsilon_i\right)$$
$$Y_i(\boldsymbol{a}) = Y_i(\mathbf{0})\left(1 + \delta a_i + \gamma\frac{\sum_{j\in[n]} G_{ij}a_j}{d_i}\right)$$

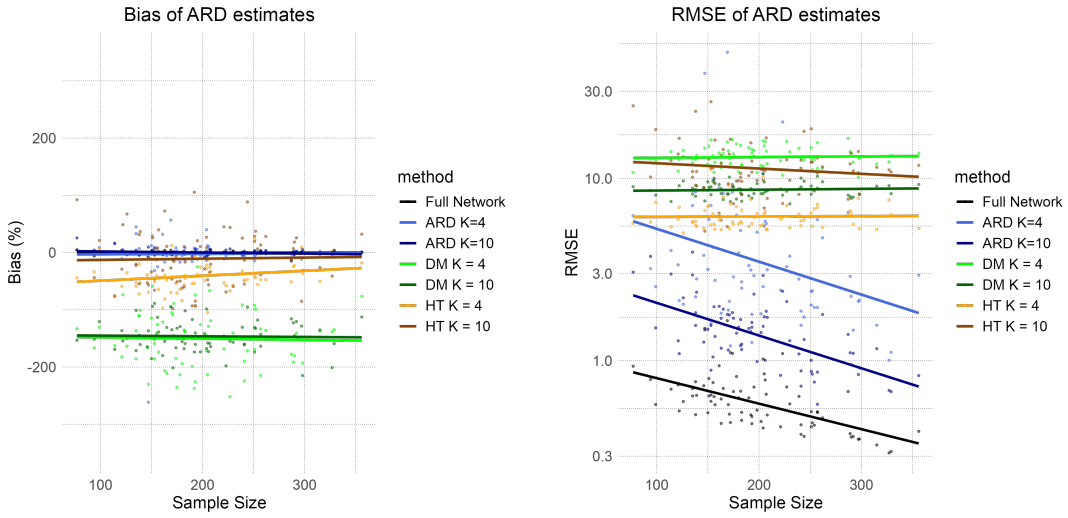
where $\epsilon_i \sim_{iid} N(0,1)$ is some independent noise, and X_i is a covariate that varies throughout the network, d_i is the degree of individual i and $\overline{d}$ is the average degree across the network. In our simulations $\alpha = 1$, $b = 1$, $\delta = 1$, $\sigma = 0.5$ and $\gamma = -0.5$. This model clearly falls into our framework where the potential outcomes are confounded by local graph statistics (degree) and node-level covariates. The exposure is the individual treatment in conjunction with the average treatment of neighbors, and the graph confounder include the degree ratio and node level covariates

$$f_V(\boldsymbol{a}; \varphi_i, G) = \left(a_i, \frac{\sum_{j\in[n]} G_{ij}a_j}{\overline{d}}\right)$$
$$f_S(\boldsymbol{X}; \vartheta_i) = \left(\frac{d_i}{\overline{d}}, X_i\right).$$

The global average treatment effect in this model is $\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{0})(\delta+\gamma) = \Psi(\boldsymbol{a} = 1|G) - \Psi(\boldsymbol{a} = 0|G)$. Though we see that this exact parameter value depends on the graph statistic through the degrees of each individual node, an application of lemma 3.8 would suggest that

$\Psi(\boldsymbol{a}=1|G)-\Psi(\boldsymbol{a}=0|G)-(\mathbb{E}[\Psi(\boldsymbol{a}=1|G)-\Psi(\boldsymbol{a}=0|G)|\theta_0])=O_P(n^{-1})$, for a densely growing graph, and thus be close to the average on any draw from the graph model θ_0.

We compare the graph cluster randomization of using a Horvitz-Thompson estimator Ugander et al. (2013) and a simple difference in means estimator under a cluster randomized design. For each of the clusters learned, we use a design where half of the clusters receive a treatment saturation of 0 and the other half receive a treatment fraction of 1. We increase the number of learned clusters from 4 to 16 though limit display to 4 and 10 in fig. 3 for clarity.



(a) Bias (b) RMSE

FIGURE 3. Comparison of GATE estimators. ARD denotes our method where the full graph is replaced using a model of the graph estimated using ARD. The Full Network method uses a regression approach. DM is the difference in means and HT is the Horvitz-Thompson estimator.

Figure 3 shows that the full data regression model performs the best, unsurprisingly as it utilizes additional information compared to the ARD approaches. The ARD version however, performs very effectively in terms of the bias (Figure 3(a)) and the RMSE (Figure 3(b)). Since the graphs we simulate are relatively dense and have few clusters, as the sample size gets large, the Horvitz-Thompson Estimator struggles as the probability that any node has no treated neighbors is exceedingly small, and therefore, the outcomes for many individuals are not used, even though the HT estimator has access to the whole network data. The difference in means estimator maintains a constant bias throughout, due to the lack of incorporation of the heterogeneous covariate information. Of course the regression approach using complete data is the most effective, however we can achieve reasonably close results using partial network data,

5.2. **Experimental Design.** We next highlight aspects of experimental design in two settings. With and without accounting for uncertainty in the estimation of the network model.

5.2.1. *A Generalized Hearing Model.* We first consider an information diffusion example based on the hearing model referenced in Section 2.2.2. This is based on an SIS model in which at each time step information is transmitted from one node to each of the neighbors with probability q_t. At each time step the previously infected nodes are susceptible again

the nodes infected in the last round will infect their neighbors with probability q_{t+1}. We repeat this for $T = 3$ rounds. Let N_i denote the total number of infections after the process. We then sample some binary response $P(Y_i = 1|N_i) = \text{logit}(\alpha_0 + \alpha_1 N_i)$ where α_0 and α_1.

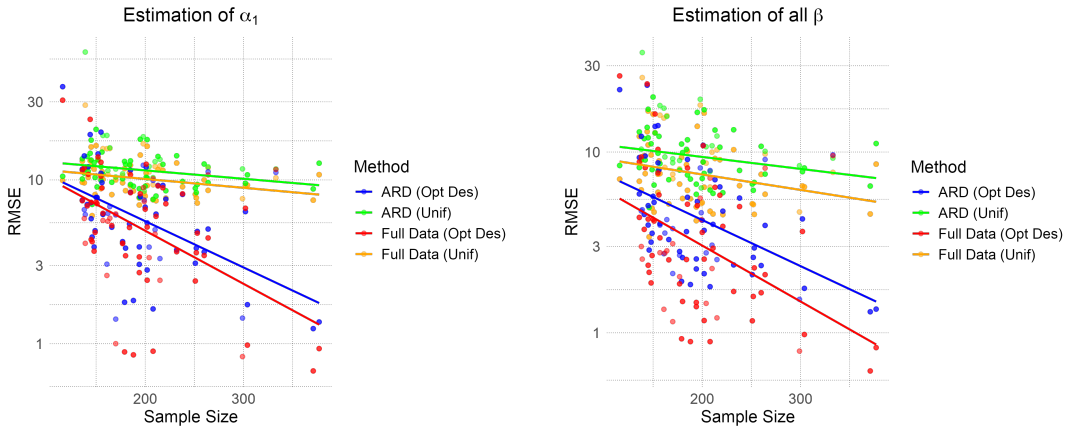
In this case, $\mathbb{E}[N_i|\boldsymbol{a}] = \sum_{t=0}^{3} \beta_t \boldsymbol{a}(G^t)_i$ where $\beta_t = \prod_{j=1}^{t} q_j$. Our goal is to estimate α_1, as well as the remaining coefficients in the model. In this case, $\beta = (0, 0.5, 0.05, 0.005)$. We estimate the coefficients in each of these cases letting $V_i = \mathbb{E}[N_i|\boldsymbol{a}]$ be the exposure mapping.

We then generate the outcomes according to the exposure received

$$\mathbb{E}[Y_i|S_i, V_i] = \Lambda(\alpha_0 + \alpha_1(\sum_{t=0}^{3} \beta_t (G^t)_i \boldsymbol{a}))$$

where $\Lambda(\cdot)$ is the logistic function.

In the actual dataset, seeds are assigned under a uniform randomization with 3 or 5 seeds per network. In this case, we compute the optimal seed allocations according to our procedure in Section 4[7], where no cluster can receive more than the total number of seeds in the actual experiment (3 or 5) depending on the network. We simulate this process 500 times over each of the villages in the dataset. We compare the estimates for α_1 as well as the total set of parameters in the model in fig. 4. We observe in Figure 4 and for estimating α_1 as well as the all of the model parameters, there is often much more to be gained by a more experimental design, than using the graph parameters directly.



(a) Estimation of α_1.

(b) Estimation of all model parameters.

FIGURE 4. Estimation of parameter α_1 and all model parameters β using the naive and optimized seeding. We observe that the potential gain found using a more efficient design is much greater than simply collecting more network data.

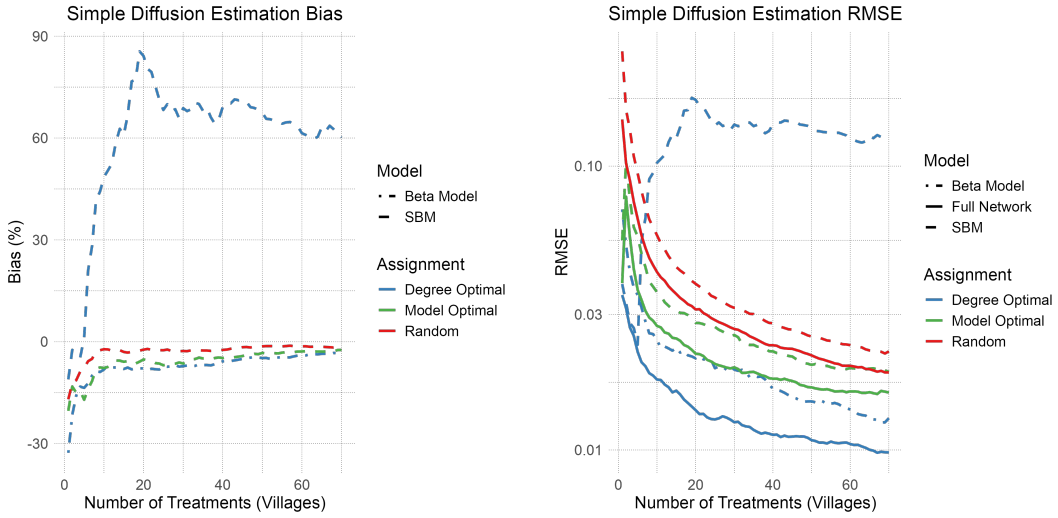
5.2.2. *Local Diffusion.* We next consider an example using a local diffusion process. We suppose that seed nodes are placed at time 0 and that outcomes are measured at time $T = 1$, allowing for diffusion to only take place to the immediate neighbors with a fixed probability q. In this case, for non-seed nodes the probability of infection is related to the total number of treated neighbors through the following link function. Under this model let $V_i \in \{0, 1\}$ denote the exposure as to whether one of their neighbors have received the treatment, i.e. $V_i = I(\sum_j G_{ij} a_j > 0)$. Then

$$\mathbb{E}[Y_i|V_i, S_i] = qV_i$$

[7]Our Bayesian optimization procedure which first randomly samples the target space 20 times and then uses 20 iterations to find a better saturation.

In this experiment, a single individual is seeded in each network. Our goal is to identify the best individuals in each of the network to seed and rank them by the expected variance of the estimator. We compare this to random seeding of individuals in the network as well as seeding by only the highest degree nodes. We use the networks constructed by the union of all connections of Banerjee et al. (2019). We construct estimates of the stochastic blockmodel as the partial data example using $K = 3$ in each case. We construct the traits using ARD responses based on number of connections with the following traits outlined in the appendix in Section 10.2. We also include an alternative where a beta-model (Chatterjee and Diaconis, 2011) is used in place of the SBM for the degree seeding where further details on estimation are included in Section 10.1. We then draw samples of the graph using the parametric bootstrap to obtain a resampled distribution of ARD $\{\boldsymbol{X}^{*(b)}\}_{b=1}^{B}$ for $B = 1000$. We identify the optimal treatment block for each parameter according to Section 4.1.1. We simulate 1000 draws of the draws in the diffusion process for each true, and plot the associated bias and RMSE of the seeding strategies in fig. 5 with a true diffusion parameter $q = 0.2$.

In the full data case, the optimal strategy would be to seed the highest degree node in each of the networks and measure whether each of their neighbors are infected at time $T = 1$. However, this poses a problem for the stochastic blockmodel as we are essentially picking an outlier to seed, which is different than a typical member of the block over draws of the process. This can be corrected for using a model which accounts for degree heterogeneity, in our case, the beta model. In our optimal seeding strategy, we find that the RMSE is lower in both the degree optimized strategy with the beta model, as well as the block optimized strategy with the SBM, than even the full data version with a completely randomized allocation, hence highlighting the role of the interplay of the model of the graph and the experimental design. This behavior is observed in Figure 5.



(a) Bias of Full and Partial Data Diffusion Parameter Estimates

(b) RMSE of Full and Partial Data Diffusion Parameter Estimates

FIGURE 5. RMSE and bias of estimating parameter q using random seeding, and the optimal seed for each village.

5.3. **Optimal Seeding.** We lastly consider the approach to optimal seeding. We consider two approaches, firstly, one where the outcome model is estimated and then used for optimal

treatment allocation. And secondly, one where the outcome model is assumed and used for optimal seeding, replicating the results of Beaman (2012). In contrast to the optimal design approach, rather than minimizing the variance of the estimator, we wish to maximize the outcome across networks.

5.3.1. *Estimated outcome model.* In this example, we consider a problem of optimal treatment assignment after the outcome model is estimated. In this example we suppose that there is some benefit $\beta_1 > 0$ to receiving a treatment, and some smaller benefit based on the fraction of the neighbors treated $0 < \beta_2 < \beta_1$. We wish to assign treatments in a way that will maximize the expected outcome $\Psi(\boldsymbol{a}|G)$ for each network.

$$Y_i = \beta_0 + \beta_1 a_i + \beta_2 q_1 + \epsilon_i$$

Where $q_i := \frac{1}{d_i}\sum_{j=1}^{n} G_{ij}a_j$ denotes the normalized number of treated neighbors. We simulate the data with $\beta_0 = 1$, $\beta_1 = 1$ and $\beta_2 = 1/2$ with $\sigma_i \sim N(0,1)$.[8]

We suppose that in each example there is only a budget for $B \in \{10, 20, 40, 80\}$ treatments for each of the villages. The goal is to maximize the overall expected outcome. We consider the following competing procedures. In this case, we suppose that we have a single pilot network where we can learn the model and the goal is to maximize the benefit on the remaining networks. We use the same gossip diffusion networks as in sections 5.2.2 and 5.2.1.

We compare the following seeding strategies.

(1) Random assignment to all individuals in the network
(2) Equal assignment amongst clusters.
(3) Assign treatments ordered by the highest degree of the nodes.
(4) Maximize the total expected outcome by maximizing $\max_{\boldsymbol{a};||a||_1 \leq B} \Psi(\boldsymbol{a}; \widehat{\beta}, \widehat{\theta})$

Let $\mathbb{E}[Y_i|\boldsymbol{a}] = \beta_0 + \beta_1 a_i + \beta_2(1-a_i)\sum_{k'=1}^{K} \widehat{P}_{\widehat{k}_i k'} n_{t,k}$ and let $n_{t,k} = \sum_{j:k_j=k} a_j$. Therefore, the objective function.

$$\Psi(\boldsymbol{a}|\beta,\theta) = \beta_0 + \frac{1}{n}\beta_1 \mathbf{1}^T \boldsymbol{n}_t + \frac{1}{n}\beta_2 \zeta^T \boldsymbol{n}_t$$

where $\zeta = \frac{1}{d_i}\sum_{i=1}^{n} \mathbf{P}_{k_i,\cdot}$ and $\boldsymbol{n}_t = (n_{t,1}, n_{t,2}, \ldots, n_{t,K})$. In general, given a conditional model, one may fine tune the optimization approach to the particular challenges of evaluating the optimal treatment allocation. We partition each network into 6 blocks.

We plot the expected average outcome under each of the treatment allocations for the remaining 68 networks after learning a model from the first pilot network. We repeat this for the total number of treatments $B \in \{10, 20, 40, 80\}$.

In Figure 6 we find that based on our method, we can achieve higher average outcomes than simple models based on the block positioning alone, emphasizing the importance of considering the potential outcome model when optimal targeting.

5.3.2. *Complex Contagion (Technology Adoption).* We lastly highlight our methodology as applied to the seeding problem in Beaman et al. (2021). In their work, they propose that diffusion of a pit-planting technology for farmers in Malawi is governed by a complex contagion process. In this setting the outcome model is pre-defined by a model of behaviour of the process $Y_i = f_Y(S_i, V_i, \boldsymbol{\epsilon}_Y)$

Under complex contagion, each individual has a threshold $\varsigma_i \sim N_{[0,\infty)}(\lambda, 0.1)$[9] for which infection is propagated at time t if at least ϵ_i of their neighbors are infected at time $t-1$,

[8] We choose this form of a response function since it will be simple to solve with an off the shelf mixed-integer programming approach using CVXR (Fu et al., 2020).

[9] We use 0.1 to emphasize the complex contagion part of the diffusion

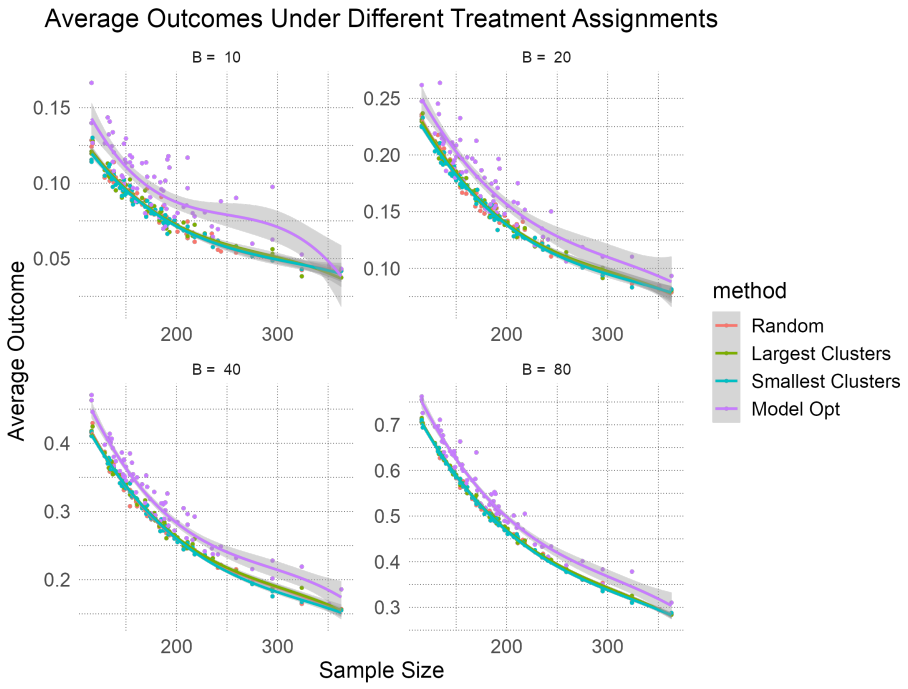
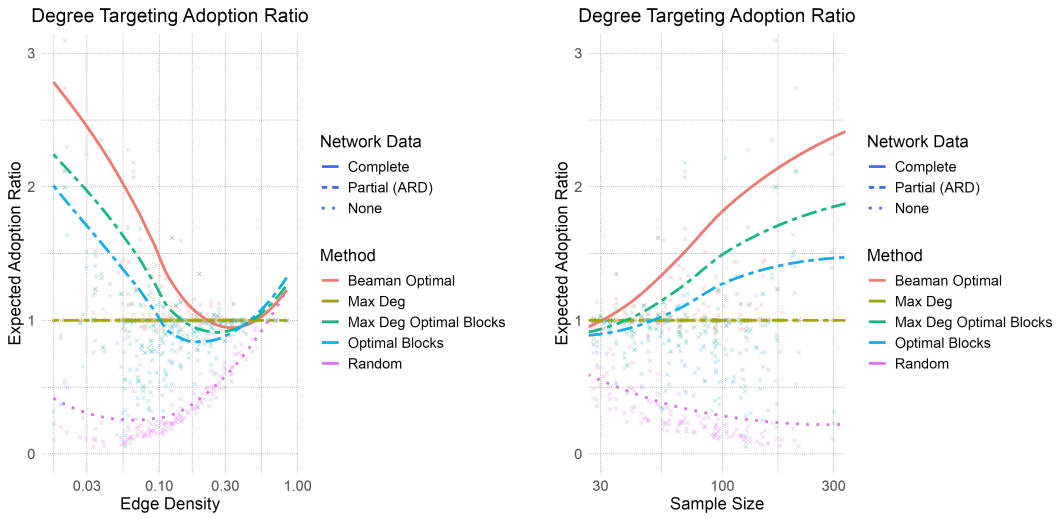


FIGURE 6. Our method, model based optimal treatment allocation (Model Opt) compared to random assignment and assignment to largest and smallest clusters respectively. The larger the values represent larger average outcomes in each of the networks. Curves are fit using cubic splines. The model based optimal design tends to give a higher value at each of the sample sizes at each treatment budget. For example, at a sample size of 150 and a treatment budget of 10, our methods leads to a 30% increase in the average outcome.

where $N_{[a,b]}$ refers to the truncated normal distribution. They simulate this process for $t = 3$ time periods in order to match their experimental time-period. To match the design in their study, we set $\lambda = 2$. We repeat this process 2000 times for $K = 8$ and identify the optimal clusters under this seeding outcome. We illustrate two methods of seeding. First simply randomly assigns a seed to the top two members of the optimal blocks, while the second seeds using the largest observed degree nodes within the optimal blocks.

We focus on the comparison to degree targeting a baseline comparison to seeding in practice. We observe the ratio of adopted nodes tends to be largest for our max degree within blocks method, and is comparable to the optimal seeds as chosen in their simulations. Further analysing these trends we observe in Figure 7 that in comparison to degree targeting, our methods tend to be more effective in villages that are large and tend to be more sparse. In highly dense networks and very small networks, the difference between the targeting strategies is minimal.

Method	Average Ratio of Adoption
Beaman Optimal	**1.50 (0.16)**
Max Degree	1.00 (-)
Optimal Blocks	1.13 (0.12)
Optimal Block (Max Degree)	**1.28 (0.13)**
Uniform Random	0.39 (0.03)



(a) Adoption ratio as a function of edge density. (b) Adoption ratio as a function of village size.

FIGURE 7. Comparison of degree seeding to other methods of partial and full data network seeding under complex contagion. In each case, using the model based targeting of the optimal blocks tends to be more effective than the degree seeding, and can further be improved by targeting the highest degree nodes within each of these blocks.

6. CONCLUSIONS

We present a framework for comprehensively identifying causal effects under interference by modeling potential outcomes based on a structural causal model, allowing for inference with partial network data. We develop estimators for the relevant model parameters, and plug-in parameters. We also establish and connect rates of estimating the stochastic blockmodel under various data types, and establish bounds on the estimated parameters due to misspecfication of the blockmodel in a smooth graphon class. By formulating structural causal models of the interference process, we can estimate a wide range of causal effects, in contrast to methods that typically focus on estimating a single causal effect like the GATE. This modelling allows us to tailor methods of experimental design to minimize the variance of relevant model parameters using our datatype. We demonstrate the effectiveness of our framework through several semi-synthetic problems, showcasing its performance comparable to methods using fully observed data and its superiority in certain scenarios.

Our approach suggests that when estimating effects under interference, a more direct modeling of the proposed interference mechanisms can yield numerous estimation benefits, including faster estimation rates and more obvious implications for seeding and experimental design.

Future work may go deeper into discussing potential outcome models useful for interference, along with the data requirements for learning these models, whether full, partial, or no network information is necessary. Additionally, investigating the interplay of design factors such as staggered rollouts or saturation randomized designs could enhance our understanding of effective experimental strategies, whether that include staggered rollouts, simple Bernoulli randomized designs, or cluster randomized designs.

Future work may consider nonparametric estimators for the graph model, such as low rank model used in Alidaee et al. (2020), however, this will likely require more carefully designed

estimators, and the consideration of semiparametric theory such as in Auerbach (2022). Additionally, other structured assumptions on the potential outcomes such as those made in Belloni et al. (2022) could potentially be estimated using similar methods and studied from a semiparametric perspective. However, the design problems would be distinct, and such a model may not allow for additional structure on the potential outcomes.

References

[1] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. The American Economic Review, 105(2):564–608, 2015.

[2] Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. Advances in Neural Information Processing Systems, 26, 2013.

[3] Hossein Alidaee, Eric Auerbach, and Michael P Leung. Recovering network structure from aggregated relational data using penalized regression. arXiv preprint arXiv:2001.06052, 2020.

[4] Zack W Almquist. Random errors in egocentric networks. Social networks, 34(4): 493–505, 2012.

[5] Attila Ambrus, Markus Mobius, and Adam Szeidl. Consumption risk-sharing in social networks. American Economic Review, 104(1):149–182, 2014.

[6] Jock R Anderson and Gershon Feder. Agricultural extension. Handbook of agricultural economics, 3:2343–2378, 2007.

[7] Donald WK Andrews. Consistency in nonlinear econometric models: A generic uniform law of large numbers. Econometrica: Journal of the Econometric Society, pages 1465–1471, 1987.

[8] Peter M. Aronow and Cyrus 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.

[9] Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. Journal of the American Statistical Association, 113(521):230–240, 2018.

[10] Eric Auerbach. Identification and estimation of a partially linear regression model using network data. Econometrica, 90(1):347–365, 2022.

[11] Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. arXiv preprint arXiv:2105.03810, 2021.

[12] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. Science, 341(6144):1236498, 2013.

[13] Abhijit Banerjee, Emily Breza, Arun G Chandrasekhar, and Benjamin Golub. When less is more: Experimental evidence on information delivery during india's demonetization. Technical report, National Bureau of Economic Research, 2018.

[14] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. The Review of Economic Studies, 2019.

[15] Aaron J Baraff, Tyler H McCormick, and Adrian E Raftery. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. Proceedings of the National Academy of Sciences, 113(51):14668–14673, 2016.

[16] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. Dynamical processes on complex networks. Cambridge university press, 2008.

[17] L.A. Beaman. Social networks and the dynamics of labour market outcomes: Evidence from refugees resettled in the u.s. Review of Economic Studies, 79 (1):128–161, 2012.

[18] Lori Beaman, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. Can network theory-based targeting increase technology adoption? American Economic Review, 111(6):1918–43, 2021.
[19] Alexandre Belloni, Fei Fang, and Alexander Volfovsky. Neighborhood adaptive estimators for causal inference under network interference. arXiv preprint arXiv:2212.03683, 2022.
[20] Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.
[21] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. Journal of econometrics, 150(1):41–55, 2009.
[22] Jennifer Brennan, Vahab Mirrokni, and Jean Pouget-Abadie. Cluster randomized designs for one-sided bipartite experiments. arXiv preprint arXiv:2210.16415, 2022.
[23] Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 2020.
[24] Emily Breza, Arun G Chandrasekhar, Shane Lubold, Tyler H McCormick, and Mengjie Pan. Consistently estimating network statistics using aggregated relational data. Proceedings of the National Academy of Sciences, 120(21):e2207185120, 2023.
[25] Giulia Cencetti, Diego Andrés Contreras, Marco Mancastroppa, and Alain Barrat. Distinguishing simple and complex contagion processes on networks. Physical Review Letters, 130(24):247401, 2023.
[26] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. American journal of Sociology, 113(3):702–734, 2007.
[27] A.G. Chandrasekhar, H. Larreguy, and J.P. Xandri. Testing models of social learning on networks: Evidence from a framed field experiment. mimeo: Stanford University, 2013.
[28] Arun Chandrasekhar and Randall Lewis. Econometrics of sampled networks. Unpublished manuscript, MIT.[422], 2011.
[29] Arun G Chandrasekhar, Matthew O Jackson, Tyler H McCormick, and Vydhourie Thiyageswaran. General covariance-based conditions for central limit theorems with dependent triangular arrays. arXiv preprint arXiv:2308.12506, 2023.
[30] S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. Arxiv preprint arXiv:1102.2650, 2011.
[31] Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. The Annals of Applied Probability, pages 1400–1435, 2011.
[32] Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. Journal of Causal Inference, 7(2), 2019.
[33] Mayleen Cortez, Matthew Eichhorn, and Christina Yu. Staggered rollout designs enable causal inference under interference without network knowledge. In Advances in Neural Information Processing Systems, 2022.
[34] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5):1–9, 2006.
[35] Giacomo De Giorgi, Michele Pellizzari, and Silvia Redaelli. Identification of social interactions through partially overlapping peer groups. American Economic Journal: Applied Economics, 2(2):241–275, 2010.
[36] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological), 39(1):1–22, 1977.
[37] Dennis Epple and Richard E Romano. Peer effects in education: A survey of the theory and evidence. In Handbook of social economics, volume 1, pages 1053–1163. Elsevier,

2011.

[38] Peter I Frazier. A tutorial on bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

[39] Linton C Freeman. Centered graphs and the structure of ego networks. Mathematical Social Sciences, 3(3):291–304, 1982.

[40] Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. Journal of Statistical Software, 94(14):1–34, 11 2020. ISSN 15487660. doi: 10.18637/jss.v094.i14. URL https://CRAN.R-project.

[41] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. The Annals of Statistics, pages 2624–2652, 2015.

[42] Sharad Goel and Matthew J Salganik. Respondent-driven sampling as markov chain monte carlo. Statistics in medicine, 28(17):2202–2229, 2009.

[43] Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences, 107(15):6743–6747, 2010.

[44] P. Goldsmith-Pinkham and G. Imbens. Social networks and the identification of peer effects. Journal of Business and Economic Statistics, 31:3:253–264, 2013.

[45] AKB Green, TH McCormick, and AE Raftery. Consistency for the tree bootstrap in respondent-driven sampling. Biometrika, 107(2):497–504, 2020.

[46] Viet Ha-Thuc, Avishek Dutta, Ren Mao, Matthew Wood, and Yunli Liu. A counterfactual framework for seller-side a/b testing on marketplaces. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2288–2296, 2020.

[47] Kathleen Mullan Harris, Carolyn Tucker Halpern, Eric A Whitsel, Jon M Hussey, Ley A Killeya-Jones, Joyce Tabor, and Sarah C Dean. Cohort profile: The national longitudinal study of adolescent to adult health (add health). International journal of epidemiology, 48(5):1415–1415k, 2019.

[48] Xiaoqi He and Kyungchul Song. Measuring Diffusion Over a Large Network. The Review of Economic Studies, page rdad115, 12 2023. ISSN 0034-6527. doi: 10.1093/restud/rdad115. URL https://doi.org/10.1093/restud/rdad115.

[49] Douglas D Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. Social problems, 44(2):174–199, 1997.

[50] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. https://doi.org/10.1198/016214502388618906, 97(460):1090–1098, 12 2002. ISSN 01621459. doi: 10.1198/016214502388618906. URL https://www.tandfonline.com/doi/abs/10.1198/016214502388618906.

[51] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. Journal of the American Statistical Association, 103(482):832–842, 2008.

[52] Kosuke Imai, Zhichao Jiang, and Anup Malani. Causal inference with interference and noncompliance in two-stage randomized experiments. Journal of the American Statistical Association, 116(534):632–644, 2021.

[53] Liwei Jing, Chengyi Qu, Hongmei Yu, Tong Wang, and Yuehua Cui. Estimating the sizes of populations at high risk for HIV: a comparison study. PloS ONE, 9(4):e95601, 2014.

[54] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. Management Science, 68(10):7069–7089, 2022.

[55] Matthew James Keeling and Pejman Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2008.

[56] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international

conference on Knowledge discovery and data mining, pages 137–146, 2003.
[57] Peter D Killworth, Christopher McCarty, H Russell Bernard, Gene Ann Shelley, and Eugene C Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation Review, 22(2):289–308, 1998.
[58] László Lovász and Balázs Szegedy. Limits of dense graph sequences. Journal of Combinatorial Theory, Series B, 96(6):933–957, 2006.
[59] Blake MacDonald, Pritam Ranjan, and Hugh Chipman. Gpfit: An r package for fitting a gaussian process model to deterministic simulator outputs. Journal of Statistical Software, 64:1–23, 2015.
[60] Anup Malani, Phoebe Holtzman, Kosuke Imai, Cynthia Kinnan, Morgen Miller, Shailender Swaminathan, Alessandra Voena, Bartosz Woda, and Gabriella Conti. Effect of health insurance in india: a randomized controlled trial. Technical report, National Bureau of Economic Research, 2021.
[61] Charles F Manski. Identification of endogenous social effects: The reflection problem. The review of economic studies, 60(3):531–542, 1993.
[62] Tyler H McCormick and Tian Zheng. Latent surface models for networks using aggregated relational data. Journal of the American Statistical Association, 110(512): 1684–1695, 2015.
[63] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. Handbook of econometrics, 4:2111–2245, 1994.
[64] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. Journal of the American Statistical Association, pages 1–15, 2022.
[65] Panagiotis Papastamoulis. label. switching: An r package for dealing with the label switching problem in mcmc outputs. arXiv preprint arXiv:1503.02271, 2015.
[66] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. Reviews of modern physics, 87 (3):925, 2015.
[67] Judea Pearl. Causality. Cambridge university press, 2009.
[68] Jean Pouget-Abadie, Vahab Mirrokni, David C Parkes, and Edoardo M Airoldi. Optimizing cluster-based randomized experiments under monotonicity. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2090–2099, 2018.
[69] Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Testing for arbitrary interference on experimentation platforms. Biometrika, 106(4):929–940, 2019.
[70] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.
[71] Sebastien Roch and Karl Rohe. Generalized least squares can overcome the critical threshold in respondent-driven sampling. Proceedings of the National Academy of Sciences, 115(41):10299–10304, 2018.
[72] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Detecting network effects: Randomizing over randomized experiments. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1027–1035, 2017.
[73] Fredrik Sävje, Peter Aronow, and Michael Hudgens. Average treatment effects in the presence of unknown interference. Annals of statistics, 49(2):673, 2021.
[74] O Scutelniciuc. Network scale-up method experiences: Republic of kazakhstan. Consultation on estimating population sizes through household surveys: Successes and

challenges (New York, NY), 2012.
[75] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995, 2009.
[76] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. Statistical methods in medical research, 21(1):55–75, 2012.
[77] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. Scientific reports, 9(1):5233, 2019.
[78] Viet Chi Tran and Thi Phuong Thuy Vo. Estimation of dense stochastic block models visited by random walks. Electronic Journal of Statistics, 15(2):5855–5887, 2021.
[79] Johan Ugander and Hao Yin. Randomized graph cluster randomization. Journal of Causal Inference, 11(1):20220014, 2023.
[80] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F128815:329–337, 8 2013. doi: 10.1145/2487575.2487695.
[81] A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, 10 1998. doi: 10.1017/cbo9780511802256. URL /core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.
[82] Mark J van der Laan. Causal inference for networks. 2012.
[83] Davide Viviano. Experimental design under network interference. arXiv preprint arXiv:2003.08421, 2020.
[84] CF Jeff Wu. On the convergence properties of the em algorithm. The Annals of statistics, pages 95–103, 1983.
[85] Yachen Yan. rBayesianOptimization, 2021. URL https://github.com/yanyachen/rBayesianOptimization.
[86] Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. Proceedings of the National Academy of Sciences, 119(44):e2208975119, 2022.

7. Supplement

7.1. **A discussion on the frameworks of interference.** We contrast the approaches of a fixed outcome approach as in (8) to a structural causal model approach. In the former approach, each individual has a distinct outcome under an exposure v, $Y_i(v)$. Though such an approach is robust for learning parameters such as average treatment effects $\frac{1}{n}\sum_{i=1}^{n} Y_i(v)$, the information in an individual i's potential outcome is completely distinct from individual j. This important details has important downstream implications.

Consider the simple contagion model from the example in Section 2.2.1 which takes place in a single time period ($T = 1$). Consider the nodes i, j in Figure 8 with seeded nodes in blue. Suppose that at time $T = 1$, that each neighbour of a treated node is infected with probability q. Since each one has only a single treated neighbor the distribution of the infection probability $P(Y_i = 1|\boldsymbol{a}, G)$ i and j are equivalent as their exposures are identical (i.e. they are each connected to a single seed node). However, in the finite sample framework the potential outcomes of any two nodes with a single treated neighbor can be arbitrarily different ($Y_i(v) \neq Y_j(v)$).

This nonparametric structure imposed on the potential outcomes later imposes restrictions on the degree of influence of others a node can have for estimation, thereby limiting this framework to examples with local dependencies (a phenomena also seen in (64)).

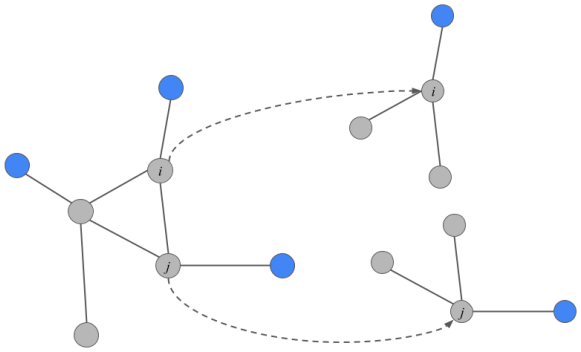


FIGURE 8. Equivalence of distribution of potential outcomes of nodes i and j are equivalent under this given treatment assignment as all of the rooted networks are equivalent.

7.2. **Proofs of paper theorems:**

7.2.1. *Proof of lemma 3.9.*

Proof. The proof is straightforward application of Hoeffding's inequality. Given an m node subsample of the full graph, and given their known types. Since $\widehat{\mathbf{P}}_{kk'} = \frac{1}{\rho_{kk'}m}\sum_{i,j} G_{ij}I(k_i = k, k_j = k')$, then the final result is a direct application of Hoeffding's inequality.

For the missing data case, we can plug-in the estimate of the edge sampling $P(G_{ij} = 1|X_{ij} = x)$ in order to correct for the missingness of the edges. If $\sup_x |\widehat{P}(G_{ij} = 1|X_{ij} = x) - P(G_{ij} = 1|X_{ij} = x)| = o_P(m^{-1})$ then the estimation of the propensity is negligible and we can correct for the missingness of edges. □

7.2.2. *Proof of lemma 3.10.*

Proof. Under the stochastic blockmodel assumption, the true latent traits are some discrete type $k_i \in \{1, 2, \ldots, K\}$. Then the mean connection probability Z_{ck} is simply a mixture over the connection probabilities, weighted by $P(k_j = k'|t_j = t)$. Let N_k denote the set of individuals with group k membership. Furthermore, let $n = |N_k|$. Denote analogous quantities for the trait memberships N_t as well as the intersection of k and t by N_{kt}. When we have a correct clustering. Denote $\widehat{P}_{kt} = \frac{1}{n_k}\sum_{i\in N_k}\frac{1}{n}\widetilde{Y}_{it}$. Assuming independent samples conditional on the graph clusters, let $P_{kt} = \frac{1}{n_t}\sum_{k'\in[K]}\sum_{i\in N_{tk'}} P_{kk'}$ denote the mean probability of connection averaged over the clusters conditional on their latent traits. Let $\omega_{kt} = \frac{n_{kt}}{n_t}$.

We can express $\widetilde{P}_{kt} = P(G_{ij} = 1|k_i = k, t_j = t)$ as a weighted sum of the connection probabilities from the constituent distributions. If the true clusters are known, then these

proportions ω_{kt} are known exactly from the data. Then

$$\begin{aligned}
\widetilde{P}_{kt} &= P(G_{ij} = 1 | k_i = k, t_j = t) \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1 | k_i = k, k_j = k', t_j = t) P(k_j = k' | t_j = t) \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1 | k_i = k, k_j = k', t_j = t) \omega_{k't} \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1 | k_i = k, k_j = k') \omega_{k't} \\
&= \sum_{k'=1}^{K} P_{kk'} \omega_{k't}
\end{aligned}$$

Expressing this relationship over the whole set of matrices, we have:

$$\widetilde{P} = \Omega P$$

Where $\Omega_{tk} = \frac{n_{tk}}{n_k}$,

We can solve this system as long as the columns of Ω are linearly independent. Therefore:

$$P = (\Omega^T \Omega)^{-1} \Omega^T \widetilde{P}$$

We next bound the estimation error in Frobenius norm of the true cross-cluster probabilities

$$\begin{aligned}
||\widehat{P} - P||_F &= ||(\Omega^T \Omega)^{-1} \Omega^T (\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq ||(\Omega^T \Omega)^{-1} \Omega^T||_F ||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq \sqrt{||(\Omega^T \Omega)^{-1} \Omega^T||_F^2} ||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq \sqrt{\mathrm{Tr}\left((\Omega^T \Omega)^{-1} \Omega^T \Omega (\Omega^T \Omega)^{-1}\right)} ||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&= \sqrt{\mathrm{Tr}\left((\Omega^T \Omega)^{-1}\right)} ||(\widehat{\widetilde{P}} - \widetilde{P})||_F
\end{aligned}$$

Since we assume that Ω's column's are linearly independent, then $\Omega^T \Omega$ is invertible. Therefore, what remains is bounding the Frobenius norm of $||(\widehat{\widetilde{P}} - \widetilde{P})||_F$.

For each element, let

$$\begin{aligned}
\widehat{\widetilde{P}}_{tk} &= \frac{1}{n_k n_t} \sum_{i \in N_k} \widetilde{Y}_{ik} \\
&= \frac{1}{n_k n_t} \sum_{i \in N_k} \sum_{j \in N_t} G_{ij}
\end{aligned}$$

Therefore, applying Hoeffding's inequality

$$P(|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq \epsilon) \leq 2 \exp\left(-2\epsilon^2 n_k n_t\right)$$

Letting $\rho_k = \frac{n_k}{n}, \rho_t = \frac{n_t}{n}$, then

$$P(|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq \epsilon) \leq 2 \exp\left(-2\epsilon^2 \rho_k \rho_t n^2\right)$$

Therefore, by a union bound,

$$P(\max_{k,t}|\widehat{\widetilde{P}}_{tk}-\widetilde{P}_{tk}|\geq\epsilon)\leq 2KT\exp\left(-2\epsilon^2\rho_k\rho_t n^2\right)$$
$$\implies P(\sum_{k,t}|\widehat{\widetilde{P}}_{tk}-\widetilde{P}_{tk}|\geq KT\epsilon)\leq 2KT\exp\left(-2(KT)^2\epsilon^2\rho_k\rho_t n^2\right)$$

Therefore,

$$||\widehat{\widetilde{P}}_{tk}-\widetilde{P}_{tk}||_1=\mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

Hence

$$||\widehat{P}-P||_2=\mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

Lastly, we show that as n grows, the probability of achieving a correct clustering of the true block memberships approaches 1. Recall that $n_t=\rho_t n$, and let $\underline{\rho_T}=\min_t\rho_t$. By Hoeffding's inequality: $P(||X_i^\dagger-Z_{k_i}||>\epsilon_n)\leq 2\exp(-2\epsilon_n^2 n/\underline{\rho_T})$. Taking a union bound over all response vectors, $P(\max_i||X_i^\dagger-Z_{k_i}||>\epsilon_n)\leq 2n2\exp(-2\epsilon_n^2 n/)\to 0$ for all $\epsilon_n=o(\sqrt{\log(n)/n})$.

Therefore, as n grows, the normalized response vectors in each cluster become well separated, and once $\epsilon_n<\min||Z_k-Z_{k'}||/2$, then all clusters will be well separated and naively hierarchical agglomerative clustering will consistently group the blocks together for K clusters. Therefore for example, if we let $\epsilon_n=\log(n)n^{-1/2}$,then $P(\max_i\{\widehat{k}\neq k\}=O(\frac{1}{n}))$. Of course the labels learned are only consistent up to permutation. We exploit the fact that as referred to in (24), the clustering problem gets easier as the sample size grows. Let $\mathcal{E}$ be the event that $\widehat{k}_i=k_i$ up to permutation for all $i\in\{1,2,\ldots,n\}$, i.e. $P(\max_i|\widehat{k}_i=k_i|>0)=1-P(\mathcal{E})\leq\frac{1}{n}$. Since the estimators are not necessarily independent of the event of perfect classification.

$$\begin{aligned}
P(||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||>\epsilon)&=P(||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||>\epsilon|\mathcal{E})P(\mathcal{E})+P(||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||>\epsilon|\mathcal{E}^c)P(\mathcal{E}^c)\\
&\leq P(||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||>\epsilon,\mathcal{E})+P(\mathcal{E}^c)\\
&\leq P(||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||>\epsilon,\mathcal{E})+\frac{1}{n}\\
&=P(||\widehat{P}_{\boldsymbol{k}}-P_{\boldsymbol{k}}||>\epsilon,\mathcal{E})+\frac{1}{n}\text{ Since }\mathcal{E}\text{ indicates the correct classification}\\
&\leq P(||\widehat{P}_{\boldsymbol{k}}-P_{\boldsymbol{k}}||>\epsilon)+\frac{1}{n}\\
&\leq\sqrt{\text{Tr}\left((\Omega^T\Omega)^{-1}\right)}2KT\exp\left(-2(KT)^2\epsilon^2\rho_k\rho_t n^2\right)+\frac{1}{n}
\end{aligned}$$

Therefore

$$||\widehat{P}_{\widehat{\boldsymbol{k}}}-P_{\boldsymbol{k}}||=\mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

□

7.2.3. *Proof of theorem 3.4.*

Proof. We emphasise that in general, the outcomes $\boldsymbol{Y}$ may be dependent, and this is reflected through correlation in the estimating functions (or the residuals in the case of OLS). We will partition the proof into two sections. First, we will prove the consistency of the estimator $\widehat{\beta}$ and secondly, we will prove the central limit theorem.

Consistency: The following result hinges on a typical consistency proof for the M or Z estimators using a structure similar to those found in Chapter 5 of (81). First, we denote that:

$$\begin{aligned} m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) - g_n(\boldsymbol{Z};\widehat{\beta}) &= m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) - m_n(\boldsymbol{Z};\widehat{\beta},\theta_0) \\ &\leq b_n(\boldsymbol{Z})||\widehat{\theta} - \theta_0|| \\ &= O_P(1)o_P(s(n)) \\ &= o_P(s(n)) \end{aligned}$$

Next, we can see that, based on this expansion,

$$\begin{aligned} & m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) = 0 \\ \implies & 0 = (m_n(\boldsymbol{Z},\widehat{\beta},\widehat{\theta}) - g_n(\boldsymbol{Z},\widehat{\beta})) + g_n(\boldsymbol{Z};\widehat{\beta}) \\ & = o_P(s(n)) + g_n(\boldsymbol{Z};\widehat{\beta}) \text{ By A2} \end{aligned}$$

At this point, we can now treat this as a standard Z-estimation problem. Therefore, by A2 and A1, then $\widehat{\beta}$ is a solution to the estimating function g and is therefore consistent by an application of Theorem 5.9 of (81).

Asymptotic Normality: We illustrate asymptotic normality through a Taylor series expansion argument. As we saw in the consistency part of the proof

$$g_n(\boldsymbol{Z};\widehat{\beta}) = o_P(s(n))$$

For brevity in notation, we suppress the dependence on $\boldsymbol{Z}$, which is implicit for functions, with the subscript n. Using a Taylor expansion around β_0, and let $\widetilde{\beta}_j \in [\beta_{0,j}, \widehat{\beta}_j]$ for $\beta_{0,j} \leq \widehat{\beta}_j$ and $\widetilde{\beta}_j \in [\widehat{\beta}_j, \beta_{0,j}]$ otherwise.

$$\begin{aligned} g_n(\widehat{\beta}) &= g_n(\beta_0) + D_n(\boldsymbol{Z};\beta_0)(\widehat{\beta} - \beta_0) + \sum_{jk} \frac{\partial^2}{\partial\beta_j\beta_k} g_n(\boldsymbol{Z};\widetilde{\beta})(\widehat{\beta}_j - \beta_{0,j})(\widehat{\beta}_k - \beta_{0,k}) \\ &= g_n(\beta_0) + D_n(\boldsymbol{Z};\beta_0)(\widehat{\beta} - \beta_0) + o_P(s(n) + ||\widehat{\beta} - \beta_0||) \end{aligned}$$

by the application of the consistency and A2. Therefore, we focus on main terms. By Assumption C1.

Therefore:

$$\Gamma_n^{-1/2} D_n(\boldsymbol{Z};\beta_0)(\widehat{\beta} - \beta_0) = \Gamma_n^{-1/2} g_n(\beta_0) + o_p(\frac{s(n)}{r(n)})$$

Noting that $D_n(\beta_0) - D(\beta_0) = o_P(1)$, by an application of Slutsky's lemma:

$$\Gamma_n^{-1/2} D(\beta_0)(\widehat{\beta} - \beta_0) \to_d N(0, I_p)$$

and therefore, the proof is complete. □

7.3. **Proof of theorem 3.5.**

Proof. We first we expand the form of the OLS estimator.

$$\begin{aligned}
\widehat{\beta}_{ols} &= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})Y_i \\
&= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\theta_0)\beta_0 + u_i) \\
&= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\widehat{\theta})\beta_0 + (\widetilde{H}_i^T(\theta_0) - \widetilde{H}_i^T(\widehat{\theta}))\beta_0 + u_i) \\
&= \beta_0 + \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\theta_0) - \widetilde{H}_i^T(\widehat{\theta}))\beta_0}_{(A)} \\
&+ \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0))u_i}_{(B)} + \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\theta_0)u_i}_{(C)}
\end{aligned}$$

We next bound terms (A) and (B) after which, the asymptotic distribution of (C) will be apparent.

We note that the Hessian evaluated at the true model parameters can be evaluated $\mathsf{H}_n(\widehat{\theta}) = \mathsf{H}_n(\theta_0) + o_P(s(n))$ by assumptions item D2 and item D1. By the continuous mapping theorem $\mathsf{H}_n(\widehat{\theta}) = \mathsf{H}_n(\theta_0) + o_P(s(n))$. We see that $(A) = o_P(s(n))$ by item D4, item D2 and item D1. Next, by the stochastic boundedness of the error item D5 and applying Hölder's inequality.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0)) &\leq (\frac{1}{n}\sum_{i=1}^{n}|u_i|)\max_i||\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0)|| \\
&= o_P(s(n))
\end{aligned}$$

Therefore:

$$\Gamma_n^{-1/2}\mathsf{H}_n(\widehat{\theta})(\widehat{\beta}_{ols} - \beta_0) = \Gamma_n^{-1/2}\sum_{i=1}^{n}\widetilde{H}_i(\theta_0)u_i + o_P\left(\frac{s(n)}{r(n)}\right) \to_d N(0, I_p)$$

by item E1 and Slutsky's Lemma, completing the proof. □

7.3.1. *Proof of lemma 3.6.*

Proof. The proof follows from an application of the delta method, with the additional caveat that we must account for the estimation of the model parameters θ_0 In this case:

$$\begin{aligned}
|\Psi(\widehat{\beta}, \widehat{\theta}) - \Psi(\widehat{\beta}, \theta_0)| &\leq \frac{1}{n}\sum_{i=1}^{n} b_i||\widehat{\theta} - \theta_0|| \\
&= o_P(s(n))
\end{aligned}$$

The remainder of the proof follows from a simple application of the delta method using the plug-in estimator $\Psi(\widehat{\beta}, \theta_0)$. See Theorem 3.1 of (81). □

7.3.2. *Proof of lemma 3.11.* We first include a useful lemma for bounding the approximation of the error of the graphon model.

Lemma 7.1 (Lemma 2.1 of (41))**.** *Denote $k_i \in \{1, 2, \ldots, K\}$ are the block memberships of a stochastic-blockmodel with average connection probabilities across blocks $\overline{\eta}_{ij} = P_{k,k'} = \frac{1}{n_k n_{k'}} \sum_{i,j:k_i=k,k_j=Z_i} \sum_{l:Z_l=Z_j} \eta_{kl}$. If the true graphon $g \in \mathcal{H}_\alpha(M)$, then, there exists some membership vector $\boldsymbol{k}$ and corresponding average across block probabilities P_0 such that:*

$$\frac{1}{n^2} \sum_{ij} (\eta_{ij} - \overline{\eta}_{ij})^2 \leq M^2 \left(\frac{1}{K^2} \right)^{\alpha \wedge 1}$$

We now proceed with a the proof of the lemma.

Proof. We firstly use a Taylor expansion of $L_n(\beta_0, \eta_*)$ where $\widetilde{\beta}$ is an element-wise intermediate value of β and $\widetilde{\beta}$

$$\begin{aligned} L_n(\beta_0, \eta_*) = L_n(\beta_*, \eta_*) + \frac{\partial}{\partial \beta} L_n(\beta, \eta_*) \Big|_{\beta=\beta_*} (\beta_0 - \beta_*) \\ + \sum_{jk} \frac{\partial^2}{\partial \beta_j \partial \beta_k} L_n(\widetilde{\beta}, \eta_*)(\beta_{0j} - \beta_{*j})(\beta_{0k} - \beta_{*k}) \end{aligned}$$

$$\frac{\partial}{\partial \beta} L_n(\beta, \eta_*) \Big|_{\beta=\beta_*} (\beta_0 - \beta_*) = -L_n(\beta_0, \eta_*) + \sum_{jk} \frac{\partial^2}{\partial \beta_j \partial \beta_k} L_n(\widetilde{\beta}, \eta_*)(\beta_{0j} - \beta_{*j})(\beta_{0k} - \beta_{*k})$$

Since we assume $L_n(\beta, \eta_*)$ is twice continuously differentiable in β, and $\mathcal{B}$ is compact, then $\frac{\partial^2}{\partial \beta_j \partial \beta_k} L_n(\widetilde{\beta}, \eta_*)$ is bounded. Therefore,

$$||\beta_0 - \beta_*||_2 \leq \frac{|L_n(\beta_0, \eta_*)|}{\lambda \sqrt{p}} + O(||\beta_0 - \beta_*||_2^2)$$

Lastly, by our continuity assumptions, $|L_n(\beta_0, \eta_*)| \leq L||\eta_0 - \eta_*||_2/n \leq LMK^{-(\alpha \wedge 1)}$. After applying this, our proof is complete.

□

8. Additional Methodological details

8.1. **An EM algorithm for Logistic Regression.** Here we elaborate on the computation of a Z estimator. In general, an estimator may require specific implementation, we provide an illustrative example with logistic regression. Recall the characterization of the average estimating function $m_i(Y_i, \boldsymbol{a}, \boldsymbol{X}; \beta, \theta) = \mathbb{E}[\widetilde{m}(Y_i, S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta) | \boldsymbol{Y}, \boldsymbol{a}, \boldsymbol{X}; \theta]$. Under this model, $P(Y_i = 1 | S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G)) = \Lambda(\widetilde{h}(S_i, V_i)^T \beta)$.

In order to compute the new estimating function, we need to be able to consider the distribution of the graph, conditional on the observed outcome Y_i. Specifically.

$$\begin{aligned} P(G | Y_i, \boldsymbol{a}, \boldsymbol{X}, \beta, \theta) &= \frac{P(Y_i | G, \boldsymbol{a}, \boldsymbol{X}; \beta) P(G | \boldsymbol{a}, \boldsymbol{X}, \theta)}{P(Y_i | \boldsymbol{a}, \boldsymbol{X}, \beta, \theta)} \\ &= \frac{P(Y_i | S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta) P(G | \theta)}{P(Y_i | \boldsymbol{a}, \boldsymbol{X}, \beta, \theta)} \end{aligned}$$

In a standard missing data problem, one would impute the missing covariates directly, however, due to the dependence through the graph, this can be very difficult to achieve in practice. However, it will be straightforward to sample from the graph model $P(G|\theta)$. Using

a simple approach, we can compute the maximizer exploiting standard software methods using an EM algorithm (36; 84). Suppose that we draw a sample of graphs from the generative model $\{G^{(l)}\}_{l=1}^{L} \sim_{iid} P(G|\theta)$.

Let $w_i(Y_i, G; \beta)$ define the weight of an observation.

$$w(Y_i, G; \beta) = \frac{P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)}{P(Y_i|\boldsymbol{a}, \boldsymbol{X}, \beta, \theta)} \approx \frac{P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)}{\frac{1}{L}\sum_{l=1}^{L} P(Y_i|S_i(\boldsymbol{X}, G^{(l)}), V_i(\boldsymbol{a}, G^{(l)}); \beta)}$$

The EM algorithm can now be defined as follows.

(1) Sample $\{G^{(l)}\}_{l=1}^{L} \sim_{iid} P(G|\widehat{\theta})$ denote a sample from the graph model and initialize parameters $\widehat{\beta}^{(0)}$

(2) For $t \in \{1, 2, \ldots, T\}$

(a) (E-step) Compute the weighted empirical estimating function

$$m_n^{(t)}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{X}, \beta, \widehat{\theta}) = \frac{1}{L}\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n} \widetilde{m}(Y_i, S_i(\boldsymbol{X}, G^{(l)}), V_i(\boldsymbol{a}, G^{(l)}); \beta) w(Y_i, G^{(l)}; \widehat{\beta}^{(t-1)})$$

(b) (M-step) Solve the new estimating function by solving:

$$m_n^{(t)}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{X}, \widehat{\beta}^{(t)}, \widehat{\theta}) = 0$$

In practice, this allows for one to use standard solvers for the (M-step), after sampling a single time with the (E-step).

Additionally, one can include correlations across the observations Y_i through the use of a generalized estimating equation approach. In other generalized linear models, additional assumptions may be required in order to model the full conditional distribution $P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)$ such as a dispersion component.

8.2. **Optimal design for a Z-estimator.** Here we illustrate the optimal design approach for Z-estimators. In this example, the variance itself may depend on the a parameter β, and thus one can include a working candidate for the parameter β'.

Algorithm 3 Saturation Randomized Design Variance.

Inputs: Working covariance Γ_n, model estimate $\widehat{\theta}$, working parameter β'
Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L} \sim \widehat{\theta}$
Sample R treatments $\{\boldsymbol{a}_r\}_{r=1}^{R}$ according to the block-saturations $\boldsymbol{\tau}$.
for $r \leftarrow 1$ **to** R **do**
    Compute $\widehat{D}_r(\boldsymbol{a}) = \frac{1}{nL}\sum_{l=1}^{L}\sum_{i=1}^{n} \nabla_\beta m_i(Y_i, S_i V_i; \widehat{G}^{(l)}, \beta')$
    Compute the variance for a single draw $\boldsymbol{a}_r$:

$$\upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta}) = \phi^T \widehat{D}_r(\boldsymbol{a})^{-1} \Gamma_n \widehat{D}_r(\boldsymbol{a})^{-1T} \phi$$

end for
Average over each of the draws $\overline{\upsilon}(\boldsymbol{\tau}; \widehat{\theta}) = \sum_{r=1}^{R} \upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta})$

9. Additional Simulations

9.1. **Coverage of the GATE.** In our simulation setup in Section 5.1 we can also compute confidence intervals based on the regression $Y_i = \beta^T \mathbb{E}[\widetilde{h}(S_i, V_i)] + \epsilon_i$ where we apply the Eicker-Huber-White sandwich estimator of the variance. We then compute the corresponding plug-in estimator of the variance using the covariates observed and lemma 3.6. Since

the covariates in the true regerssion model behave like averages over the graph, we expect lemma 3.8 to hold and therefore the difference between the GATE for any one draw of the graph, and the true GATE is very small. We see in fig. 9 that the coverage tends to be larger than the nominal 95%, though in general, due to model misspecification of the true-graph, there can be additional uncertainty due to the misspecification of the graph model. However, we see in this simple example that the coverage performs well with an off-the-shelf implementation.

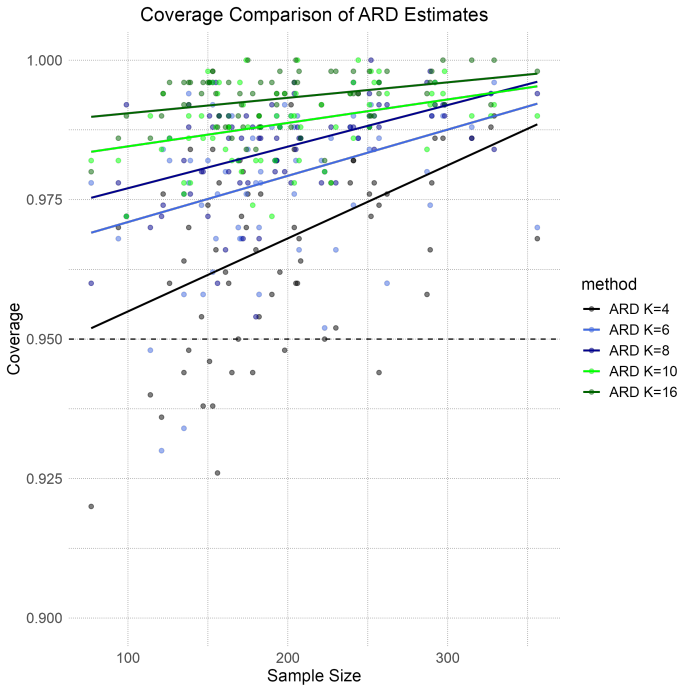


FIGURE 9. Coverage of the GATE using Eicker-Huber-White estimates of the variance.

9.2. **Inference for evidence of complex contagion with partial network data.** We can also replicate the results of (18)'s study on the evidence of pitplanting. They consider 3 measures of information diffusion. Firstly, if an individual has heard of pitplanting, second, if they know how to pitplant, and thirdly whether they adopt pitplanting in their practice. In order to control for one's position in the network, the authors consider the distance between the optimal seeds using two other targeting methods, simple diffusion, and geo targeting as well as complex contagion. They then compare the increased odds of con

$$Y_{iv} = \alpha + \beta_1 I(1TSeeds) + \beta_2 I(2TSeeds) + \beta_3 I(1Simple)_{iv} + \beta_4 I(2Simple)_{iv} + \beta_5 I(1Complex)_{iv} + \beta_6 I(2Complex)_{iv} + \beta_7 I(1Geo)_{iv} + \beta_8 I(2Geo)_{iv} + \delta_v + \epsilon_{iv}$$

Again, we generate synthetic covariates and apply a stochastic blockmodel in order to estimate $K = 8$ blocks within each of the networks. We plot the coefficients for the connection to exactly 1 seed, 2 seeds and within radius 2 of at least 1 seed in Figure 10[10].

[10]We note that we run the same regression as in (18), however, some since the full network data includes some additional noise top preserve anonymity, we do not have the exact same estimates of the coefficients as in their paper, however, the conclusions are substantively the same.

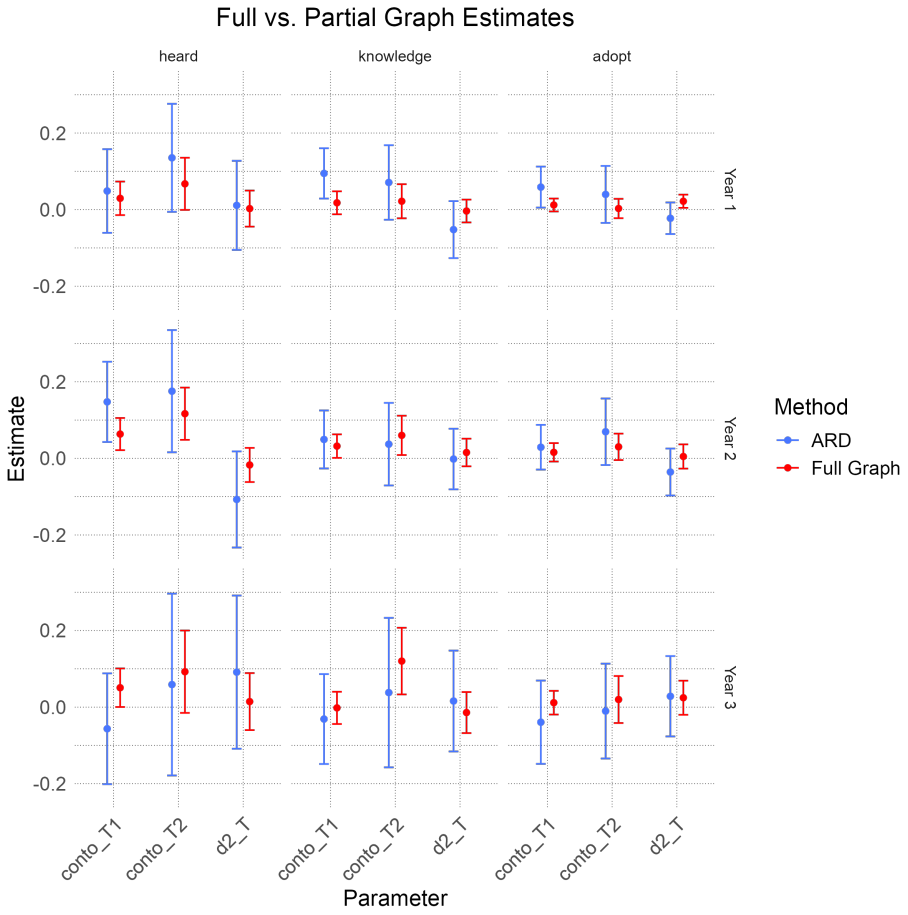


FIGURE 10. Replication of regression coefficients using aggregated relational data and associated 95% confidence intervals.

10. Additional Experimental Details

10.1. **Beta Model Estimation.** Another common model utilized for random graph formation is the beta model coined by (31). Namely these are a class of models that can be learned based on their degree sequence. We consider a version where each node has an affinity parameter ν_i and the probability of connection between each pair of nodes is $P(G_{ij} = 1) = \nu_i \nu_j$. Let $\nu_n = \sum_{i=1}^{n} \nu_i$ Therefore, $\mathbb{E}[d_i = d] = \sum_{j \neq i} P(G_{ij} = 1) = \nu_i(\nu_n - \nu_i)$. The set of parameters $\{\nu_i\}_{i=1}^{n}$ can be estimated using an iterative solution to the fixed point equation:

$$\nu_i^{(t+1)} = d_i / (\nu_n^{(t)} - \nu_i^{(t)})$$

10.2. **ARD Questions.** We utilize the measured traits to construct responses for ARD questions for each individual for the networks in (14). The constructed ARD include traits which ask "How many people do you know ..."

- that are in each sub-caste?
- that are Farmers, Shop owners, Domestic workers etc. ?
- that own their house?
- that have a house with at least 3 rooms?
- that have access to electricity?

For the estimation of the GATE using (12), we use Leiden clustering and denote the clusters traits. When replicating the results of (18), only a subset of nodes have available covariate. As was done in our examples with (12), we construct synthetic traits using the clusters observed from Leiden clustering for $K = 10$. ARD is then constructed based on the connections to nodes of each trait.

10.3. **GATE Estimators.** The two estimators we compare for estimation of the global average treatment effect are the difference in means estimator $\widehat{\tau}_{DM}$ and the Horvitz-Thompson estimator $\widehat{\tau}_{HT}$. Let E_{i0} and E_{i1} denote the events that all neighbours of i are untreated (including i themselves) and treated respectively.

$$\widehat{\tau}_{DM} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i a_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-a_i)$$
$$\widehat{\tau}_{HT} = \frac{1}{n}\sum_{i=1} \frac{Y_i I(E_{i1})}{P(E_{i1})} - \frac{Y_i I(E_{i0})}{P(E_{i0})}$$

In general, the Horvitz-Thompson estimator will be unbiased, however, it can often suffer from high variance for two reasons. Firstly, the probabilities of the events that all nodes are treated may be exceedingly low, inflating this variance, and also, relatively few nodes receive the exposures under which all of their neighbours are treated or none of them are.

In the case where the spillover effects are relatively mild, often a difference in means approach to the estimator is preferred. The effect of cluster randomization on the MSE of this estimator has been further studied in the complete network(22; 83).